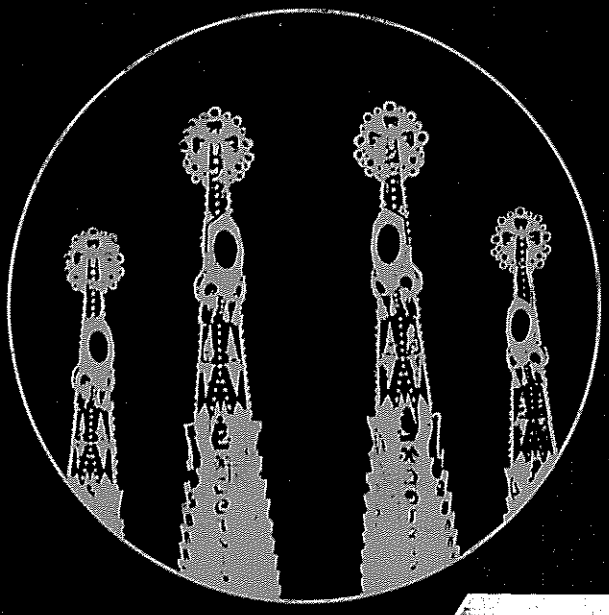
Volume II

SIGNAL PROCESSING V

THEORIES AND APPLICATIONS



L. Torres
E. Masgrau
M.A. Lagunas
editors

Elsevier

UT 6.7
Inv.Nr. 6520b

Torres, L.; Masgrau, E.; Lagunas, M. (eds.):
5th European Signal Processing Conference, Sept.18-21,1990, Barcelona. Vol.2.

Amsterdam: Elsevier Science Publishers, 1990.

Eigentum
des Inst. f. Nachrichtentechnik
und Hochfrequenztechnik
Technische Universität Wien
Inventar Nr. F6522 6 10.82

SIGNAL PROCESSING V
THEORIES AND APPLICATIONS

SIGNAL PROCESSING V
THEORIES AND APPLICATIONS

Proceedings of EUSIPCO-90
Fifth European Signal Processing Conference
Barcelona, Spain, September 18–21, 1990

Edited by

Luis TORRES
Enrique MASGRAU
Miguel A. LAGUNAS

Department of Signal Theory and Communications
ETSIT-UPC
Barcelona, Spain

EURASIP

Eigentum
des Inst. f. Nachrichtentechnik
und Hochfrequenztechnik
Technische Universität Wien
Inventar Nr. F 6520 6 11.9.92.

VOLUME II

1990

ELSEVIER
AMSTERDAM • NEW YORK • OXFORD • TOKYO

ELSEVIER SCIENCE PUBLISHERS B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

Distributors for the United States and Canada:

ELSEVIER SCIENCE PUBLISHING COMPANY INC.
655 Avenue of the Americas
New York, N.Y. 10010, U.S.A.

ISBN: 0 444 88636 2

© Elsevier Science Publishers B.V., 1990
© British Crown Copyright, 1990: pp. 433-436

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publisher, Elsevier Science Publishers B.V./Physical Sciences and Engineering Division, P.O. Box 1991, 1000 BZ Amsterdam, The Netherlands.

Special regulations for readers in the U.S.A. – This publication has been registered with the Copyright Clearance Center Inc. (CCC), Salem, Massachusetts. Information can be obtained from the CCC about conditions under which photocopies of parts of this publication may be made in the U.S.A. All other copyright questions, including photocopying outside of the U.S.A., should be refferred to the copyright owner, Elsevier Science Publishers B.V., unless otherwise specified.

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein.

pp. 123-126, 177-180, 309-312, 433-436, 441-444, 633-636, 757-760, 813-816, 817-820, 913-916, 971-974, 975-978, 1059-1062, 1175-1178, 1255-1258, 1287-1290, 1291-1294, 1499-1502, 1539-1542, 1575-1578, 1595-1598, 1671-1674, 1679-1682, 1707-1710, 1739-1742, 1747-1750, 1791-1794, 1807-1810, 1839-1842, 1847-1850, 1875-1878, 1951-1954, 1995-1998, Copyright not transferred.

This book is printed on acid-free paper.

Printed in The Netherlands

FOREWORD

EUSIPCO-90, the European Signal Processing Conference, is the fifth of the International Conferences promoted and organized by EURASIP, the European Association for Signal Processing. This book (in three volumes) presents the Proceedings of the Conference. The conference was held September 18-21, 1990 in Barcelona, Spain.

EUSIPCO-90 consisted of 52 sessions organized in 6 parallel programs. The Scientific Committee reviewed over 710 submitted abstracts to select the 515 that were accepted for presentation at the conference. Each abstract was reviewed by at least 2 independent reviewers. In addition, 11 tutorials were given by well known experts in the areas.

The Technical Sessions were organized in 7 broad topics according to the following distribution:

A. THEORY OF SIGNALS AND SYSTEMS
1. Detection. 2.Estimation. 3.Filtering. 4.Spectral estimation. 5. Adaptive systems.
6. Modelling. 7.Prediction
B. IMAGE PROCESSING
1. Coding. 2.Enhancement. 3.Restoration and reconstruction. 4.Biomedical processing
C. SPEECH PROCESSING
1.Coding. 2.Synthesis. 3.Recognition and understanding. 4.Enhancement.
5. Aids for the handicapped
D. MULTIDIMENSIONAL SIGNAL PROCESSING
1. Array processing. 2.Digital transforms. 3.Digital filtering. 4.Geophysical and seismic processing
E. IMPLEMENTATIONS
1. Hardware. 2.Software. 3.VLSI. 4.Novel architectures
F. KNOWLEDGE ENGINEERING AND SIGNAL PROCESSING
1.Expert systems. 2.Pattern recognition. 3.Signal interpretation. 4.Neural networks
G. APPLICATIONS
1. Radar. 2.Sonar. 3.Communications. 4.Digital audio. 5.Sensing. 6. Robotics

The volumes contents are as follow:
Volume 1: Theory of Signals and Systems. Multidimensional Signal Processing
Volume 2: Image Processing. Speech Processing
Volume 3: Implementations. Knowledge Engineering and Signal Processing. Applications

We would like to thank all the participants of EUSIPCO-90, as well as the sponsor institutions and exhibitors. Without any of them, EUSIPCO-90 would not have had place. Elsevier Science Publishers B.V. (North-Holland) provided all the help and care we needed. Angela Noguera and Silvia Soriano from the Department of Signal Theory and Communications, Barcelona, did an outstanding job in clerical work. Our gratitude to them.

Although many times we asked ourselves "why did we accept to organize EUSIPCO-90?", we have been very pleased to accept this commitment and hope these Proceedings will be helpful to the Signal Processing Community.

Barcelona, September 1990

Luis Torres
Enrique Masgrau
Miguel A. Lagunas

CONFERENCE COMMITTEE

M. A. Lagunas
Chairman

J. Fernández
A. Figueiras
A. Gasull
J.B. Mariño
E. Masgrau
A. Moreno
C. Nadeu
L. Torres
F. Vallverdú
G. Vázquez

SCIENTIFIC COMMITTEE

ACHA, J.
ALMEIDA, L.B.
AMENGUAL, M.
BAJON, J.
BELLANGER, M.
BERTRAN, M.
BIEMOND, J.
BOHME, J.
BOITE, R.
BRACCINI, C.
CARAYANNIS, G.
CASACUBERTA, F.
CASALS, A.
CASAR, J.R.
CASTANIE, F.
CISNEROS, G.
CLARKE, R.J.
COLLA, A.M.
CONSTANTINIDES, A.G.
CORTELAZZO, G.
COWAN, C.F.N.
DOCAMPO, D.
DUBUISSON, B.
ELIAS, A.
FARINA, A.
FARRIER, D.
FETTWEISS, A.
FIGUEIRAS, A.R.
FLANDRIN, P.
GALAND, C.
GARCIA, R.
GARCIA, N.
GAUDENZI, R.
GILGE, M.
GLANGEAUD, F.
GRANLUND, G.H.
GRIFFITHS, J.W.R.
HEES, W.
HEUTE, U.
HOFFMANN, J.C.
JUTTEN, J.C.
KUNT, M.
LABARTA, J.
LACAZE, B.
LACOUME, J.L.
LLABERIA, J.M.
LUCAS, R.
MACCHI, O.
MARCOS, S.
MARTHON, M.
MARTIN, N.
MARTINEZ, J.A.
MECKLENBRAUKER, W.
MUÑOZ, C.
NEUVO, Y.
NIEMANN, H.
NOLL, P.
ORTIGUEIRA, M.
PARDO, J.M.
PICINBONO, B.
RENDAS, M.Y.
RIBEIRO, I.
ROCCA, F.
ROUX, C.
RUBIO, A.J.
SALLENT, S.
SAN FELIU, A.
SANTOS, J.
SANZ, A.
SCAGLIOLA, C.
SCHUSSLER, H.W.
SECILLA, J.P.
SICURANZA, G.L.
SMITH, S.G.
THEODORIDIS, S.
TORRAS, C.
TRANCOSO, I.
VARY, P.
VENTURA, J.
VICENZI, C.
VIDAL, E.
VIOLA, R.
WOLF, D.

ORGANIZATION

Organizer:

European Association for Signal Processing
EURASIP

Sponsors

Dirección General de Investigación Científica y Técnica
Ministerio de Educación

CIDEM Centre d'Informació i Desenvolupament Empresarial
Generalitat de Catalunya

UPC. Universitat Politècnica de Catalunya

CIRIT
Generalitat de Catalunya

IBM

CESELSA

Grupo AMPER

TELETTRA España, S.A.

PAGE IBERICA, S.A.

COMELTA, S.A.

DIGITAL Equipment Corporation. España

TID. Telefónica Investigación y Desarrollo

ELSEVIER- Science Publishers B.V.

HEWLETT PACKARD

EXHIBITORS

BOIXAREU Editores, S.A.
COMELTA, S.A.
CONTROL SYS
DIGIMETRIE
ELSEVIER- Science Publishers B.V.
FUJITSU
HORIZON TECHNOLOGIES
LOUGHBOROUGH SOUND IMAGES Ltd.
NEC Electronics
OROS
SGS - Thomson
SIGNAL TECHNOLOGY Inc.
TEXAS INSTRUMENTS
VTE Digitalvideo GmbH

TECHNICAL PROGRAM SCHEDULE

Tuesday, September 18

Time	Room 7	3	5a	5b	6	8	4
8h 40 – 9h 10 – 9h 15	Opening Ceremony						
10h 35 – 11h 00	L1 Speech Enhancement	L2 Time Delay Estimation	L3 Signal Enhance. and noise Reduction	L4 Geophys. and Seismic Process.	L5 Novel Architec.	L6 Sensing/ Robotics	P1 Image Sequence Coding
13h 00 – 14h 40							
		Tutorial 1			Tutorial 2		
15h 10 – 15h 15							
16h 55 – 17h 15		L7 Neural Nets I	L8 Speech Synthes.	L9 Time-Frequen. Signal Analysis	L10 Radar Signal Process. I	L11 Image Enhance. and Restorat.	P2 Communications I
18h 15							

Wednesday, September 19

Time	Room 3	5a	5b	6	8	4
8h 40 – 9h 10	Tutorial 3			Tutorial 4		
9h 15 – 10h 35 – 11h 00	L12 Adaptive Filtering I	L13 Image Coding	L14 Speech Coding	L15 Underwat. Acoustics	L16 VLSI Implementation	P3 Communications II
13h 00 – 14h 40						P4 Pattern Recognit./ Signal Interpr.
	Tutorial 5			Tutorial 6		
15h 10 – 15h 15						
16h 55 – 17h 15	L17 Neural Nets II	L18 Speech Recognit. In Noise	L19 Echo Cancelling and Deconvol.	L20 Estimation and Identification	L21 Transform Image Coding	P5 Hardware Implem./ Software Tools for VLSI
18h 15						

Thursday, September 20

Time	Room 3	5a	5b	6	8	4
8h 40 – 9h 10	Tutorial 7			Tutorial 8		
9h 15 – 10h 35 – 11h 00	L22 Speech Coding I	L23 Adaptive Filtering II	L24 Array Process.: Spatial-Spectrum Estimat.	L25 Biomedical Image Processing	L26 Dynamic Scene Analysis	P6 Speech Recognition I
13h 00 – 14h 40						P7 Hardware and Software for DSP
	Tutorial 9			Tutorial 10		
15h 10 – 15h 15						
16h 55 – 17h 15	L27 Array Process.	L28 Knowled. Engineer.	L29 Acoustic Echo Control	L30 Speech Processing	L31 Detection	P8 Image Process. and Machine Vision
18h 15						

Friday, September 21

Time	Room 3	5a	5b	6	8	4
8h 40 – 9h 10	Tutorial 11					
9h 15 – 10h 35 – 11h 00	L32 Line Detection	L33 Speech Coding II	L34 Modelling	L35 VLSI for Multidimensional Signal Processing	L36 Speech Recognition II	P9 Multidimens. Filtering
13h 00 – 14h 40						P10 Adaptive Filtering III
15h 10 – 15h 15						
16h 55 – 17h 15	L37 Array Process: Adaptive Beamform.	L38 Spectral Estimat.	L39 Radar Signal Process. II	L40 Vector Quantizat. Image Coding	L41 Modelling/ Signal Theory	P11 Filtering
18h 15						

L = Lecture P = Poster

VOLUME I

OPENING SESSION

TUTORIALS

THEORY OF SIGNALS AND SYSTEMS

TIME - DELAY ESTIMATION

SIGNAL ENHANCEMENT AND NOISE REDUCTION

TIME - FREQUENCY SIGNAL ANALYSIS

ADAPTIVE FILTERING I

ESTIMATION AND IDENTIFICATION

ADAPTIVE FILTERING II

DETECTION

LINE DETECTION

MODELLING

ADAPTIVE FILTERING III

MODELLING / SIGNAL THEORY

FILTERING

MULTIDIMENSIONAL SIGNAL PROCESSING

GEOPHYSICAL AND SEISMIC PROCESSING

ARRAY PROCESSING: SPATIAL-SPECTRUM ESTIMATION

ARRAY PROCESSING

MULTIDIMENSIONAL FILTERING

ARRAY PROCESSING: ADAPTIVE BEAMFORMING

VOLUME II

IMAGE PROCESSING

IMAGE SEQUENCE CODING

IMAGE ENHANCEMENT AND RESTORATION

IMAGE CODING

TRANSFORM IMAGE CODING

BIOMEDICAL IMAGE PROCESSING

DYNAMIC SCENE ANALYSIS

IMAGE PROCESSING AND MACHINE VISION

VECTOR QUANTIZATION IMAGE CODING

SPEECH PROCESSING

SPEECH ENHANCEMENT

SPEECH SYNTHESIS

SPEECH ANALYSIS

SPEECH RECOGNITION IN NOISE

SPEECH CODING I

SPEECH RECOGNITION I

SPEECH PROCESSING

SPEECH CODING II

SPEECH RECOGNITION II

VOLUME III

IMPLEMENTATIONS

NOVEL ARCHITECTURES

VLSI IMPLEMENTATION

HARDWARE IMPLEMENTATION / SOFTWARE TOOLS FOR VLSI

HARDWARE AND SOFTWARE FOR DSP

VLSI FOR MULTIDIMENSIONAL SIGNAL PROCESSING

KNOWLEDGE ENGINEERING AND SIGNAL PROCESSING

NEURAL NETS I

PATTERN RECOGNITION AND SIGNAL INTERPRETATION

KNOWLEDGE ENGINEERING

APPLICATIONS

SENSING/ROBOTICS

RADAR SIGNAL PROCESSING I

COMMUNICATIONS I

UNDERWATER ACOUSTICS

COMMUNICATIONS II

ECHO CANCELLING AND DECONVOLUTION

ACOUSTIC ECHO CONTROL

RADAR SIGNAL PROCESSING II

AUTHOR INDEX

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Software Architecture for TV/HDTV Codec Simulation*

Narciso García, Fernando Jaureguízar, José I. Ronda, and Alberto Sanz

Grupo de Tratamiento de Imágenes, E.T.S. Ingenieros Telecomunicación
Universidad Politécnica de Madrid, E-28040 Madrid, Spain

The design of digital codecs for TV/HDTV signals require flexible and powerful simulation workbenches to conduct algorithmic studies. A SW architecture considering the data-flow paradigm and using Abstract Data Types (ADT) as a main software implementation paradigm has been proposed, and a simulator based on it has been designed and constructed. It runs on a general purpose computer architecture and allows for important modifications and short tunings for diverse algorithmic analysis and testing, while keeping very low execution times. Besides the description of the signal processing and SW architectures, several application examples are provided. Achieved results are also presented.

1. Introduction

Image communications are stemming as digital technologies are able to provide new acquisition, processing, transmission, storage, and display facilities. Nevertheless, quality demanding visual services generate signals with so incredible huge amounts of raw information that the available transmission or storage facilities cannot cope with them. Therefore, there is a need for systems, that maintaining the visual information, are able to reduce the associated data amount required for its representation. An initial idea is to recover the same visual information, after reducing and restoring the associated data, assuring that there is no mathematical difference between the original and the data-reduced visual signals. Usually, a dissimilarity between these signals is allowed to achieve a higher data-reduction factor, arising a trade-off design between compression (data-reduction capabilities) and quality (similarity). Depending on the considered communications service, two transmission classes are considered: contribution-quality (the data-reduced visual signal could be processed again) and distribution-quality (no new processing is considered).

The design of a digital codec for visual information transmission is a long process between its original inception and the building of the final system. Initially, an algorithm should be selected among the extense set of available solutions; afterwards, the encoding parameters should be matched to the considered visual signal; then, the available HW technologies should be evaluated; and, finally, the codec will be built. However, some basic research on algorithms should be conducted before one is selected, and extensive efforts should be performed to tune the selected algorithm for properly encoding the visual signal. Moreover, all these steps must be completed before HW solutions are evaluated. Therefore, simulation appears as the solution to provide a workbench for algorithm evaluation and parameter tuning.

The main requirements for a simulation environment are twofold: flexibility (easy tailoring of the simulation tools) and speed (image sequence processing in a reasonable amount of time). As a general principle, the first requirement can be addressed by a very high level application oriented language which can be enriched to include a growing number of options and new facilities. The second one can be fulfilled by a parallel computing architecture which will accelerate the calculations.

Currently, there is a worldwide effort for the standarization of the digital transmission of TV/HDTV signals [1,2]. Digital codecs for these signals require extensive simulation on algorithms and implementations for the analysis of system performances, mainly compression capabilities and subjective quality of the recovered image, as the final design will be used as an international standard for a long time. Therefore, the previously mentioned simulation requirements are critical, as extensive simulations on a wide set of complete sequences of different compression schemes should be conducted.

Here, a SW architecture for standard (conventional or parallel) computing systems is presented. It is based on the Abstract Data Types (ADT) paradigm, which allows the definition of "SW building blocks" that can be easily combined [3]. Impossing restrictions on memory management, fast ADT are obtained. Therefore, this paradigm is the basis for flexibility and speed, and, so, a good choice for a simulation tool.

A simulator based on these SW architecture and considering the data-flow paradigm has been built [4]. Its design has been done in a hierarchical way, just following the hardware solution proposed within Eureka-256 [5]. Initially, the signal processing architecture will be presented; then, the SW architecture will be discussed; afterwards, applications will be described and, finally, conclusions will be stated.

This work has been done within Eureka-256: "Bit-Rate Reduction System for HDTV Digital Transmission". It has been partially supported by the Plan Electrónico e Informático Nacional and the Comisión Interministerial de Ciencia y Tecnología of the Spanish Government.

2. Signal Processing Architecture

The visual information signal, either TV or HDTV, is processed following the approach considered within both CMTT/2 [1,2] and Eureka-256 [4,5], where a Hybrid-DCT has been selected for the compression scheme. So, prediction is applied along the temporal axis and DCT (transform encoding) is applied along the spatial ones. Figure 1 presents the overall system architecture of the encoder; the decoder is equal to the encoder feed-back loop. The figure shows the functional operation of the system, based on signal processing oriented functions. Operations are shown as modules and interchanged data are placed in between.

As DCT should be computed on blocks, preferable square, block size is the first design parameter. Size 8 rows by 8 columns has been chosen taking into account compression efficiency and computational cost. Therefore, a field holding both luminance (Y) and chrominance (C_R and C_B) can be considered as tesselated into equal sized blocks. A superblock or macroblock is the minimum structure holding luminance and chrominance. Considering the sampling lattice defined in the CCIR Recommendation 601 for TV signals, it holds an area of 8 rows by 16 columns, having two luminance blocks and two chrominance ones (one C_R and one C_B). Currently, there is no final definition for HDTV signals, but the european approach considers the same sampling lattice than that of TV, so superblocks are similar. A stripe is a full horizontal band 8 rows high.

Prediction along the time axis is performed choosing the best fit among the area to be DCT encoded and the prediction set. For the TV/HDTV signals, this prediction set is formed by four predictions: zero (intrafield, no prediction at all), the equivalent area in the last field (interfield prediction), the same area in the previous to the last field (intraframe prediction), and the motion compensated area in the previous to the last field (motion compensated prediction). The prediction selection can be done *a priori*, just evaluating a measure of goodness for all prediction fitnesses, or *a posteriori*, choosing the minimum the code length for the DCT encoding of all the alternatives. The prediction selection is performed on a superblock basis, taking into account the allowed prediction combinations between luminances and chrominances within the superblock.

Figure 1 presents an *a priori* approach for the temporal prediction, but the *a posteriori* approach has also been considered, particularly for distribution quality transmission environments.

Transform encoding is performed on every block of the superblock, as presented in Figure 2. The signal processing chain includes DCT, Scaling and Quantizing, where Scaling performs the trade between code length (compression) and quantization fineness (quality), thus allowing buffer regulation. VLC encoding reduces code length by assigning word lengths inversely proportional to word probabilities. The feed-back is provided by the *Analyze Buffer Occupancy* module which computed the transmission factor, for controling the scaling procedure, taking into account the buffer occupancy.

The algorithms for the estimation of the motion vectors and the buffer control are those under study within Eureka-256.

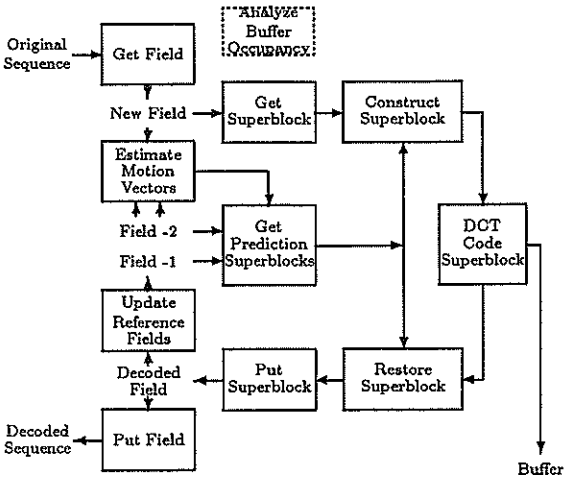


Fig. 1 – General architecture.

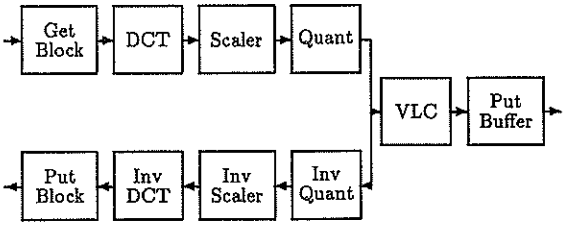


Fig. 2 – DCT_code_superblock.

3. SW Architecture

A simulator for TV/HDTV encoding has been built based on the data-flow paradigm as it presents the best fit to the visual information processing. Also, a very flexible programming paradigm, such as a functional language or an Abstract Data Type (ADT's) implementation, has been selected. The use of these paradigms would increase enormously the execution times for a practical system, and, for this reason, an efficient implementation of ADT's by means of pointers along with the use of multifunctional modules [6] and standard functions as mainline for the algorithms, was selected as basic foundation of the specification.

The software architecture is therefore built around the following module types:

- *Abstract Data Types (ADT)*, each one consisting of an object allocated in the machine heap and whose contents are accessible via an associated set of operations which have been implemented as functions. Several instances of an ADT can be generated and manipulated with the same set of operations. The basis of abstract data types functional manipulation are preserved, but some violations to this scheme have been defined in order to provide efficient implementation mechanisms, enforcing overall system performance.

- *Multifunctional Modules (MM)*, consisting of a global object or set of objects (such as ADT's), allocated inside a program, which can only be manipulated through a limited set of functions. The object associated to the MM is a unique instance, being for all practical purposes, a global element with restricted access capabilities.

- *Functions*, consisting in single defined functions which perform a specified algorithm. Within the functions, instances of ADT's will be created, manipulated and deleted, through the ADT operations. A certain algorithm specifies how this tasks are performed. Functions are the upper level of the design.

The most relevant issue of the ADT implementation is the absence of significant data movement, as ADT's with inherited memory are used extensively. This allows for a very fast operation while retaining the basic properties of ADT's in readability and maintainability. The price paid is the need for a unique allocation and deallocation of memory for the manipulated ADT's within the framework of a single function, so that new instances of ADT's are neither created nor destroyed throughout the execution of a simulation for a sequence, after the initial allocation and before the final deallocation.

Based on them, a very readable, easy to maintain, and highly reliable code has been obtained, that has a very fast execution speed.

4. Parallelization scheme for HDTV

As HDTV signals hold a extreme high binary rate, it is impossible to consider a monoprocessor operation, provided standard HW technologies are used. Therefore, either a more sophisticated HW technology is used, or the computational burden is divided among several processors. Eureka-256 follows the latter, considering individual processors formed by complete hybrid DCT TV codecs [4,5]. So, the HDTV codec is built on the association of standard TV codecs, each operating on a different area of the field.

To avoid the need of large buffers between the full HDTV image and the parallel codecs, the splitting approach should follow a procedure of line breaking into equal sized segments. Therefore, it implies a division into vertical bands. Hereafter, each one of these vertical bands will be called a subfield.

As each one of the codecs varies the quality of the decoded image depending on the buffer occupancy, a quality synchronization procedure is required, governing the quality of all the codecs through the occupancy of the common buffer. For each one of them, there is no difference with its individual operation, as it works always on an external parameter that in this case is provided by a common HDTV system control.

The general encoder architecture is presented in Figure 3. Regarding the SW design, no new ADT are required, since subfields are implemented as instances of the field abstract data type. Parallelism is simulated by interlacing stripe by stripe the operation of all codecs,

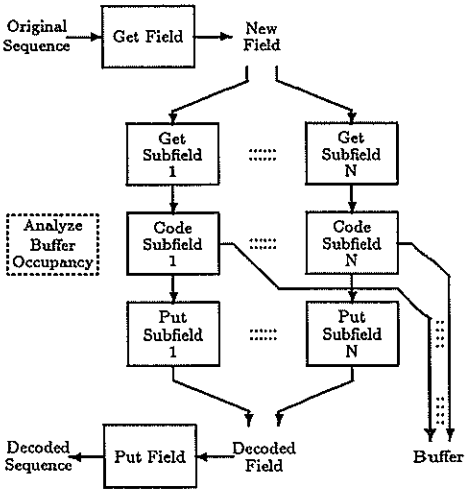


Fig. 3 – General parallel architecture.

i.e., the process starts with the first stripe of the first subfield, continues with the first stripe of the second subfield, etc. Other parallel mechanisms can be easily simulated. A real multiprocessor execution can also be performed.

5. Applications

Several applications of the simulator, built based on the previous introduced architecture, are presented. They take advantage of the flexibility achieved through the modular design and the special organization of the outputs.

The modularity of the simulator allows the testing of different algorithms with a minimal software maintenance cost. So, different approaches can be studied for DCT statistics, buffer control, motion vectors estimation, motion vector set reduction, etc. Other evaluations can also be carried out, such as the reduction of the choices for the superblock prediction mode, different estimators of the variance, interpolation filters for temporal prediction, etc.

The special output organization allows the introduction of "data-probes" within the simulator code, and the user can specify at the execution time, which ones are going to be active and provide data, and which program is going to receive and process them. With the use of these "data-probes" no recompilation is needed to obtain different configurations for the output of the simulator. Different programs can be devised to handle the data from the probes while the simulator remains unchanged, thus reducing the number of its different versions and making it easier the maintenance of the complete software package.

5.1. Statistics computation

The simulator has been used to model the TV signal in the DCT-domain, allowing the computation of coefficient density probabilities, the analysis of the inter/intra mode selection, the characterization of the code words by prediction mode and component type, ... [7].

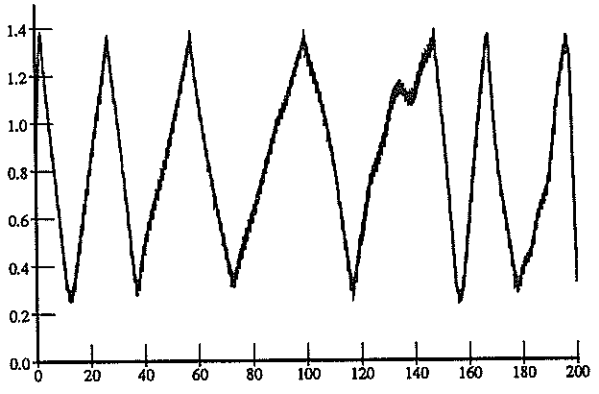


Fig. 4 – Evolution of the buffer occupancy.

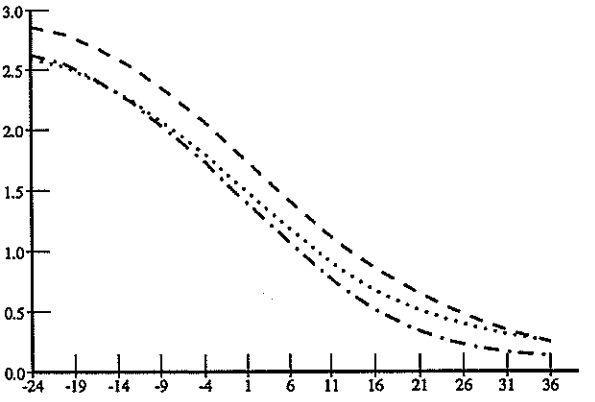


Fig. 5 – Average bit amount per field in Mbit vs. transmission factor.

5.2. Movement vectors estimation

The simulator has been used to evaluate the alternatives for the estimation of the motion vectors required for the motion compensated prediction. Several methods have been tested including the exhaustive, the hierarchical and the one at a time search (OTS) [8].

5.3. Buffer evolution analysis

The simulator has been used to analyze the evolution of the buffer occupancy and the transmission factor to evaluate different alternatives for buffer control. Figure 4 shows the evolution of the buffer occupancy for a digital TV test sequence.

Fixing the transmission factor to a pre-specified value, it is possible to compute the required bit-rate for a given quality (transmission factor). Figure 5 presents this bit-rate distorsion curve for three digital TV test sequences.

6. Conclusions

A software architecture for TV/HDTV codec simulation has been presented. The use of Abstract Data Types combined with multifunctional modules allows for a great deal of flexibility in the definition and combination of "software blocks". Structuring obtained by the approach is also responsible of the high efficiency which has been reached on non specialized computing machines. Possibilities of the scheme have been demonstrated with a number of tools built on this architecture for very diverse purposes: evolution of the transmission factor, statistics on DCT coefficients, rating of *a priori* versus *a posteriori* schemes and comparison of motion compensation techniques.

References

[1] *Draft New Recommendation: Transmission of Component-coded Digital Video Signals for Contribution-quality Applications at Third Hierarchical Level of CCITT, Recommendation G.702*, Document CMTT/303, October 1989.

[2] *Proposed Modifications to Report AD/CMTT: Digital Transmission of Component-Coded Television Signals at 30–34 Mbit/s and 45 Mbit/s.* Document CMTT/321, October 1989.

[3] J.A.Goguen, *Reusing and Interconnecting Software Components*, Computer, Vol. 19, No. 2, February 1986, pp. 16-28.

[4] N.García, F.Jaureguízar, J.I.Ronda, and A.Sanz, *HDTV Parallel Codec Simulator*, Proc. Third Int. Workshop HDTV, Torino, August 1989.

[5] G.F.Barbieri, F.Molo, and J.L.Tejerina, *A Modular and Flexible Video Codec Architecture for Application to TV and HDTV*, Proc. 16th Int. TV Symposium, Montreux, June 1989, pp. 410-420.

[6] D.L.Parnas, *On the Criteria to Be Used in Decomposing Systems into Modules*, Communications of the ACM, Vol. 5, No. 12, December 1972, pp. 1053-1058.

[7] J.I.Ronda, F.Jaureguízar, and N.García, *DCT-Domain Modelization of the TV Signal for Quantizer Design*, in these Proceedings.

[8] F.Jaureguízar, J.I.Ronda, and N.García, *Motion Compensated Prediction on HDTV*, in these Proceedings.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

THREE DIMENSIONAL ADAPTIVE LAPLACIAN PYRAMID IMAGE CODING

S.Sallent* L.Torres** L.Gils*

* Department of Applied Mathematics and Telematics, ** Department of Signal Theory and Communications

ETSETB-UPC, Apartado 30002, Barcelona 08034, Spain.

In this paper we propose a three dimensional Laplacian Pyramid coding scheme. The input image sequence is subsampled both spatially and temporally into different channels using proper three dimensional sampling structures. This results in the image sequence being represented by a series of bandpass sequences through three dimensional Gaussian and Laplacian pyramid data structures. In order to build the spatio-temporal pyramid structure several filters are discussed.

1 INTRODUCTION

The Laplacian Pyramid is a new and efficient method for image encoding (1). The method is of increasing interest as bandpass pyramids and multiresolution images are being used in other image processing applications. The pyramid image structure can be naturally adapted for progressive image transmission over low-speed channels and hierarchical image retrieving in computerized image storage.

The method obtains good compression rates and excellent visual quality for static images. It was then logical to extend the pyramid image structure using arbitary non-rectangular sampling lattices and characterize the sampling lattices by matrices, thus providing a compact and powerful notation (2).

On the other hand increasing interest is focused on image sequence coding. Applications such as videoconference, videotelephone and low bit rate image coding in general, are a key issue in current video communication systems. The paper we present proposes a new three dimensional coding scheme as an extension of the previously reported work on static images. The input image sequence is subsampled both spatially and temporally into different channels using proper three dimensional sampling structures. This results in the image sequence being represented by a series of bandpass sequences through three dimensional Gaussian and Laplacian pyramid data structures.

In order to build the temporal pyramid structure several spatial-temporal filters are discussed along with different sampling strategies. It is shown, as in the static image case, that the performance of the pyramid encoding system can be improved by proper selection of the sampling structures thus resulting in an adaptive and efficient encoding method. Motion compensation algorithms are also introduced in the scheme to further decrease the bit rate as is the case for hybrid methods.

2 GAUSSIAN AND LAPLACIAN DATA STRUCTURES

The temporal and spatial pyramid data structure represents the original image sequence into a set of code elements which are localized in spatial and temporal frequencies as well as in space and time. Each element in the new data structure is obtained by applying an appropriate three dimensional weighting function defined on an arbitrary sampling structure.

The image sequence, and particularly the video-conference sequences, are characterized by the high correlations of the neighboring pixels in the spatial and temporal dimensions. The three dimensional pyramid coding reduces the correlation by subtracting the original sequence $S_o^o(l,m,n)$ from the low-pass version sequence of itself $S_1^o(l,m,n)$, where l,m,n are the temporally and spatially coordinates respectively.

The code elements are obtained from those sequences as a suitable difference which represents the prediction error

$$D_o(l,m,n) = S_o^o(l,m,n) - S_1^1(l,m,n) \qquad (1)$$

being $D_o(l,m,n)$ more decorrelated than the original sequence.

Then rather, than encode $S^{\circ}_{o}(l,m,n)$ it encodes the set of band-pass sequences obtaining data compression. The compression is achived because the low-pass sequences are built through a decimation process associated with $\mathbf{M}_i$ matrix. Consequently the low-pass sequences are encoded at a reduced sample rate, and the high-pass sequences can be described with fewer bits.

Iterating this process over the low-pass sequence a set of low-pass $\{S^{\circ}_{i}(l,m,n)\}$ and band-pass $\{D_i(l,m,n)\}$ image sequences is obtained whose support regions are defined on sampling lattices characterized by a 3X3 $\mathbf{M}_i$ matrix.

Both data sets can be modeled as a pyramid data structures where each level is a sequence of decreasing dimension and resolution, the original sequence being the bottom of the pyramid.

The low-pass sequence set is constructed applying recursively the decimation algorithm

$$S^{\circ}_{i+1}(l,m,n) = dec\{S^{\circ}_{i}(l,m,n)\} = \sum_{o}\sum_{p}\sum_{q} W(o,p,q)\cdot S^{\circ}_{i}(M_i[l,m,n]^T + [o,p,q]^T)^T \quad (2)$$

where o, p, q belong to the support region of the three dimensional weighting function W, i is the pyramid level, with i=0 the bottom of the pyramid structure and $o \le i < L-1$.

The band-pass sequence set is built by applying recursively the interpolation algorithm

$$D_i(l,m,n) = S^{\circ}_{i}(l,m,n) - S^{1}_{i+1}(l,m,n) = S^{\circ}_{i}(l,m,n) - inter\{S^{\circ}_{i+1}(l,m,n)\} = S^{\circ}_{i}(l,m,n) - |det(M_i)| \sum_{o}\sum_{p}\sum_{q} W(o,p,q)\cdot S^{\circ}_{i+1}(M^{-1}(l-o\ \ m-p\ \ n-q)^T)^T \quad (3)$$

This algorithm is only evaluated for integer values of $S^{\circ}_{i+1}(l,m,n)$.

When the decimation and interpolation process uses like-Gaussian filters, the data sets are named Gaussian and Laplacian data structures.

3 THREE DIMENSIONAL LAPLACIAN PYRAMID CODING

The three dimensional Laplacian Pyramid Image coding with associated sampling lattices defined by a three by three matrices set $\{M_i\}$ is based on the transmission of the quantized set of band-pass sequences $\{D_i(l,m,n)\}$, L the number of total levels of the pyramid structure and $D_{L-1}(l,m,n) = S^{\circ}_{L-1}(l,m,n)$ is the top sequence of the pyramid.

The new coding scheme is implemented in four stages.

a- A set of L low-pass versions of the original sequence defined on sampling $\{M_i\}$ lattices is obtained by applying recursively the decimation algorithm. For the construction of each level i+1 an apropriate sampling lattice is chosen in order to provide the best information compaction. Thus resulting in each level having its proper sampling structure. In the frequency-domain the shape of the reciprocal unit cell associated to the sampling lattice $\mathbf{M}_i$ is compared to the frequency content of the low-pass sequence $S^{\circ}_{i}(l,m,n)$. The algorithm can be modeled as a low-pass filtering and down-sampling process.

b- The set of band-pass sequences are constructed by applying recursively the interpolation algorithm over the low-pass sequence. This algorithm can be modeled as an up-sampling process, low-pass filtering and a suitable difference. The weighting function and sampling lattice used are the same that have been used in the construction of the equivalent level in the low-pass sequence.

c- The set of band-pass sequences are quantized by laplacian quantizers. The parameters of the quantizers are selected acording to the statistics, the total chosen compression and the quality of the desired reconstructed sequence. The set of quantized

sequences $\{D_i(l,m,n)\}$ are transmited using variable length codewords.

***d*-** At the receiver, the original sequence is reconstructed applying recursively

$$S_i(l,m,n) = D_i(l,m,n) + |\det(M_i)| \cdot \sum_o \sum_p \sum_q W(o,p,q) \cdot S_{i+1}(M^{-1}(l-o \;\; m-p \;\; n-q)^{T,T}) \quad (4)$$

where $S_{L-1}(l,m,n) = D_{L-1}(l,m,n)$ and $S_o(l,m,n)$ is the reconstructed sequence. The use of arbitrary sampling lattices in the construction of low and band pass sequences allows to split the spectrum in regions of similar statistics adapting the coding process to the spatial and temporal characteristics of the sequence.

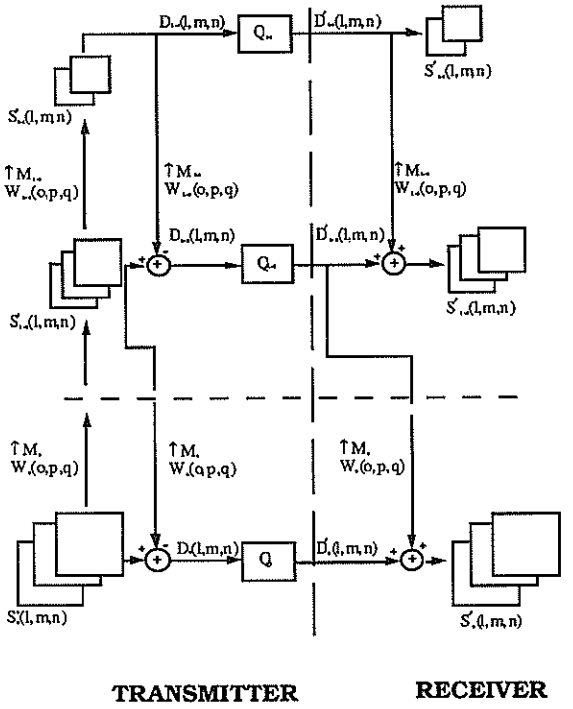


Figure 1 shows the block diagram of the Three Dimensional Laplacian Pyramid Coding.

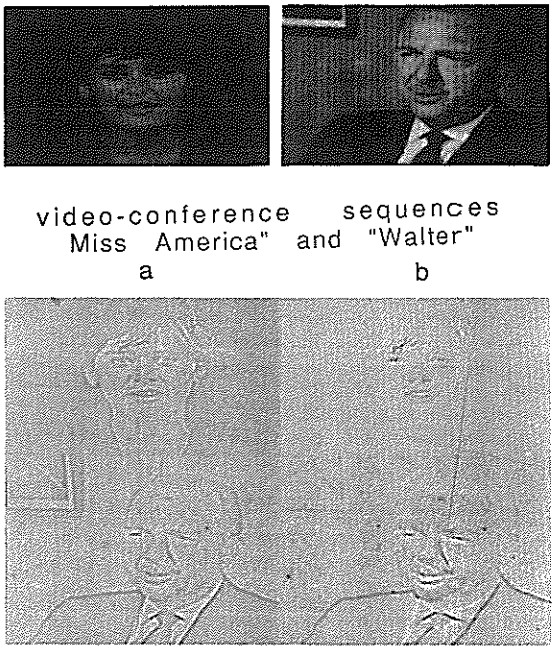
The prediction error can be improved applying motion compensation to the weighting function. A regular decomposition quadtree method (3) is applied to segment the interframe differential signal into homogeneous regions of different block sizes. Each region is characterized by a motion vector and used to correct the local weighted average of each pixel.

4 SIMULATIONS AND RESULTS

The results presented here were derived from two video sequences known as "Miss America" and "Walter" which are 256x256 pixels per frame with eight pixels per bit and twenty five frames per second.

In this coding method, the type of the filter has been chosen according to the appropriate matrix $\mathbf{M}_i$ associated to each level. For instance, figure 2 shows the first level of the Three Dimensional Laplacian Pyramid for " Miss America " sequence. In this case we use 3D spatio-temporal filter with 125 taps and 1D temporal filter 5 taps associated to

$$M_i = \begin{pmatrix} 200 \\ 020 \\ 002 \end{pmatrix} \quad \text{and} \quad M_i = \begin{pmatrix} 200 \\ 010 \\ 001 \end{pmatrix}$$



video-conference sequences Miss America" and "Walter"

a b

Figure 2 The first level of the Laplacian Pyramid $D_o(l,m,n)$ for "Miss America" and "Walter" using a) 125 tap spatio-temporal filter, and b) 5 tap temporal filter.

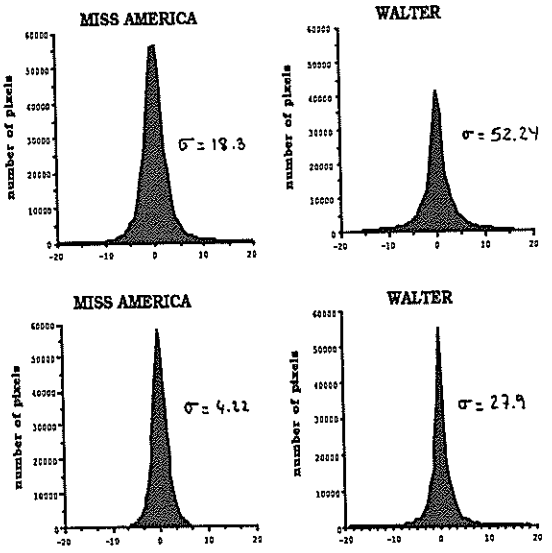


Figure 3 shows the histograms of the first level using the spatio-temporal and temporal filters described above of the walter and Miss America sequences.

Figure 4 shows results of consecutive coded frames 6, 7, 8 and 9 of the "Miss America" whith a compression ratio of 40, and frames 1, 2, and 3 of " Walter" sequence with a compression ratio of 20. Computer results are presented at 64 x 4 kbits/sec with excellent visual quality and a signal to noise ratio of 24 and 27db respectly. This rate can be lowered by applying motion compensation to the three dimensional weighting function.

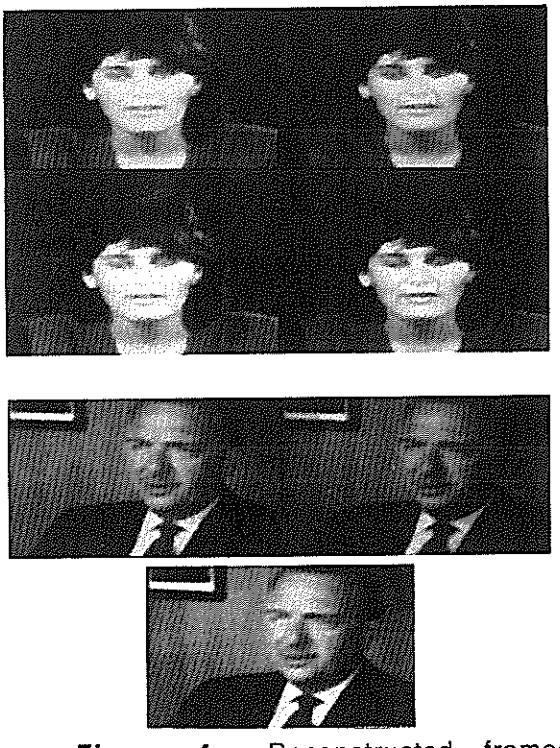


Figure 4 Reconstructed frames 6,7,8,9 of "Miss America" and 1, 2, 3 of "Walter" with a signal to noise ratio of 24 and 27 dB respectively.

The simulation results indicated that the proposed method is capable of operating very efficiently for a wide class of video applications such as 192 Kbits/second high definition video conferencing and in the range of B-ISDN hierarchies. It is shown that the scheme does not present any block effect, offers low computational complexity, and can be implemented in parallel.

5 CONCLUSION

We presented the three dimensional Laplacian Pyramid Coding used to obtain a high compression rate for wide variety of video applications. This method is the result of extending the Pyramid Coding to the three dimensions, using arbitrary sampling lattices. Also the suitability to apply motion compensation on the weighting functions was emphasized, achieving a significant improvement in the compression ratio and visual quality.

6 REFERENCES

[1] "The Laplacian Pyramid as a Compact Image Code", P.J. Burt, E.H. Adelson, IEEE Transactions on Communications, Vol. COM-31, nº4, April 1983, pp.532-540.

[2] "An Adaptive Pyramid Image Coding System", S.Sallent, L.Torres, Proceedings ICASSP 1988, New York, April 11-14, 1988.

[3] "Simulation Of A Teleconference Codec For ISDN", S.Sallent, A.Artero, J.Haro, EUSIPCO-90, Barcelona, September, 1990.

[4] "The Sampling And Reconstruction Of Time-Varying Imagery With Applications in Video Systems", E. Dubois, Proceedings IEEE 1988, Vol. 73, 502-522, April 1985.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Motion Compensated Prediction on Digital HDTV*

Fernando Jaureguizar, José I. Ronda, and Narciso García

Grupo de Tratamiento de Imágenes, E.T.S. Ingenieros Telecomunicación
Universidad Politécnica de Madrid, E-28040 Madrid, Spain

Motion compensated prediction techniques for image coding results very effective because of the large redundancy reduction in the temporal axis when applied to image sequences with limited movement between fields. Thus a hybrid motion-compensated prediction DCT coding scheme results very effective in coding digital HDTV signals. After presenting the tools for evaluating motion estimation, several algorithmic approaches have been analyzed on Digital TV and HDTV sequences, showing the feasibility of extending TV results to the HDTV field. Motion compensated prediction is very effective, giving better results for less-demanding quality applications. The two search areas currently considered have been evaluated showing that the increase in the search area increments the computation requirements while not significant increasing the efficiency.

1. Introduction

Conventional temporal prediction systems have low efficiency for moving pictures encoding. However, motion compensated prediction techniques for image coding results very effective because of the large redundancy reduction in the temporal axis when applied to image sequences with limited interfield movement. Within a Digital HDTV scheme, this temporal redundancy reduction is not enough due to the large amount of information. Besides, it is necessary to carry out a spatial redundancy reduction such a DCT transform.

Thus, a hybrid motion-compensated prediction DCT coding scheme results very effective for Digital HDTV signals encoding. So, international standardization efforts [1, 2] in Digital TV are being directed towards this hybrid DCT coding scheme.

Although the basic idea behind motion compensation is a very simple one (to find a prediction in the previous picture which best reduces the prediction error), the difficulty is to find the motion vectors of the various parts of a TV picture. There are several methods [3, 4] to estimate the motion vectors of the picture, however the differences between them are what is being matched and how the matching is carried out. The most used methods are the block-matching method, the differential method, and the phase-correlation one:

- In the block-matching method, a best match of a block of pixels is searched within the previous field or frame. The distance between the original block and its best match is the motion vector. Here it is the intensity levels of a group of pixels that is being matched. The matching is carried out using either an exhaustive method, search method, hierarchical method or another one. This methods uses a selection criterion (maximum correlation, m.s.e., m.a.e., ...).
- The differential method includes any method which makes use of the spatial and temporal differences of an image. The more used method is the pel-recursive one. The intensity level of individual pixels or a collection of them is what is being matched. The searching for a match is directed by local image gradient.
- In the phase-correlation method the 2-D DFT for two corresponding blocks in two successive images are evaluated. The peak of the inverse Fourier transform of the cross power spectrum phase indicates the velocity of the block. Conceptually, this method can be taken as a special implementation of the block-matching method with the search accomplished by convolution.

As the Digital HDTV signal holds a very high binary rate, the use of only one processor for the encoding implies an extremely high internal speed, not allowing the use of standard hardware technologies. One solution is the division of the computational burden among several processors. Each field is divided into several vertical bands, each one being independently processed by a complete Hybrid DCT codec [5]. Following the multiprocessor scheme, each individual processor can be formed by complete hybrid DCT Digital TV codec. So, the HDTV codec is built on the association of standard TV codecs in a parallel configuration.

This work has been done within Eureka-256: "Bit-Rate Reduction System for HDTV Digital Transmission". It has been partially supported by the Plan Electrónico e Informático Nacional and the Comisión Interministerial de Ciencia y Tecnología of the Spanish Government.

Since the aim of motion compensation is the bit rate reduction, one family of evaluation techniques must study the efficiency of the motion estimation methods to reduce this bit rate. Other family of evaluation techniques is oriented towards the measure of the difference between the real motion vectors and the estimate ones.

Some evaluation techniques for each motion estimation method are the bit rate reduction, statistics about the number of times the motion compensated prediction is chosen, the error in the estimation of the motion vectors relative to an exhaustive or "brute force" method.

2. Simulator and Sequences

A HDTV parallel codec simulator has been developed within Eureka-256 [5] to prove and study the motion estimation and compensation methods. This codec follows the Hybrid DCT algorithm proposed within CMTT [1, 2] considering the B2 code VLC option. The simulator maps a hardware architecture, thus performing all operations (including DCT) in fixed point. The field is divided into macroblocks (two luminance 8×8 blocks and the 8×8 blocks of the two corresponding chrominances), and an unique motion vector for each macroblock has been used [1].

Three TV clips and two HDTV clips, each 10 frames long, have been considered: *mobile and calendar*, *renata*, and *renata and butterfly* for TV, and *renata hdtv*, and *un bell di vedremo* for HDTV. The field dimensions are 288×720 for TV and 576×1440 for HDTV. The four first clips have been extracted from frames 0 up to 9, and the last from frames 50 up to 59 of the homonymous sequences. The first sequence was shot by the CCETT and the other four were shot by the RAI. The main characteristics are:

- In the clip *mobile and calendar* there are parts of the image (the calendar, the train, the background) showing different, well defined, nearly uniform, translating motions.
- In the clip *renata* a moving object shows a not uniform, not purely translating motion before a background showing a nearly pure translating but not uniform motion.
- In the clip *renata and butterfly* a moving object shows a not translating motion with a significant zooming effect.
- In the clip *renata hdtv* it shows mainly a significant zooming effect.
- In the clip *un bell di vedremo* several moving objects show not translating motions with a significant zooming effect.

3. Simulation Techniques

The used motion vector estimation method has been the block-matching one in the previous frame. This matching has been carried out using three techniques: the exhaustive, hierarchical, and the one_at_a_time_search (OTS) on two search areas [±7.5 lines × ±15.5 pels] and [±3.5 lines × ±7.5 pels] with 1/2 pixel precision . The exhaustive method estimates the motion vectors in every point of the search area. The hierarchical method consists of several stages of motion estimates in some predeterminated points of the search area. In our case it is estimated in one point for each four points [a subarea of (2×2) 1/2 pixels], for each nine points [a subarea of (3×3) 1/2 pixels], and for each eight points [a subarea of (2×4) 1/2 pixels]. The OTS method estimates the motion following the algorithm described in [6]. For every motion vector estimation method, the best match is found based on the minimum mean absolute error (MAE).

The motion vectors are estimated only for the two luminance blocks within a macroblock [7] as a single match, been the motion vectors for the two chrominance blocks derived from the one used for the luminance [1].

To evaluate the efficiency of the motion compensation, the resulting net video bit rate (regardless of the VLC of the motion vector differences) of each method is compared with the net video bit rate of a not motion compensated simulation. In order to achieve a better understanding of the results, the simulations have been performed at constant transmission factor (this implies a constant scaling factor).

No refresh has been applied. Nevertheless, the results are believed to be significant because of the shortness of the sequences.

4. Results on TV and HDTV

All computations have been carried out using the $[\pm 3.5, \pm 7.5]$ search area. After discussing the results, an analysis of TV extensibility is done.

4.1. TV

Within hierarchical methods, the [3_3] one has a good compromise between efficiency and speed (nine times speeder than the exhaustive method). For example, processing the clip *calendar* at a transmission factor f=60, which for this sequence implies a contribution quality (rounding 30Mb/s), the three studied hierarchical methods have an efficiency in reducing the bit rate of:

- [2_2] reduces the 26.38%
- [3_3] reduces the 26.01%
- [2_4] reduces the 25.88%

where the maximum reduction, achieved by the exhaustive, is the 26.89%, and reduction for the OTS method is the 25.88%.

Figures 1 and 2 show the buffer occupancy (at null output bit rate) versus the number of frames processed for the most relevant estimation methods, and the 3-D histogram of the estimated motion vectors, both at f=60 for the clip *calendar*. Figure 3 shows the buffer occupancy versus the f for the same clip. It can be seen that the higher the f (implying lower quality), the higher the motion compensation efficiency. The bit rate reduction with the exhaustive method at different f's is:

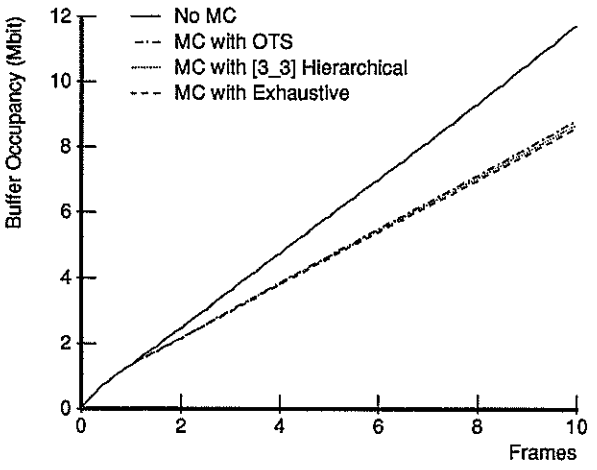


Figure 1.- Buffer Occupancy vs. Frames : Calendar at f=60.

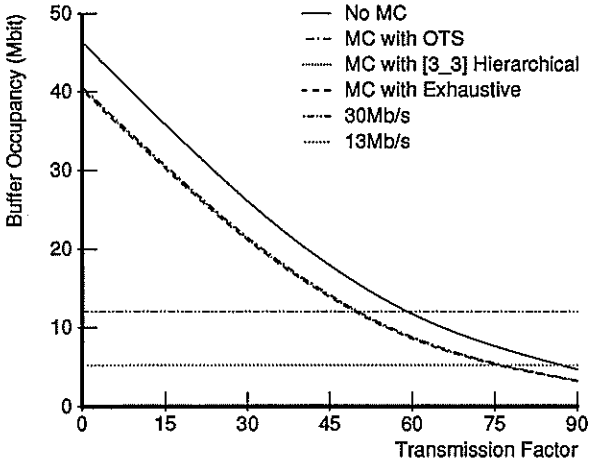


Figure 3.- Buffer Occupancy vs. Transmission Factor : Calendar.

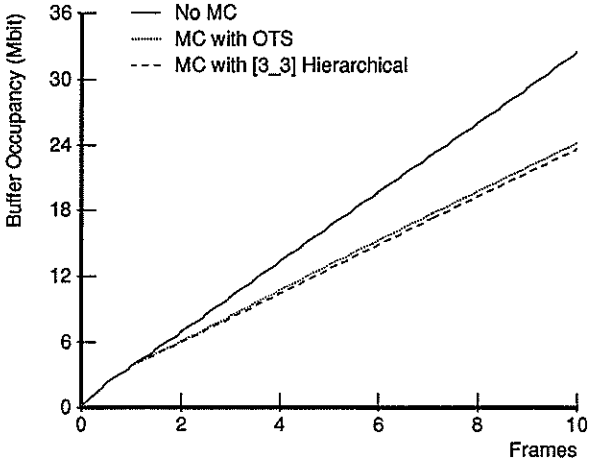


Figure 4.- Buffer Occupancy vs. Frames : Renata_hdtv at f=60.

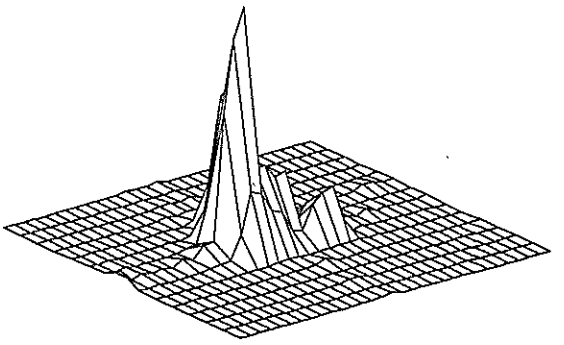


Figure 2.- Motion Vectors Histogram : Calendar.

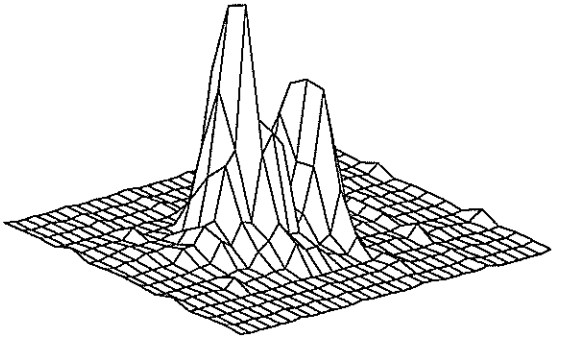


Figure 5.- Motion Vectors Histogram : Renata_hdtv.

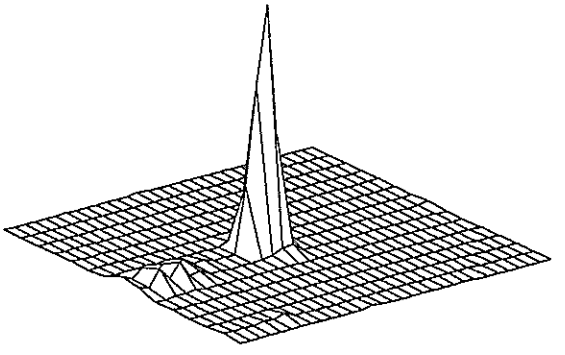


Figure 6.- Motion Vectors Histogram : Renata with [3.5, 7.5] Area.

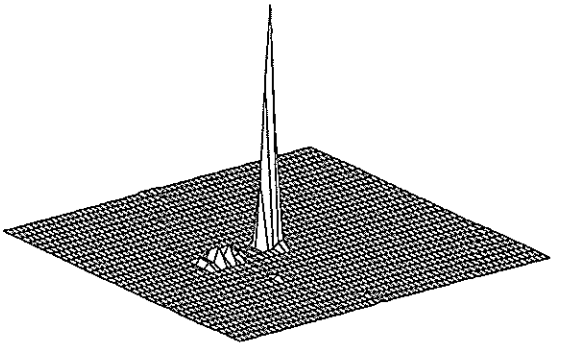


Figure 7.- Motion Vectors Histogram : Renata with [7.5, 15.5] Area.

- f=0 reduces the 13.23%
- f=30 reduces the 19.05%
- f=60 reduces the 26.89%
- f=90 reduces the 32.06%

For the three clips, the reduction at approximately the contribution quality is:

- *calendar* at f=60 reduces the 26.89%
- *renata* at f=50 reduces the 20.47%
- *renata and butterfly* at f=42 reduces the 19.28%

4.2. HDTV

The [3_3] hierarchical method is also a good compromise between efficiency and computational velocity for the studied sequences. Figure 4 shows the temporal evolution of the buffer occupancy for the clip *renata hdtv*, and for the most interesting estimation methods under study ([3_3] hierarchical and OTS) at a f=60, which approximately implies HDTV distribution quality (rounding 80Mb/s). Here, as in TV case, the exhaustive method is slightly more effective than the [3_3] hierarchical one. The motion compensation reduces the bit rate in the next amounts:

- 27.13% for [3_3] hierarchical
- 25.37% for OTS

Figure 5 shows the 3-D histogram for the [3_3] hierarchical method. It can be seen that for this clip, motion vectors are more widespread than in figure 2. For the clip *un bell di vedremo*, at approximately the same distribution quality the reduction for both estimation methods is:

- 10.69% for [3_3] hierarchical
- 8.56% for OTS

The vector histogram for this clip is very widespread out of the origin.

4.3. TV Extensibility

Viewing the results, it can be said that the motion compensation has similar behavior both in TV and HDTV. Therefore, it is computational time interesting to study motion compensation in HDTV using TV sequences.

5. Search Area Analysis

It had been studied the influence of the dimensions of the search area in the efficiency of the compression in the three TV sequences. The obtained results were very interesting. The bit rate reduction gained using a $[\pm 7.5, \pm 15.5]$ search area instead of a $[\pm 7.5, \pm 15.5]$ one was: 0.04% in *calendar*, 0.05% in *renata*, and 0.48% in *renata and butterfly*. The percentage of motion vectors of the bigger area which are lying inside the smaller one is: the 97.68% for *calendar*, the 91.40% for *renata*, and the 73.53% for *renata and butterfly*. Figures 6 and 7 show the 3-D histogram for the most favorable case, *renata*, at f=60 with the [3_3] hierarchical method and for the two search areas.

6. Conclusions

- Motion compensation is very effective in reducing the bit rate, or increasing the quality when maintaining the same bit rate.
- The lower the image quality, the higher the bit rate reduction with motion compensation.
- Due to the similar results in TV and HDTV, motion estimation and compensation in HDTV can be studied using TV sequences.
- The efficiency loss in the OTS method is compensated with his computational speed. It seems to be a good estimation method.
- Increasing the search area from $[\pm 3.5, \pm 7.5]$ to $[\pm 7.5, \pm 15.5]$, quadruples the computational time without significant bit rate reduction.

References

[1] "Draft New Recommendation: Transmission of Component-coded Digital Video Signals for Contribution-quality Applications at Third Hierarchical Level of CCITT, Recommendation G.702", Document CMTT/303, October 1989.

[2] "Proposed Modifications to Report AD/CMTT: Digital Transmission of Component-Coded Television Signals at 30–34 Mbit/s and 45 Mbit/s", Document CMTT/321, October 1989.

[3] H. G. Musmann, P. Pirsch and H.-J. Grallert, "Advances in Picture Coding", Proc. IEEE, vol. 73, no. 4, pp. 523-548, April 1985.

[4] S. F. Wu and J. Kittler, "2-D Motion Parameter Estimation: A Survey", Proc. 6th Scand. Conf. on Image Analysis, pp. 1043-1050, Oulu, June 1989.

[5] N. García, F. Jaureguizar, J. I. Ronda and A. Sanz, "HDTV Parallel Codec Simulator", Proc. Third International Workshop on HDTV, Torino, August 1989.

[6] R. Srinivasan and K. R. Rao, "Predictive Coding Based on Efficient Motion Estimation", IEEE Trans. Commun., vol. COM-33, no. 8, pp. 888-896, August 1985.

[7] K. A. Prabhu and A. N. Netravali, "Motion Compensated Component Color Coding", IEEE Trans. Commun., vol. COM-30, no. 12, pp. 2519-2527, December 1982.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

Backward Predictive Motion Compensated Image Sequence Coding *

J.N. Driessen, R.A.F. Belfor and J. Biemond

Delft University of Technology,
Department of Electrical Engineering, Information Theory Group,
P.O. Box 5031, 2600 GA Delft, The Netherlands.

Abstract

In this paper a motion compensated image sequence coding strategy is presented that does not require the transmission of any motion information. The strategy relies on motion detection and estimation on reconstructed data that is already known at the receiver. This backward predictive *strategy is applied in a pel-recursive DPCM-based coding environment, where a MAP-detector together with a Wiener-based estimator perform the motion detection/estimation. The prediction error is quantized using a MAP-detector for the regions that are predicted satisfactory and Max-Lloyd quantization of the remaining regions. The source coding is achieved by block-runlength and Huffman coding. Comparison with earlier proposed pel-recursive coding techniques shows the advantages of the proposed scheme especially for relatively low bit-rates.*

1 Introduction

The coding of image sequences is necessary for all applications where the capacity of the transmission channel or the storage medium is much lower than the bandwidth of the signal. Since spatial and temporal correlations between the pixel intensities are naturally present in image sequences, both intra- and interframe coding techniques are utilized. Often the interframe techniques rely on the compensation for object motion present in image sequences [1]. Prior to this compensation, the motion has to be estimated from the consecutive frames and, besides a motion estimation procedure, a motion detection procedure is involved. Unless stated otherwise, the term *motion estimation* is here used to indicate the total algorithm to determine the motion. If one or more original frames are involved in the estimation procedure, the estimated motion information needs to be transmitted. The need to transmit as efficiently as possible the prediction errors as well as the motion information consists of obviously conflicting goals and leads to a combined optimization problem. Moreover, the transmission of the motion information limits the data reduction achievable, since for a given sequence a lower bound on the bit-rate is determined by the bit-rate required for the transmission of the motion information.

It is the scope of this paper to present a predictive motion compensated coding strategy that does not require the transmission of any motion information. The tradeoff between such backward predictive schemes and forward schemes is discussed and the application op the proposed strategy in a pel-recursive coding environment is outlined. Two predictive pel-recursive motion compensated coding schemes have been proposed in literature [2,3]. It is shown that they are only partly based on the backward predictive principle.

*This research is supported by the Netherlands Technology Foundation (STW).

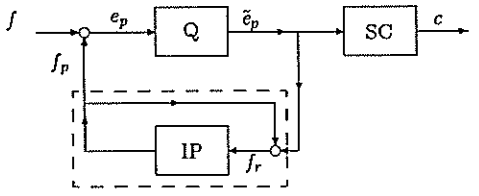


Figure 1: General Predictive Coding Scheme.

2 Motion Compensated Coding

In this section a discussion on the tradeoff between forward and backward predictive motion compensated coding schemes is presented. Although such a discussion is valid for fairly general coding schemes, it is based here on the image sequence coding scheme in Fig. 1. In this DPCM coding scheme, an intensity predictor (IP) forms an estimate f_p of the actual intensities f on the basis of previously reconstructed intensities f_r. The prediction error e_p is formed by subtracting the original and predicted intensities. The quantized (Q) prediction error $\tilde{e}_p$ is source coded (SC) to generate code words c with a bit-rate approaching the entropy of $\tilde{e}_p$. It is emphasized that this scheme together with the discussion following in this section is valid for both frame-, line- and pel-recursive algorithms.

2.1 Forward Predictive Schemes

If the temporal part of the intensity predictor compensates for motion between consecutive frames, the overall scheme has to contain a motion estimator. In Fig. 2, a common motion compensated prediction scheme is sketched. It consists of a motion compensated intensity predictor (MCIP) that operates along the estimated motion trajectories and a motion estimator (ME) that supplies the motion trajectories. This motion estimator operates on the original actual and the reconstructed previous

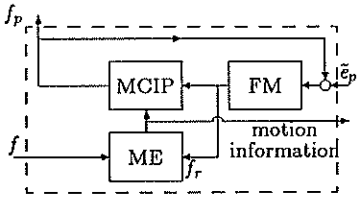


Figure 2: Forward predictive motion compensation.

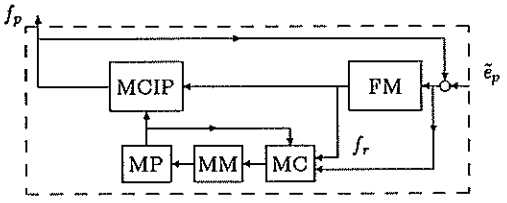


Figure 3: Backward predictive motion compensation.

frames. Since the receiver has not available the original frames, the motion information needs to be transmitted.

The advantage of this forward predictive scheme is that the estimated motion is accurate since it is estimated partly on the original frames. The disadvantage is that the motion information has to be transmitted, which causes an explicit lower bound on the bit-rate and a tradeoff in balancing the amount of prediction errors and motion information to be transmitted. As an illustration of this tradeoff it is assumed that the motion is represented by a so-called *motion field* which represents local displacements in the image plane. To provide an accurate intensity predition a high spatial resolution of the motion field is required to accurately capture discontinuities in the motion field due to differently moving objects. Additionally a fractional pixel accuracy of the vectors is required to account for non-integer displacements. These requirements imply a large amount of motion information. Common choices to represent the motion field are based on coarse resolutions and integer pixel accuracies which inevitably reduce the amount of motion information, but increase the prediction error.

2.2 Backward Predictive Schemes

A possible solution to such a problem is provided by the backward predictive scheme drawn in Fig. 3. The motion estimator is split into a motion predictor (MP) and a motion corrector (MC). The predictor forms an estimate of the motion trajectory based on previously estimated motion that is present in a motion memory (MM). The motion compensated intensity predictor (MCIP) forms an estimate of the actual intensities based on already reconstructed intensities and the predicted motion trajectory. The error between the original and the predicted intensities is quantized and after reconstruction of the intensities, a motion corrector calculates an update to the predicted motion. Since no original frames are involved in this motion estimation procedure, the receiver can perform the estimation.

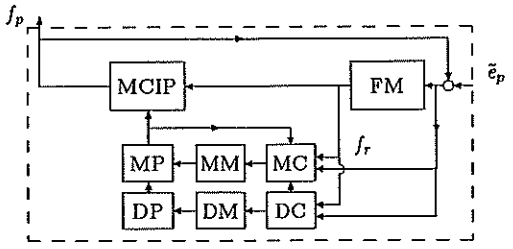


Figure 4: Backward predictive motion detector/estimator.

A disadvantage of such a scheme is the increase of the prediction error when compared with the forward predictive scheme since the backward predictive motion estimator is less accurate. This is due to the one-step prediction and the correction on the basis of reconstructed frames involved in the estimator. However, the main advantage of this backward predictive strategy is the absence of the need to transmit the motion information.

An obvious question now is which scheme will perform better: the forward or the backward predictive coder? Unfortunately, in general this question cannot be answered theoretically. It will depend on the amount of motion information to be transmitted in the forward schemes and on the increase in the intensity prediction error in the backward scheme. This increase, in turn, depends on the actual intensity prediction schemes, the motion prediction/correction scheme and the reconstruction quality or the bit-rate.

3 Application in a Pel-Recursive Coder

Schemes that are partly based on the backward predictive ideas are the pel-recursive coding schemes of [2,3]. In both coders, the total motion estimation algorithm consists of a temporal change detector and an estimator that is switched on in "changed" areas. The estimator operates according to the backward predictive principle, but the detector operates on original actual intensities. Therefore, both schemes require the transmission of control information for the estimator part. In the algorithm of Netravali and Robbins [2] the side information are the addresses of the moving regions and in the algorithm of Walker and Rao [3] the side information are resets to control the estimator. In this section, the backward prediction is also applied to the detector part.

3.1 The Motion Compensated Predictor

In Fig. 4 the new pel-recursive scheme is sketched. It is similar to the scheme in Fig. 3 except for the extension with the detector part. In the detection predictor (DP) an estimate of the detection result is formed based on previous results that are available in a memory (DM). This predicted detection is used to control the motion trajectory predictor (MP) which operates according to the description in the previous section. After the quantization of the intensity prediction error, the actual intensity is reconstructed and a correction of the detection is made based on these intensities. The corrected result is used to control the motion corrector (MC). Now the prediction and

correction equations are presented into more detail.

The detection result is represented by a binary mask $m(x,y)$ with value 1 if the spatial location (x,y) belongs to temporaly changed region and 0 otherwise. The prediction of the detection result $m_p(x,y)$ is based on a logical combination of previous detection results according to:

$$m_p(x,y) = m_c(x-1,y) \vee m_c(x,y-1), \quad (1)$$

with $m_c(.)$ the corrected detection mask. Thus the actual pixel is labeled as "changed" if the previous pixel or the previous line pixel has been labeled as "changed". The motion information is represented by a motion field $\vec{d}(x,y)$ that represents the local displacements between consecutive frames. The prediction of the displacement vectors, denoted by $\vec{d}_p(x,y)$, is based on a linear combination of previous displacement vectors controlled by the predicted detection:

$$\vec{d}_p(x,y) = \begin{cases} 0, & m_p(x,y) = 0 \\ \sum a(k,l)\vec{d}_c(x-k,y-l) & m_p(x,y) = 1 \end{cases}, \quad (2)$$

with $a(k,l)$ adaptive prediction model coefficients that depend on the binary detection mask $m_c(x,y)$. Here, the following prediction coefficients are used:

$$a(k,l) = \frac{m_c(x-k,y-l)}{\sum_{(i,j)\in S_m} m(x-i,y-j)}, \quad \forall (k,l) \in S_m, \quad (3)$$

with $S_m = \{(1,0),(0,1)\}$. The intensity prediction formula is based on the assumption that intensity changes between consecutive frames are entirely due to local displacements:

$$f_p(\vec{x},t) = f_r(\vec{x} - \vec{d}_p(\vec{x}), t - dt). \quad (4)$$

with $f_p(.)$ the predicted and $f_r(.)$ the reconstructed intensities. The detection corrector is based on a Markov Random Field model for the regions to be detected and is performed by a M(aximum) A P(osteriori) detector based on the Viterbi algorithm[5]. The correction or update of the predicted displacement in the changed regions is obtained via the Wiener-based approach of Biemond et al. [4]. The complete update formula depends on the detector as follows:

$$\vec{d}_c(\vec{x}) = \begin{cases} 0, & m_c(x,y) = 0 \\ \vec{d}_p(\vec{x}) - (G^tG + \mu I)^{-1}G^t\vec{z}, & m_c(x,y) = 1, \end{cases} \quad (5)$$

where G is a matrix containing gradient information within a support window located in the previous frame at displaced locations, $\vec{z}$ is a vector containing the displaced frame differences within the same support window and where μ is a so-called damping.

3.2 Description of the Remaining Coder Parts

The quantizer consits of a two-step procedure [6]. The intensity prediction error is expected to contain a large number of clustered "near zero" elements in regions where the motion estimator/detector and the intensity predictor operate accurately. A MAP detector [5] is used to extract these regions and the prediction error inside these regions is set to zero. The prediction error outside these regions are quantized by a Max-Lloyd quantizer.

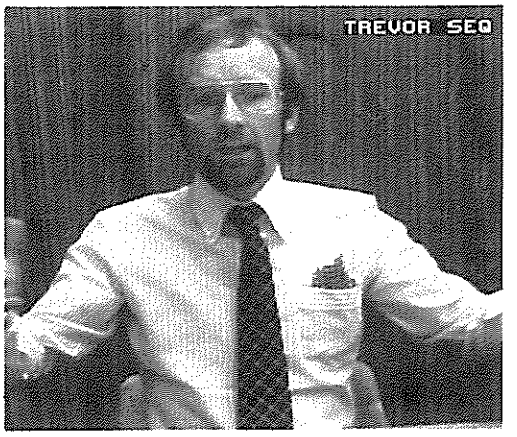


Figure 5: Original frame of the Trevor White sequence.

The source coding is performed by block run-length coding of the zeroes in the quantized prediction error with block-sizes of 8. This is an approximation to optimal run-length coding that could result in very large, and therefore impracticle, code books. The resulting symbols and the Max-Lloyd quantized values are coded via a Huffman coding technique [6].

4 Experimental Results

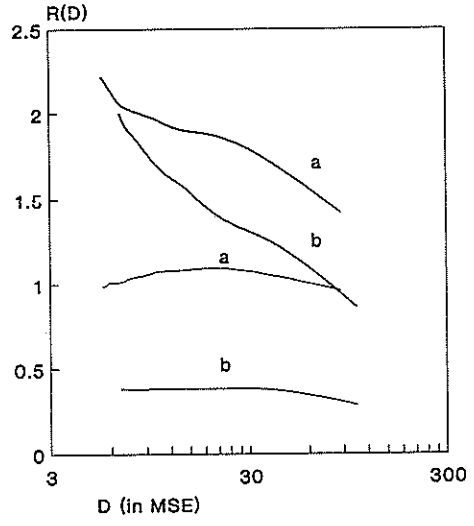
The experiments presented in this section were performed on the typical low bit-rate image sequence "Trevor White" shown in Fig. 5. This sequence contains the upper part of the body of a talking and heavily gesticulating person.

The first experimental result is the improvement of the coding efficiency of the codec of Netravali and Robbins [2] by using a forward predictive MAP-based change detector together with a Wiener-based motion estimator. In Fig. 6, the results are shown as rate-distortion curves, where the thick lines represent the total bit-rate of the codec output and the thin lines represent the bit-rate required for the detection mask. The overall improvement in efficiency can be seen to be due mainly to the improvement in the detection mask. For low bit-rates, however, the transmission of this mask requires a still large bit-rate.

Secondly, the proposed algorithm is compared with the improved existing algorithms. In Fig. 7 the rate-distortion curve is shown for the three algorithms. The new scheme is better than the other schemes for almost al coding rates. However for very high coding rates, the efficiency of the new scheme becomes worse than the efficiency of the schemes of Walker and Rao. In Fig. 8 a reconstructed version of a frame of the Trevor sequence is shown produced by the proposed algorithm. The mean bit-rate for the sequence was 0.7 bit/pixel and some visual artifacts can be noticed. These artifacts are less visible when the sequence of reconstructed frames is observed.

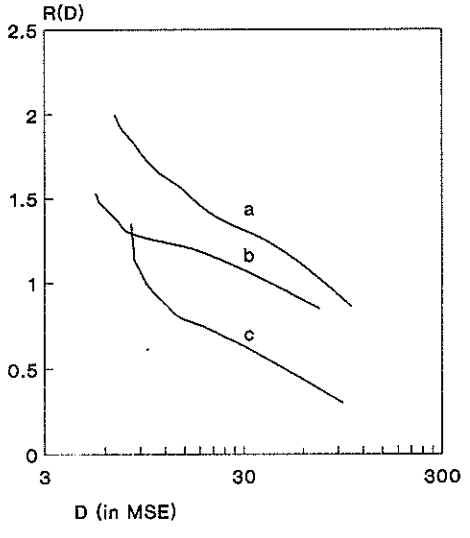
5 Discussion and Future Research

In this paper a backward predictive motion compensated image sequence coding strategy has been proposed as opposed to the existing forward predictive strategies. The main advantage is



a. Netravali and Robbins
b. Improved version

Figure 6: Improvement of the coder of Netravali and Robbins.



a. Netravali and Robbins
b. Walker and Rao
c. proposed algorithm

Figure 7: Comparison of the improved and the new scheme.

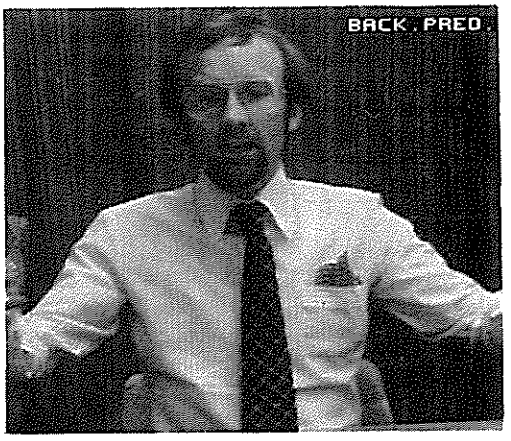


Figure 8: Reconstructed frame via backward predictive coding algorithm.

the absence of the need to transmit motion information. A minor disadvantage is the increase in the prediction error. The ideas are applied within a pel-recursive scheme which is shown to be superior to existing pel-recursive schemes for relatively low bit-rates.

Future research will focuss on adaptive quantization techniques, the inclusion of intraframe relations and the extension of the coding scheme to color sequences.

References

[1] Musmann, H.G., P. Pirsch and H.-J., Grallert, "Advances in Picture Coding", *Proceedings of the IEEE*, vol. 73, no. 4, April 1985, pp. 523-548.

[2] Netravali, A.N. and J.D. Robbins, "Motion-Compensated Television Coding: Part I", *Bell System Technical Journal*, BSTJ-58, no. 3, March 1979, pp. 631-670.

[3] Walker, D.R. and K.R. Rao, "Motion-Compensated Coder", *IEEE Transactions on Communications*, vol. COM-35, no. 11, November 1987, pp. 1171-1178.

[4] Biemond, J., L. Looijenga, D.E. Boekee and R.H.J.M. Plompen, "A Pel-Recursive Wiener-based Displacement Estimation Algorithm", *Signal Processing*, vol. 13, December 1987, pp. 399-412.

[5] Driessen, J.N., J. Biemond, D.E. Boekee, "A Pel-Recursive Segmentation and Estimation Algorithm for Motion Compensated Image Sequence Coding", in: *Proceedings ICASSP '89*, Glasgow, Scotland, May 23-26, 1989, pp. 1901-1904.

[6] Belfor, R.A.F., *Predictive Coding of Image Sequences using Pixel-Recursive Motion Compensation*, MS thesis, Delft University of Technology, Dept. of EE, Information Theory Group, The Netherlands (in dutch).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A study of a hybrid image sequence coder employing advanced motion compensation

John Håkon Husøy
Tor A. Ramstad
The Norwegian Institute of Technology
Department of Electrical and Computer Engineering
N-7034 Trondheim-NTH
NORWAY

ABSTRACT

For low bit rate image sequence coding, the currently most popular approach is a hybrid scheme incorporating motion compensated prediction and a Discrete Cosine Transform (DCT). The motion estimation is commonly performed using a block matching technique. A question that has received very little attention in the image coding community is: *Can we obtain better performance in a sequence coder by using a more sophisticated motion estimation technique?* In this paper we present a preliminary investigation on this topic. The algorithm studied is an extension of the one described by Horn and Schunck. By presenting several performance measures we show that this approach has a potential for improved performance in motion compensated coders.

1 Introduction

For low bit rate image sequence coding, the currently most popular approach is a hybrid scheme incorporating motion compensated prediction and a Discrete Cosine Transform (DCT). The motion estimation is commonly performed using a block matching scheme [1]. This is e.g. the approach adopted by CCITT in reference model 7 [2] that is applicable to the coding of teleconferencing scenes at bit rates of $64p$ kbps where $p = 1, \ldots, 32$.

In this paper we propose a profound departure from the traditional way of doing motion estimation in a hybrid coder. Rather than employing block matching we employ an extension of Horn and Schunck's algorithm [3]. This algorithm gives rise to a smooth displacement vector field with one estimated displacement vector for each pel location in the picture. The rationale for doing this is as follows: In hybrid coders based on block matching, we obtain motion vectors that indeed contribute significantly to the decrease of the RMS value of the displaced frame difference (motion compensated prediction error) as compared to the RMS value of the non-motion compensated frame difference. Typically one obtains one motion estimate for each image block of size 8x8 pels. These motion vectors (displacement vectors) are encoded *losslessly* and transmitted to the decoder as side information. Although this approach has been used with success in hybrid coders, it is also known that straight forward block matching has inferior performance if the prime objective is the estimation of the *true* motion in a scene. Therefore, in the computer vision community much attention has been directed to methods trying to estimate the true motion in a scene. Thus one might ask: *Is there more to be gained in motion compensated hybrid coders by employing more advanced motion estimation techniques where emphasis is put on estimating the true motion in the scene?* Here we present a preliminary investigation into this question.

As an introductory example, we show in Fig. 1 and Fig. 2 displacement fields for a portion of the face in the CCITT sequence CLAIR. In Fig. 1 is the displacement field obtained by block matching and in Fig. 2 is the result of applying the proposed extended Horn and Schunck algorithm. Note that the latter motion vector field is characterized by its *smoothness and by its intuitive plausibility.* This is in contrast with the motion field of Fig. 1. Also note that in the latter case certain estimates are obviously wrong although they may still contribute to a reduced displaced frame difference.

By utilizing this more accurate displacement estimate in a motion compensated predictive coder we therefore expect smaller prediction errors which in turn makes additional coding gains possible. The price for this is increased computational complexity. Also, as in schemes based on block matching, the displacement estimates have to be transmitted to the decoder as *side*

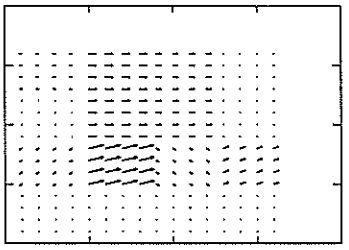


Figure 1: Displacement field obtained by block matching.

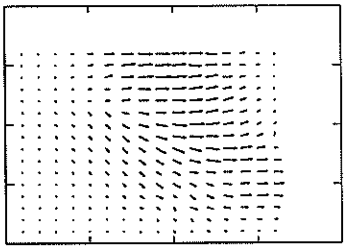


Figure 2: Displacement field obtained by the extended Horn and Schunck algorithm.

information . Since in our scheme we have one displacement vector for each pel location rather than one vector for each block, – we would expect to spend more bits on the representation of the displacement field. We should keep in mind, however, that the displacement vector field is very smooth and as such lends itself to data compression (not necessarily lossless!). Results on the bit efficient representation of such displacement fields in the context of motion compensated interpolation of missing frames were recently reported by Hung [4]. We feel that this should also be possible in our context although for the purposes of this preliminary study we assume that an exact representation of the estimated motion field is available.

2 Coder Structure

In Fig. 3 we show the overall structure of our coder. The displacement estimates are used by the predictor so that the motion compensated prediction errors can be found. The prediction error signal is transformed by a DCT. The coefficients are uniformly quantized and run length coded before being transmitted to the decoder.

With the exception of the motion estimation procedure, the major blocks in our coder structure follows closely the specifications as set out by CCITT's ref. model 7 [2].

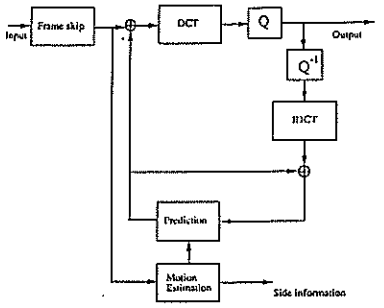


Figure 3: Coder structure

3 Displacement estimation

In this section we will first present Horn and Schunck's original algorithm [3] for the computation of the displacement field. We will then describe some modifications for making the algorithm more suitable in situations where there may be substantial motion from one image frame to the next.

3.1 Horn and Schunck's algorithm

Based on the assumption that the brightness of an object point is constant along its direction of motion, Horn and Schunck arrive at the following recursive algorithm for the computation of the displacement field [3]:

$$d_x^{n+1} = \overline{d_x^n} - \frac{E_x^2\overline{d_x^n} + E_xE_y\overline{d_y^n} + E_xE_t}{\alpha^2 + E_x^2 + E_y^2} \quad (1)$$

$$d_y^{n+1} = \overline{d_y^n} - \frac{E_yE_x\overline{d_x^n} + E_y^2\overline{d_y^n} + +E_yE_t}{\alpha^2 + E_x^2 + E_y^2} \quad (2)$$

where E_x, E_y and E_t are the finite differences used for estimating partial derivatives of the intensities with respect to spatial and temporal coordinates. d^{n+1} denotes the displacement vector at iteration number $n+1$. $\overline{d^n}$ is a local average over some neighborhood of the displacement at iteration n and α is a constant. The iteration is performed for every pel in the image until some convergence criterion is satisfied.

3.2 Improved displacement estimation

It is well known that Horn and Schunck's algorithm does not work very well when there is large motion in the scene ($\sim$ 10 pels). The reason for this is twofold. Firstly, a straight forward estimation of the temporal derivative (i.e. by using frame differences) tends to give large errors when the displacement is large. Secondly, the spatial gradient estimates, computed based on pel

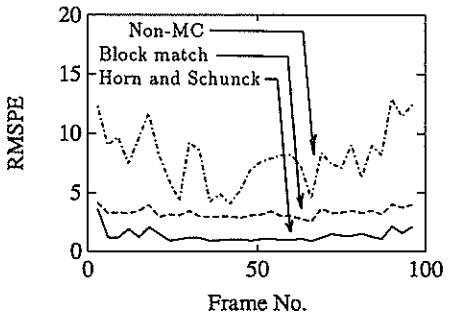


Figure 4: RMS prediction error for CLAIR, 10 frames per second.

values from a small neighborhood, are not always reliable. The first problem has prompted the usage displaced frame differences rather than straight frame differences in estimating the temporal derivative in the recursive equations. Also, displaced samples are used in computing the spatial gradient estimates. The second problem has to some extent been solved by the introduction of a hierarchical multi resolution technique which can be explained as follows. On a coarse level (lowpass filtered image) a rough estimate of the motion of the major features of the scene is obtained. This coarse estimate is used as an initial estimate on a finer level (lower degree of lowpass filtering). This successive refinement is repeated until the full resolution estimate is obtained. In our simulation we employ a strategy similar to the one described by Konrad [5].

4 Experimental results

We have implemented a software simulation of the hybrid coder employing an improved Horn and Schunck algorithm for the displacement estimation. Here we show some figures indicative of its performance.

The RMS prediction error has been used extensively by other researchers as a measure of the quality of motion compensated predictive schemes. We show, in Fig. 4, the RMS prediction errors associated with schemes in which there is no motion compensation, motion compensation using block matching with full search and motion compensation using the improved Horn and Schunck algorithm. The curves are obtained from the well known CCITT sequence commonly referred to as CLAIR. The frame rate is 10 frames per second. Also in Fig. 5 we show the same quantity for the sequence MISS AMERICA. We note that in both cases there is a distinct improvement, – on the average, in RMS prediction error to be gained by employing the modified Horn and Schunck algorithm as opposed to the block matching scheme.

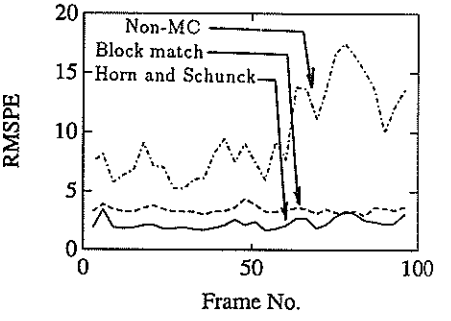


Figure 5: RMS prediction error for MISS AMERICA, 10 frames per second.

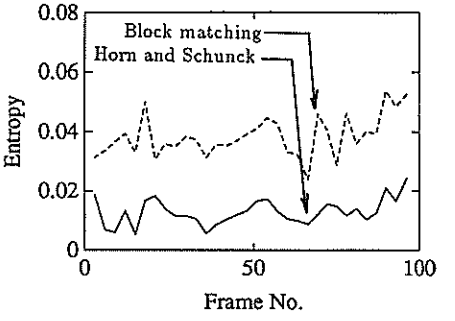


Figure 6: Entropy rate for CLAIR

The bit rate of our coding scheme is estimated by calculating the entropy per pel resulting from the transform coefficients that have to be transmitted. We selected a fixed uniform quantizer with stepsize $g = 32$ leading to a variable bit rate coder. In Fig. 6 and Fig. 7 we show the entropy per pel for each frame of the previously mentioned sequences. We note a pronounced reduction in bit rate by using the proposed approach.

The subjective and objective quality of the sequences when coded employing our proposed motion estimation scheme was considerably better than what we obtained when block matching was used. This is illustrated by the SNR curves in Fig. 8 and Fig. 9. Finally, Fig. 10 shows the superior subjective quality obtained.

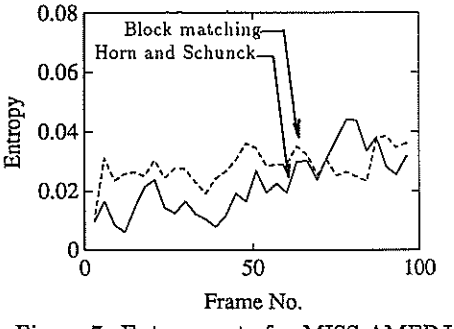


Figure 7: Entropy rate for MISS AMERICA

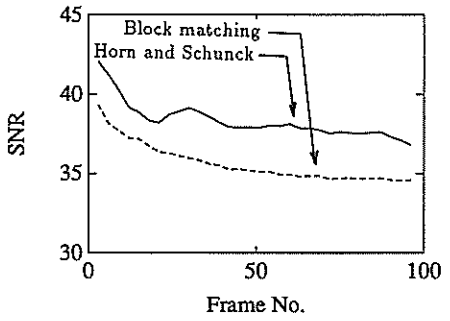


Figure 8: Peak SNR for CLAIR

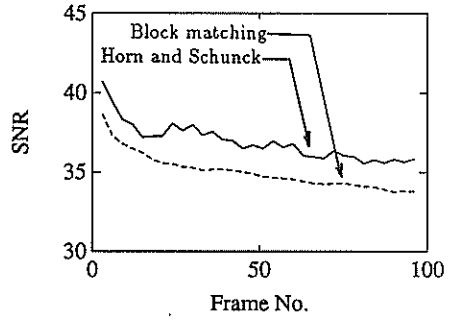


Figure 9: Peak SNR for MISS AMERICA

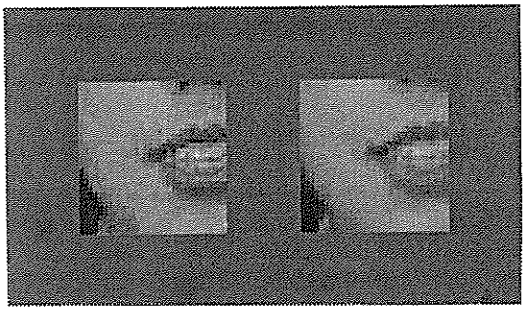


Figure 10: Left: Detail of decoded CLAIR employing block matching. Right: Detail of decoded CLAIR employing the proposed algorithm.

5 Conclusion

In this paper we have discussed the application of an enhanced version of Horn and Schunck's algorithm in a motion compensated hybrid coder. We have illustrated its salient properties in this context. Research on the representation of the motion field is planned for the near future and will be reported in a subsequent paper.

Acknowledgement The first author would like to thank Janusz Konrad of INRS-Telecommunications at The University of Quebec, Montreal for sharing his knowledge on motion estimation techniques. This work was supported by The Norwegian Council for Technological and Scientific Research (NTNF).

References

[1] J. R. Jain and A. K. Jain, "Displacement measurement and its application in interframe image coding," *IEEE Trans. Commun.*, vol. COM-29, pp. 1799–1808, Dec. 1981.

[2] CCITT Study Group XV, Working Party XV/4, "Description of ref. model 7 (rm7), document no. 446," Jan. 21 1989.

[3] B. Horn and B. Schunk, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

[4] H. Q. Nguyen, "Représentation des champs de déplacement pour le codage numériqe des séquences vidéo," Master's thesis, Université du Québec, INRS-Télécommunications, 1989.

[5] J. Konrad, "Motion-compensated interpolation for tv frame rate conversion," Tech. Rep. 88-26, INRS-Télécommunications, Université du Québec, Oct. 1988.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

REGION-ORIENTED CODING OF MOVING VIDEO – MOTION COMPENSATION BY SEGMENT MATCHING

Wolfgang Guse, Michael Gilge, Christoph Stiller

Institute for Communication Engineering, Technical University Aachen
Melatener Str. 23, 5100 Aachen, West Germany
Tel.: +49-241-807680, Fax: +49-241-804413

Abstract

Starting from a given image segmentation the result of motion estimation is used to modify this image segmentation with the aim to reach a segmentation in which each segment consist of an entire object. The term 'object' is defined as an image part described with the same motion parameters. The partition is changed according to the estimated motion parameters and the transmitted prediction error image yielding an actual segmentation of each frame of the scene into background and differently moving objects.

Introduction

Coding moving video at low data rates can only be achieved if not only spatial but also temporal correlation is taken into account. Additionally, prediction can be improved if motion compensation is used. A survey of motion estimation techniques can be found in [1]. As the twodimensional image sequence is a projection of the threedimensional real world scene, rigid motion can be approximated by translational and rotational motion of rigid objects [2]. Any residual errors are due to shape changes and uncovering of background or occlusion of objects. Matching techniques have to be restricted to simple motion models e.g. translational motion, because the computational effort is proportional to the amount of possible matchpositions (the extension to rotations would enlarge the necessary computational requirements extremly [3]). But even with this basic model of translational motion of planar objects, matching algorithms cannot completely succeed in estimating the motion if the image content is disregarded. Rectangular blocks as in block matching may contain parts of differently moving objects and nevertheless a single motion vector is assigned. If we expand the state-of-the-art block matching motion estimation algorithm to arbitrarily shaped segments the goal seems to be within reach. There are mainly three possible types of segments:

- **A:** Segments representing an entire object are matched exactly (within the assumed model, e.g. translational movements).
- **B:** Segments containing only a part of an object have to be merged with adjacent segments.
- **C:** Segments containing parts of different moving objects have to be subdivided into several parts, which are matched independently from each other.

For each segment one vector is calculated. Because the segments may include more pixels as compared to blocks in a block matching algorithm, e.g. one background segment, less motion vectors have to be transmitted thus saving data rate. In the first part of the paper it is shown, that a binary object mask is suitable to improve the standard block matcher. In a next step the matching algorithm is carried out with segments instead of blocks thus trying to reach a segmentation which is identical to the contours of the different moving objcts in the scene. To reach this aim a mean to prove the quality of motion estimation is introduced. Algorithms for merging and splitting segments are necessary in order to adapt the segmentation to the actual frame and to improve motion estimation. The last part is dedicated to the processing of the prediction error image and its effect on the segmentation.

Improved Block Matching Using an Object Mask

There have been several attempts to solve the problems arising from the interdependencies between object detection and motion estimation. A basic approach can be made with the assumption, that moving objects in a scene are represented by those areas with high amplitude in the difference-picture between successive frames. A change detector applied on this difference-picture and filtering generates a binary object mask [4]. This object mask can only distinguish between stationary background and moving object(s) in the scene, but cannot recognize two different objects. This is sufficient for most

of the scenes in the field of low bit-rate coding for instance. With the help of this object mask block matching algorithms can be improved significantly, because all blocks containing parts of moving objects as well as stationary background can be marked according to the object mask. The match criterion is calculated only for those pixels in the block, belonging to the object mask thus yielding a reliable vector for this block, which is not biased by stationary background. Additionally the object mask which is known at the transmitters and the receivers side is used to improve the prediction image. The implementation of background memory becomes an efficient tool with the help of the object mask, because the signalling information whether a part of the new frame may be copied from the background memory or not is valid for every single pixel without transmitting a single bit. All pixel outside the object mask are copied from the background memory. Nevertheless motion estimation and compensation is still based on parts of rectangular blocks.

Motion Estimation Using Image Segments

The first and easiest approach to motion estimation is a modification of a full search block-matching algorithm. The algorithm is carried out with segments instead of rectangular blocks. In order to avoid the transmission of contour information the first segmentation must be known at the receivers side as well, thus leaving two reasonable opportunities: The whole frame is one segment or the whole frame is subdivided into a regular segmentation e.g. triangles or rectangular blocks. In contrast to block-matching algorithm the segmentation of the new frame is unknown. The block oriented estimater approximates the new frame with that information from the old frame (backward) which yields the best prediction image and for every block in the new frame such information is certainly found. The segment oriented estimater matches forward and is not able to cope with the problems of new segments or uncovered background. Significant advantages are the fact that the motion parameters are exact and uncovered background is precisely detected.

Because the image sequence is only a two dimensional projection of a three dimensional scene some important information about the location of the objects relativ to each other is lost. This becomes obvious if one object moves in front of another one, thus occluding parts of that object. This effect cannot be described by motion vectors nor by deformation parameters. In order to solve this problem we made the following assumption: The moving object in the foreground yields the best match. It is compensated first and the other segments follow according to the quality of their estimated motion vectors.

Quality-Measure for Motion Vectors

The function of the match criterion over the search positions is the only information available to decide if a motion vector is reliable or not.

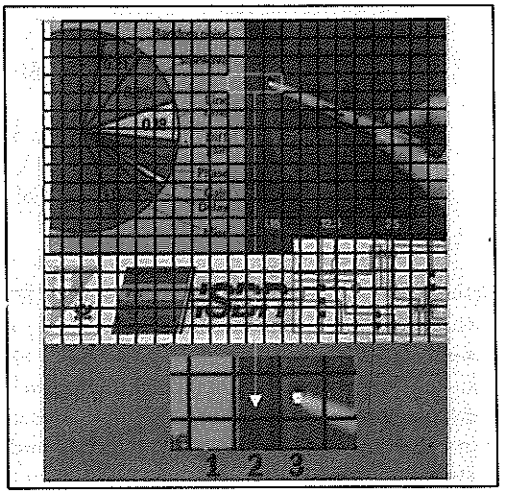


Fig. 1 : Block segmentation with 3 sample blocks marked

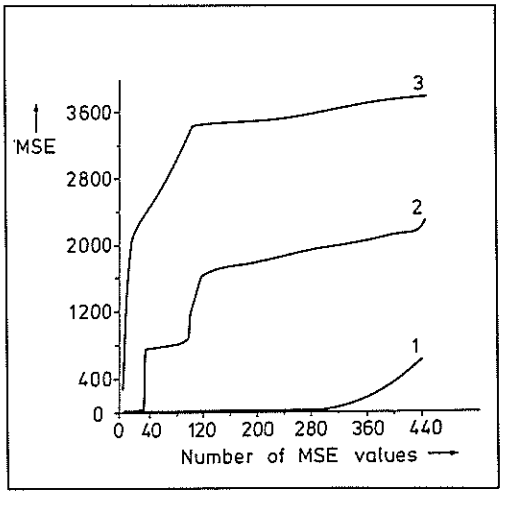


Fig. 2 : MSE-functions for the marked blocks from Fig. 1

Fig. 2 shows this function for three adjacent blocks in a sample frame of the scene *'SWING'*. The values of the MSE are sorted according to their amplitudes. Block 3 has a sharp minimum and can be matched exactly. Block 2 contains an edge which may be moved in vertical direction without any significant error. Block 1 contains no structure, almost any position yields a good minimum. The area below the MSE curves is regarded as a measure for the quality of the motion estimation. MSE values below a certain noise adaptiv treshold are neglec-

ted, because they are due to camera noise. The area is normalized using the standard area from an optimal match of one white point in a black area.

$$QM = \frac{1}{N \cdot (255^2 - noise)} \cdot \sum_{i=1}^{N} (MSE_i - AbsMin)$$

AbsMin is the smallest MSE or if this is below the camera noise level the noise level itself. N is the number of MSE values taken into consideration.

Merging and Splitting of Segments

Those segments which are not definitely predicted i.e. the MSE function shows no sharp minimum, are taken into consideration for being merged with adjacent segments. The sequence of merging is cannot be predetermined. The neighbouring segment with the most likely mean grey value is examined first. This assumption holds, because uncertain segments are unstructured and show a high DC coefficient. Merging is controlled with the help of the MSE functions. The minimum of the MSE function for the new segment consisting of the uncertain segment and its neighbour is determined. If the MSE values for the single segments at the position of this minimum does not differ significantly from the MSE minima of the single segments, the merging is regarded as being correct. To evaluate the merging of the segments i and j to the segment i+j a cost factor K is introduced

$$K = \frac{MSE_i(x_0, y_0)}{MinMSE_i} + \frac{MSE_j(x_0, y_0)}{MinMSE_j}$$

with (x_0, y_0) the position of the MSE function of new segment. The smallest K=2 holds if the postion of the best match is identical for each single segment and for their combination. If the merge results in a segment having a bad MSE the composed segment is splitted up again and another neighbouring segment is examined. The merge algorithm stops if either no uncertain segments are left or all adjacent segments have been inspected. If the motion of a segment cannot be described by the estimated parameter a splitting of the segment may be necessary. Two cases can be distinguished: First, the motion parameters are estimated using the given segmentation. In this case those parts of the segment with great prediction error must be separated from the segment and the motion estimation must be carried out again. This iterative procedure requires a great amount of computation. Second, the motion parameters are estimated indepently from the segmentation. In this case the motion parameters themselfs may be used to split the segment. The information which segments have been merged and the dividing line of splitted segments has to be transmitted. In our simulations we left the splitting of segments to the update of the prediction error image. We estimate the motion parameter using the segmentation and avoid the transmission of the dividing line be just using one iteration step i.e. the segment part with great prediction error is updated after motion compensation.

Segmentation of Update Areas

Motion compensation is carried out using the modified segmentation and the motion parameters. The prediction error image is segmented and filtered as described in [6]. A threshold operation with iterativly changed threshold divides the prediction error image into regions with significant error amplitude and regions with neglectable error. Small regions are eliminated i.e. the are converted to the state of their surrounding. The resulting update regions are coded with a generalized orthogonal transform [5]. Basis-functions are orthogonolized with respect to the form of the individual update segment, the transform coefficients are determined, quantized and transmitted. These update segments are integrated into the segmentation which has been used for motion compensation. We seperate three types of reasons for update segments:

- New information in the scene e.g. new objects or uncovered background which could not be copied from the background memory.
- The motion estimater failed in calculating the correct motion parameter.
- The used segmentation was incorrect and did not describe an entire object.

Only in the first case a new segment is necessary. If the motion parameter or the segmentation apears to be wrong the update segment must be used to modify the old segmentation i.e. the pixels should be distributed among the surrounding segments. With the help of a grey level segmentation the update segments are subdivided into homogenous parts. These small segments are merged with an adjacent segment which has the best fitting mean grey value.

Simulation Results and Conclusion

The figures show the segmentation results for the testsequences *'SWING'* and *'SALESMAN'*. Please note that no contour information has been transmitted, because in this simulation all segmentation results were based on information already available to the receiver and no splitting of segments has been carried out. Therefore, improvements over block-oriented schemes are accomplished without additional overhead information.

The number of segments used varies from 6 to 15 for the sequence *'SWING'* and from 80 to 120 for *'SALESMAN'*. Even if only translational motion is regarded the results of segment oriented motion estimation and compensation in connection with a region oriented update seem very promising. Only in the case of segment splitting (if it is done during motion estimation) and for update purposes contour information must be transmitted. The next step of investigation is an enhanced motion estimation which is able to cover the problems of rotation and deformation.

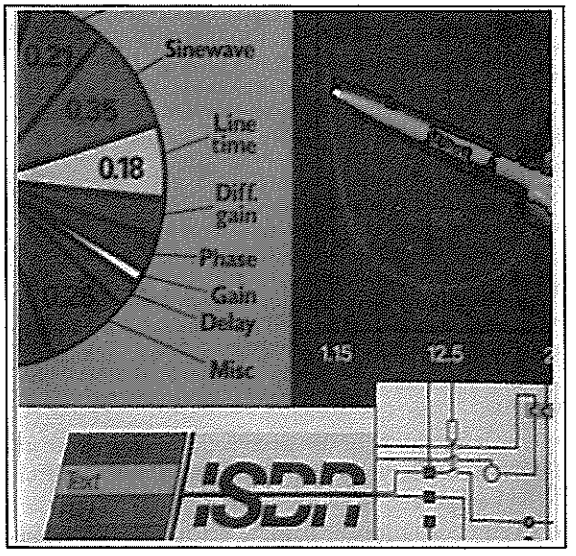
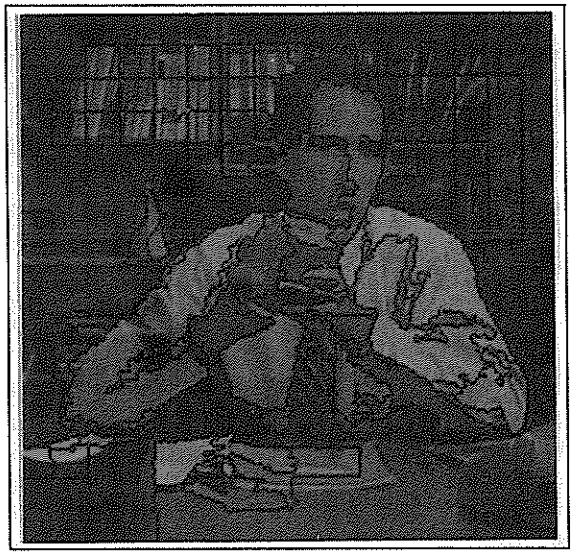


Fig. 3 : Segmentation result after 5 coded frames, sequence *'SWING'*



References

[1] H.G. Musmann, P. Pirsch, H. Grallert: "Advances in Picture Coding." Proceedings of the IEEE, VOL. 73, NO. 4, April 1985 pp. 523-548

[2] M. Hoetter, R. Thoma: "Image Segmentation Based On Object Mapping Parameter Estimation." Signal Processing 15 (1988) North-Holland pp. 315-334

[3] S.N. Jayaramamurthy, R. Jain: "An Aproach to the Segmentation of Textured Dynamic Scenes." Computer Vision, Graphics and Image Processing Vol 21, 1983 pp. 239-261

[4] C. Stiller, M. Gilge, W. Guse, : "Region-Oriented Coding of Moving Video – Compatible Quality Improvement by Object Mask Generation." Proc. 5th Europian Signal Processing Conf. Eusipco 90, Barcelona Spain, Sept. 90 this Volume

[5] M. Gilge, T.Engelhardt, R.Mehlan :"Coding of Arbitrarily Shaped Image Segments Based on Generalized Orthorgonal Transform." Signal Processing: Image Communication vol. 1, Oct. 1989, pp. 153-180, Elsevier Sience Publishers

[6] M. Gilge, B.Hürtgen: "Segmentation of Prediction Error Images." 2nd International Workshop on 64 kbit/s Coding of Moving Video. Hannover, West Germany, Sept. 1989

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Sequence Coding By Gabor Decomposition

Touradj Ebrahimi Todd R. Reed Murat Kunt

Laboratoire de Traitement des signaux
EPFL-Ecublens (DE)
CH-1015 Lausanne
Switzerland

A data compression technique based on a pyramidal decomposition is discussed. An efficient method for calculating the decomposition is also introduced. The decorrelating properties of the pyramidal and classical Gabor decompositions are compared. Simulation results show that 256x256 pixel image sequences can be transmitted at a bit rate of less than 64 Kbit/s using this method, with good image quality and without temporal subsampling .

1. Introduction:

Transform based coding methods are very popular in data compression because of the good quality of the results obtained, for a given compression ratio, compared to other methods. Among the existing transform based methods, those based on the discrete cosine transform (DCT) are the most widely used, both because of the decorrelative property of the DCT and the existence of fast algorithms for computing the DCT, implemented in VLSI. However, DCT coding methods suffer from the fact that the DCT is not local in the spatial domain. To take advantage of the local spatial correlations in images, the DCT must then be calculated over a number of separate regions, which leads to blocking effects for very low bit rate transmission.
In this paper a compression method based on a Gabor decomposition is developed. The use of such a decomposition is motivated by the fact that Gabor functions have optimal localisation in both spatial and frequency domains [1]. In addition, according to recent experiments, the majority of the receptive field profiles of the mammalian visual system can be fit quite well by this type of function [2].
Another reason to use Gabor functions is that it has been shown that the Gabor decomposition reduces the entropy of the data. This means that, like the DCT, the Gabor transform has the beneficial property of decorrelation.
In this paper a very fast and easily implemented method is proposed to compute the weighting factors of the decomposition. This method is based on the least mean squares error fitting criterion (LMS). The solution to the LMS problem shows that the weighting factors can be extracted by a simple multiplication between a matrix and the vector of data. If the set of Gabor functions is chosen independent of the image, the multiplicative matrix is constant. The reconstructed data is given by another multiplication between the matrix of Gabor functions and the vector of weighting factors. Our algorithm is designed in such a way that the weighting factors of the decomposition can be found for any set of basis functions, not just the Gabor functions.
The choice of elementary functions in the decomposition is very important. In our current system, the elementary functions are chosen so that the resulting decomposition has a pyramidal structure. The pyramidal structure is motivated both by observations of the early visual system [3] and the fact that it lends itself easily to progressive transmission.
The coding of the image sequence is done by decomposing each individual image and by transmitting the differences between the coefficients of successive frames, for significantly varying coefficients.

2. The Gabor and the pyramidal Gabor functions:

In the 1D case, a Gabor function can be expressed as:

$$g_{m:n}(t) = \hat{g}_D(t-mD) \cdot e^{jnWt}$$

where

$$\hat{g}_D(t) = e^{-\pi (t/D)^2}$$

is a gaussian function, D determines the scale of the gaussian in the time domain, and $W = 2\pi/D$ is the scale in the frequency domain. The Fourier transform of a Gabor function can be expressed as :

$$G_{m:n}(\omega) = \frac{\sqrt{2\pi}}{W} \hat{g}_W(\omega - nW) \cdot e^{-jmD(\omega - nW)}$$

The real parts of some typical Gabor functions as well as their power spectra are shown in Figure 1.

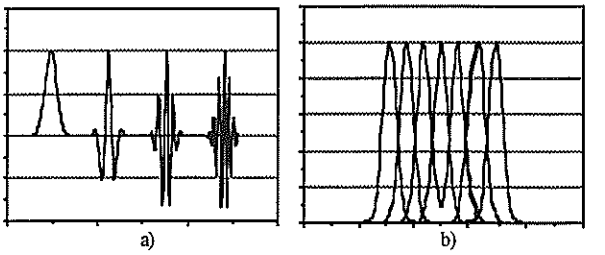


Figure 1 : Some typical Gabor functions presented in a) the time domain, b) the frequency domain.

It has been shown that the power in natural images is spread more or less uniformly within octave frequency bands [4]. In addition, it is believed that octave band division is performed in the human visual system [3]. The above considerations encourage the introduction of a pyramidal Gabor decomposition with basis functions expressed in the time domain as :

$$h_{m:s}(t) = \begin{cases} \hat{g}_D(t-mD) & \text{for } s<0 \\ \hat{g}_D(2^s t - 2^{-s} mD)\cdot e^{jk(s)Wt} & \text{for } s\geq 0 \end{cases}$$

and in the frequency domain as :

$$H_{m:s}(\omega) = \begin{cases} \frac{\sqrt{2\pi}}{W}\,\hat{g}_W(\omega)\cdot e^{-jmD\omega} & \text{for } s<0 \\ \frac{\sqrt{2\pi}}{W}\,\hat{g}_W\left(\frac{\omega-k(s).W}{2^s}\right)\cdot e^{-j4^{-s}mD(\omega-k(s)W)} & \text{for } s\geq 0 \end{cases}$$

where $k(s) = \frac{3*2^s-1}{2}$.

The extent in the frequency domain of this class of functions is concentrated in octave bands, and since they are Gabor functions,they are optimally localised in both domains. These functions have a number of interesting properties for coding purposes, which will be discussed later in this paper. Figure 2 shows the real parts of some typical pyramidal Gabor functions and their power spectra.

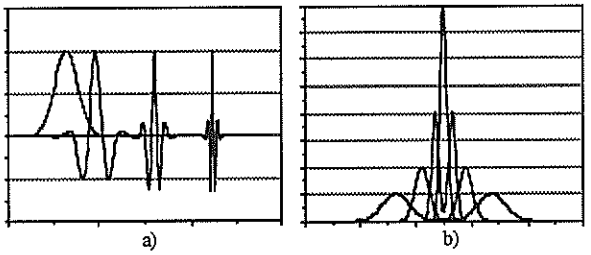


Figure 2 : Some typical pyramidal Gabor functions presented in a) the time domain, b) the frequency domain.

3. Calculation of the decomposition.

3.1. The 1D decomposition

Consider the sampled 1D signal $\{f(k)\}$and its decomposition into a weighted sum of elementary functions $\{g_l(k)\}$.

$$f(k) = \sum_{l=0}^{N-1} g_l(k).x_l \quad (k=0,\ldots,M-1;\ N\leq M) \qquad (1)$$

The above decomposition can be expressed in matrix notation as:

$$\underline{f} = G\cdot\underline{x}$$

where

$$\underline{f} = \begin{bmatrix} f(0) \\ \cdots \\ f(M-1) \end{bmatrix} \quad G = \begin{bmatrix} g_0(0) & \cdots & g_{N-1}(0) \\ \cdots & \cdots & \cdots \\ g_0(M-1) & \cdots & g_{N-1}(M-1) \end{bmatrix} \quad \underline{x} = \begin{bmatrix} x_0 \\ \cdots \\ x_{N-1} \end{bmatrix}$$

If N<M the relation (1) cannot be satisfied exactly. Hence a criterion should be defined to find the best solution to (1). In such situations the least squares error criterion is the most commonly used method. Using this criterion, the solution to equation (1) is the vector $\hat{\underline{x}}$ minimizing $(G.\hat{\underline{x}} - \underline{f})^T (G.\hat{\underline{x}} - \underline{f})$, which can be found analytically to be:

$$\hat{\underline{x}} = (G^TG)^{-1}G^T\underline{f}$$

The NxN matrix G^TG is not always invertible. It can be singular or close to singular if the elementary functions in (1) are interdependent or highly similar. To avoid the problem of matrix inversion in such a situation, the singular value decomposition can be used. It should be noted that if M=N and the elementary functions $\{g_l(k)\}$ are independent the decomposition becomes an invertible transform, in the sense that the reconstructed data $\hat{\underline{f}} = G.\hat{\underline{x}}$ will be an exact reconstruction.

3.2. The 2D decomposition

In the 2D case, we consider the decomposition of a 2D signal (for example, an image) into a set of 2D elementary functions:

$$f(k_1,k_2) = \sum_{l_1=0}^{N_1-1}\sum_{l_2=0}^{N_2-1} g_{l_1:l_2}(k_1,k_2).x_{l_1:l_2} \qquad (2)$$

If the elementary functions $g_{l_1:l_2}(k_1,k_2)$ are separable, they can be written as :

$$g_{l_1:l_2}(k_1,k_2) = g_{l_1}(k_1)\cdot g_{l_2}(k_2)$$

In this case, (2) can be written in matrix form as:

$$F = GXG^T$$

where F is the 2D data matrix, X is the 2D decomposition coefficient matrix:

$$F = \begin{bmatrix} f(0,0) & \cdots & f(0,M-1) \\ \cdots & \cdots & \cdots \\ f(M-1,0) & \cdots & f(M-1,M-1) \end{bmatrix} \quad X = \begin{bmatrix} x_{0:0} & \cdots & x_{0:N-1} \\ \cdots & \cdots & \cdots \\ x_{N-1:0} & \cdots & x_{N-1:N-1} \end{bmatrix}$$

and G is the same decomposition matrix as before.
Again, for N<M the exact relation cannot always be satisfied. Here also the least mean squares error criterion can be used to define the best solution (according to this criterion), given by:

$$\hat{X} = (G^TG)^{-1}G^TFG(G^TG)^{-1}$$

The use of the singular value decomposition is valid here also.

4. The decorrelation property of the Gabor decomposition.

In this section we are going to show, by an example, the decorrelative quality of the pyramidal Gabor decomposition, compared to that of the Gabor decomposition. The image to be coded is first decomposed as a linear combination of separable elementary Gabor functions, as defined below:

$$g_{m1:m2:n1:n2}(k_1,k_2) = g_{m1:n1}(k_1).g_{m2:n2}(k_2)$$

The same operation is used to decompose the image into a set of 2D separable pyramidal Gabor functions given by:

$$h_{m1:m2:s1:s2}(k_1,k_2) = h_{m1:s1}(k_1).h_{m2:s2}(k_2)$$

Figure 3 shows how these functions are positioned in both the spatial and spatial-frequency domains.

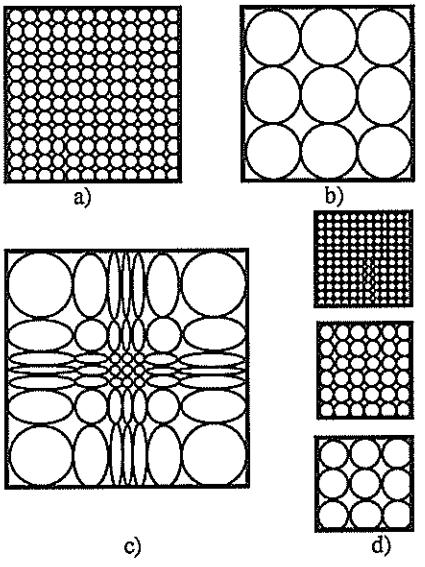


Figure 3 : Spatial and spatial-frequency partitions for 2D separable functions used in the decomposition. a) Spatial-frequency partition by Gabor functions. b) Spatial partition by Gabor functions. c) Spatial-frequency partition by pyramidal Gabor functions. d) Spatial partition by pyramidal Gabor functions.

The histograms of the decomposition coefficients obtained are shown in Figure 4. This example shows that the pyramidal Gabor decomposition provides better decorrelation compared to the Gabor decomposition. This result can be verified if one estimates the entropy in each case, defined by:

$$E = -\sum_{i=0}^{N-1} p_i.\log(p_i)$$

where p_i is the probability that a given coefficient appears. Applying this definition to our expansions, we obtain an entropy of 1.44 bits for the Gabor decomposition and an entropy of 1.42 bits for the pyramidal Gabor decomposition. These results are in accordance with the histograms shown in Figure 4.

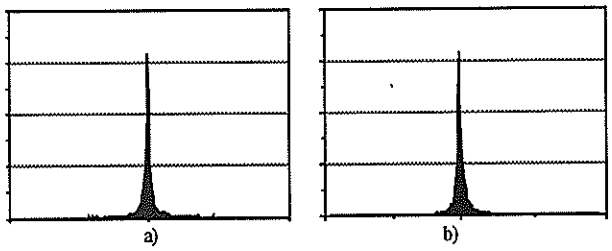


Figure 4 : The decomposition coefficients histograms obtained for a) the Gabor decomposition, b) the pyramidal Gabor decomposition.

5. Application to data compression.

5.1. Still image compression.

Figure 5 illustrates a simple codec system, which is used to demonstrate the performance of the Gabor and the pyramidal Gabor expansions for data compression. The input image is first decomposed by simple matrix multiplication as described in section 3.1. The resulting coefficients are then uniformly quantized and thresholded to retain only the most significant coefficients. The addresses of retained coefficients are run length coded by counting the number of consecutive zeroes between each pair of transmitted coefficients. The number of bits to code both the amplitude and the address of each transmitted coefficient is estimated by an entropy measure. At the receiver the inverse operation is performed to reconstruct the data. Some compression results obtained by this system are shown in Figure 6.

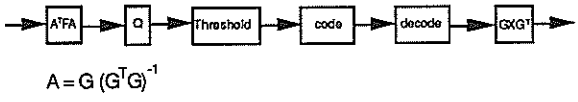


Figure 5 : The codec system used for 2D compression.

5.2. Image sequence compression.

To compress an image sequence several methods can be used. Here, as an example, we use a method based on a differential interframe coding. The principle of this coding method is given in figure 7. The sequence is decomposed frame by frame and the coefficients of each frame are uniformly quantized, similar to the previous method. In this case, however, because of the high correlation between the

successive frames, only the coefficients which differ significantly from those from the previous frame are coded and transmitted. The addresses of the transmitted coefficients are coded by a run length coder, counting the number of consecutive non-transmitted coefficients between two transmitted ones. The bit rate to code the quantized amplitude and the addresses of the transmitted coefficients is estimated by an entropy measure.

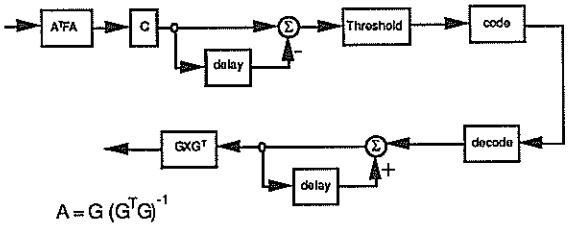


Figure 7 : The codec system used for the image sequence compression.

Four frames of a compressed sequence obtained by this system are shown in Figure 8.

6. Conclusion

A technique for coding image sequences, based on a pyramidal Gabor decomposition, was discussed. The elementary functions of the decomposition form a class which covers the frequency domain in octave bands. The results obtained show the potential of compression techniques based on localised transforms, especially when they are derived based on knowledge of the human visual system . These results can still be improved by introducing the time response of the human visual system in the design of new 3D elementary functions instead of the current 2D functions. Future investigations will be in this direction.

References

[1] D. Gabor. " Theory of communication.", J. Inst. Elect. Engr. 93, pp 429-457, 1946.

[2] S. Marcelja. " Mathematical description of the responses of simple cells.", J. Opt. Soc. Am. 70(11), pp 1297-1300, 1980.

[3] H.R. Wilson. " Psychophysical evidence for spatial channels.", In Braddick O.J. and Sleigh A.C., editors, Physical and biological processing of images., Springer-verlag, 1983.

[4] D. Field. " Relation between the statistics of natural images and the response properties of cortical cells.", J. Opt. Soc. Am. A Vol. 4, pp 2379-2394, 1987.

Figure 6 : Compression results using the method of section 5.1. for still images, by a) Gabor decomposition, the estimated compression ratio is 5:1 and the SNR = 28.6 dB, b) pyramidal Gabor decomposition, the estimated compression ratio is 5:1 and the SNR = 30 dB, c) Gabor decomposition, the estimated compression ratio is 9:1 and the SNR = 24 dB, d) pyramidal Gabor decomposition, the mated compression ratio is 9:1 and the SNR = 24 dB.

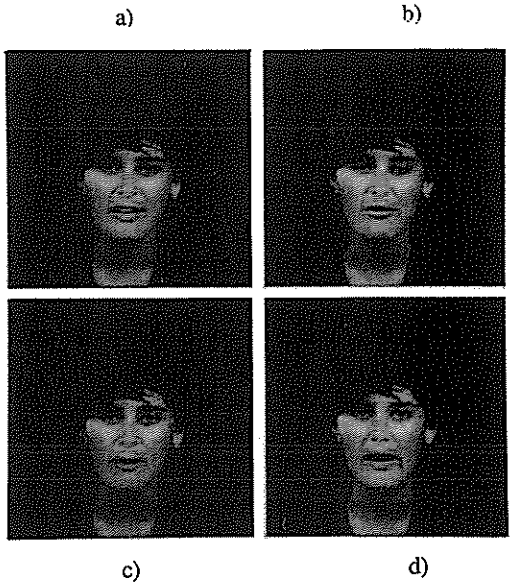


Figure 8 : Four frames of a reconstructed sequence obtained by the method of section 5.2. The estimated compression ratio is 224:1, this gives a bit rate of 59 Kbits/sec. The original sequence is the COST 211-bis test sequence, Miss America, with 256x 256 pixels per frame, PCM coded with 8 bits per pixel, and 25 frames per second. This result is obtained without any temporal subsampling.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

IMAGE SEQUENCE CODING BASED ON EDGE AND LINE DETECTION

Gaetano GIUNTA Todd R. REED Murat KUNT

Laboratoire de Traitement des Signaux, EPFL-Ecublens (DE)
CH-1015 Lausanne, Switzerland

A second generation method of image sequence coding is presented. A law for sharing the spatial lowpass and highpass components is derived from a minimum cost/resolution optimality criterion. The detection of directional elements (namely, edges and lines) is carried out by using both linear and non-linear (median) filtering. The coding is based on near optimal estimators, and is well suited for differential pulse coding modulation. The method has been applied to standard one-second sequences, obtaining estimated compression ratios of about 320:1 with a reasonably good reconstruction quality.

1. INTRODUCTION

Second generation image coding [1] is based on current knowledge of the neurophysiology of the human visual system. According to this approach, the 2D signal is projected onto a subjective visual subspace in order to code only the actually useful information. In particular, the directional sensitivity of the spatial response of the first neurons, together with their time response, play an important role for analytically specifying such a perceptual subspace.

Directional methods can be applied not only for coding single images, but also sequences. In this case, several differences arise. First, a single image must be represented quite exactly in order to allow foveal analysis on each region. This is not true when dealing with a time-varying image, because the required quality depends on how long a particular structure appears. An object which appears very briefly may not need to be perfectly represented. However, some changes in the image can come from moving objects, that the human eye is able to track and follow, which then require a good quality of reconstrution. Such changes usually have a significant highpass (HP) information component. Furthermore, local average luminance is often only slightly variant in time. As a consequence, the lowpass (LP) information of image sequences often have a lower entropy than single images.

2. SPATIAL FILTERING AND "BIT SHARING LAW"

The method presented herein is based on the separated analysis of each spatial LP and HP filtered image, constituting the sequence to be coded. The filter used to separate the sequence into spatial LP and HP components should have the following features. First, the two portions should be complementary, so that the sum of LP and HP frequency responses is an all-pass filter, in order to consider in an equal manner all the frequential components. Second, the slope of the cutoff profiles should be steep enough to allow the maximum information to be included in the LP part, given the cutoff bandwidth. Third, the slope of such profiles should be slow enough to avoid ringing artifacts (Gibbs effects) in the detected directional structures. The above requirements are also important in image reconstruction because the detected locations of edges and lines can sometimes be inexact. Using profiles without pronounced sidelobes reduces the effect of such estimation error. The frequency function we have chosen to perform the filtering is shown in fig. 1.

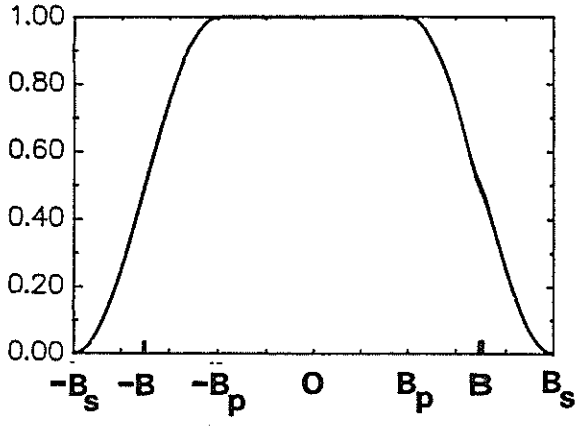


Figure 1. Frequency response of the LP filter.

Its analytical expression is:

$$W(f) = \begin{cases} 1 & \text{for } f < B_p \\ 0 & \text{for } f > B_s \\ 0.5[1 + cos(\pi \frac{f - B_p}{B_s - B_p})] & \text{otherwise} \end{cases}$$

where B_p is the bandpass frequency, B_s is the stopband frequency, and B is the halfpower bandwidth. From visual inspection, we have found that a good compromise between the two opposing requirements for the shape of the frequency response is to choose a transition bandwidth $(B_s - B_p)$ approximately equal to 1/4 of the bandwidth B.

A further question arises when we want to separate the LP and the HP components using a spatial linear filter (such as the parametric one proposed in the previous section), namely how to choose the bandwidth B of such a filter. One possible choice is to split the sequence into two complementary parts which can be coded by using the same number of bits. This choice minimizes the cost/resolution ratio. The reason we seek to minimize such a function is that we aim to have both low cost (in the number of bits) and high resolution in the reconstructed image. The proof of such *bit sharing law* is reported in the following.

Let us assume the following cost function:

$$F = \frac{cost}{resolution} = \frac{L + H}{R_L + R_H} \tag{1}$$

where L is the number of bits required for the LP components, H is the number of bits required for the HP components (in the perceptual space), R_L is the resolution of the LP reconstructed image, and R_H is the resolution of the HP reconstructed image. Our purpose is to minimize F as a function of the bandwidth B.

Let us assume the following relationships, valid only in a neighbourhood of the minimum of F: L is proportional to the square of the bandwidth, H is near constant with regard to the bandwidth, and R_L and R_H are proportional to the bandwidth. These three hypotheses are suggested by both theory and experimental evidence. The first hypothesis is correct because the LP information is subsampled, so that its cost (in bits) is proportional to number of samples of the subsampled image. The second hypothesis is quite exact in practice because the number of bits required for the perceptual parts of the HP components (i.e. the edges and lines) are approximately proportional to the number of the detected edges and lines. The number of directional elements do not vary when the variation in B is small, because edge and line extraction methods are usually thresholded. The third hypothesis becomes clearer when we consider the LP and the HP images separately. The bandwidth of the LP and HP images are proportional to their resolutions. That is, for a Dirac delta input, the output spot size for both the LP and HP filters is proportional to B.

The thesis we want to prove is that the minumum of F happens when $L = H$. In fact, let us rewrite F as a function of B.

$$F(B) = \frac{L(B) + H}{R_L(B) + R_H(B)} \tag{2}$$

which becomes, under the above three hypotheses:

$$F(B) = \frac{aB^2 + H}{cB + dB} = \frac{aB^2 + H}{(c + d)B} \tag{3}$$

The minimum is when the first derivative of $F(B)$, $F'(B)$, equals zero, and the second derivative of $F(B)$, $F''(B)$, is greater than zero.

$$F'(B) = \frac{a}{c + d} - \frac{H}{(c + d)B^2} = \frac{aB^2 - H}{(c + d)B^2} = \frac{L - H}{(c + d)B^2} \tag{4}$$

$$F''(B) = \frac{2H}{(c + d)B^3} > 0 \tag{5}$$

Then, $L = H$ is the condition for the minimum.

3. COEFFICIENT EXTRACTION

After spatial filtering by using the filter of fig. 1 (in our trials B is equal to one fourth of the sampling frequency in order to meet the bit sharing law), the following processing is performed. As a first step, the LP images are quantized using 8 amplitude levels (15 differential levels). The LP processing consists of subsampling the LP filtered and quantized images by a factor which depends on the bandwidth of the LP filter (in this case by a factor of 4 in each dimension, 16 in total). A temporal subsampling by a factor of 4 (assuming sequences composed of 25 frames per second) is then performed. The locations of the pixels in the subsampled sequence are chosen in an interlaced manner to simulate continuity of changes. The physiological reason for the above temporal subsampling is that the human eye can detect fast changes in sequences (the physiological bandwidth is 30 Hz), while it is not able to estimate the magnitude of such changes (the perception process requires some tenths of seconds at least). Therefore, the information can be sampled with a rate as low as 5 Hz, but it needs be represented as changing more continuously (more than 30 Hz are required). This is consistent with the standard representation of TV images, which are dispayed at 50 or 60 (interlaced) frames per second.

The HP part is processed to extract edges and lines of specific orientations. This simulates the behavior of the human visual system, which is particularly sensitive to

directional structures. The edge and line detection algorithms do not work on the whole image, but on an ensemble of masked static or dynamic regions defined according to the presence of relevant changes in the sequence. In order to extract edges and lines from the HP sequence, we have used directional FIR filters working on only 4 directions to limit the computational complexity of the method. The edge and line detection is separately accomplished by assuming a set of candidate edge and line points from the binarized HP image, convolved with directional median filters to help detection, and then thresholded. An algorithm of edge and line following [2] performs the connection of edge points, based on a pair of thresholds. The overlapping elements are deleted. The maximum amplitude coefficient (the square root of the energy coefficient, plus sign), the position, and the length of each edge or line are stored. The whole procedure is run once for the sequence on the static regions, while once per frame on the moving ones. The maximum amplitude coefficients of both edges and lines are quantized using 8 levels before being stored in the final lists.

4. CODING AND RECONSTRUCTION

The best way to code information is to retain only the innovation part of it (i.e., the difference between the signal and its optimally predicted value). Therefore, we have considered the use of predictors which can approximate the optimal one. The predictors used must also be simple enough to allow a fast signal reconstrution. We have found that simple time zero-order predictors (ZOP) are good approximations of optimality for sequences with no camera or background motion. Therefore, we have chosen a time ZOP for each coefficient of the sequence. This predictor is applied to the LP sequence, quantized to 8 levels. The HP coefficients which are coded are the differences between the current (8-levels) amplitude, start-position, and length of all edges and lines and the values for the closest edge or line in the previous frame. A spatial ZOP has been chosen for differentially coding the first image in the sequence and the static edges and lines. We have not chosen combined spatial-temporal predictors for both LP and HP predictors because they do not supply any significant improvement in terms of entropy reduction, while adding much more computational complexity. The techniques we have used are equivalent to using differential pulse code modulation (DPCM) for transmitting all the parameters.

The reconstruction of the image has been accomplished by adding the quantized LP and HP sequences. Reconstructing the LP images is straightforward (classical interpolation with a window function to reduce sidelobes). The HP part has been synthetically reconstructed by using parametric profiles for each edge and line location, which approximate the original ones. The edge profile was chosen in accordance with [1]:

$$w(x) = Axe^{-\frac{x^2}{2\sigma^2}} \tag{6}$$

where x is the transversal coordinate, A is the magnitude of the synthetic wavelet $w(x)$, while σ is a measure of its extent. This whole extent of $w(x)$ should be as large as the main lobe width of the impulse response of the filter used to separate the LP and HP components. Therefore, it depends on the bandwidth B. A value of the parameter σ equal to 1.6 pixels has been used in our trials from a visual fitting of average observed waveforms. Because a line can be regarded as the derivative (along the transversal direction) of an edge, the parametric line profile has been chosen as the derivative of the above expression. Finally, a non-linear post-processing based on median filtering is performed in order to remove spurious values of the reconstructed sequence.

5. RESULTS AND PERFORMANCE

The above procedure has been applied to standard test sequences in order to assess its validity. The LP bandwidth has been fixed at 1/4 of the spatial sampling frequency for these sequences. This allows a spatial subsampling of the LP images of a factor of 4. The LP sequences were also temporally subsampled by a factor of 4. For a one second period in a typical sequence, 59 moving edges and lines (on the average) per frame, and 389 static edges or lines for the entire sequence, have been detected and considered for coding.

As a performance measure we have chosen the entropy of the difference between the true value and the predicted one. This measure allows us to evaluate the optimal number of bits required according to Shannon's theory, independent of the coder implementation. As an entropy estimator we have used the approximate expression [3]:

$$H = \sum_i p(f_i) \log p(f_i) \tag{7}$$

where f_i are the quantized levels of each coded coefficient. This expression is theorically correct if the following two approximations are assumed. First, the predictor we have used is optimal in a MSE sense. This constraint implies that the estimated innovation is orthogonal (uncorrelated) to the whole past history (and therefore to the predicated value of signal). Second, the estimated realizations of the innovation process are not only orthogonal, but also independent. This is actually true only for Gaussian process. In fact, both these approximations are pessimistic because the estimated entropy values are generally higher than the correct ones [3].

The entropy evaluation has shown that the LP sequence requires 20,121 bits (6,772 for the first frame, 13,349 for the others), while the HP part needs 20,733 bits (5,296 for the static edges and lines, 15,437 for the moving ones). This is consistent with the bit allocation required by the bit sharing law. The total number of bits required for a one second sequence is 40,854, with a compression ratio of 320:1 (i.e. 0.025 bits per pixel and per frame). The original and reconstructed first, fifth, tenth, and fifteenth frames in the original sequence are shown in fig. 2, while the same reconstructed frames are shown in fig. 3. The perceived image quality is reasonable, especially for such a high compression ratio.

6. CONCLUSION

Second generation methods seem to be even more attractive for coding sequences than single images. We have shown that it is possible to obtain very high compression ratios with moderate quality degradation using this method. Future investigations include considering the use of a larger number of directional filters, the application of such filters in the 3D space, and the practical realization of the DPCM coders.

REFERENCES

[1]. Kunt M., Benard M., Leonardi R., "Recent Results in High Compression Image Coding", IEEE Trans. Circ. and Syst., vol. CAS-34, no. 11, pp.1306-1336, 1987.

[2]. Nieminen A., Kunt M., Gisler M., "Very low bit rate image sequence coding using object based approach", Int. Conf. SPIE, Visual Communications and Image Processing '88, vol. 1001, pp. 854-863, 1988.

[3] Pratt W.K., "Entropy Representation", in "Digital Image Processing", pp. 185-189, J. Wiley and Sons, New York, NY, USA, 1978.

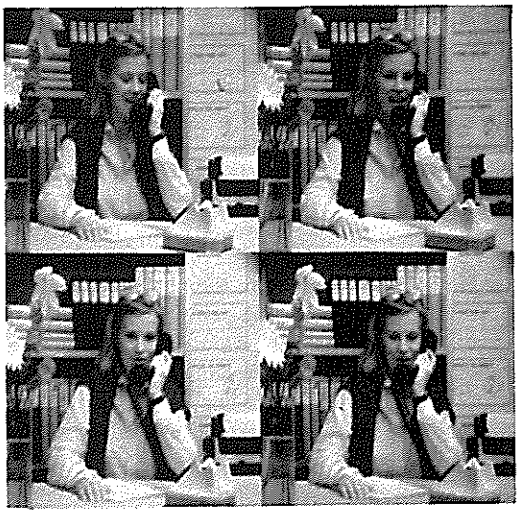


Figure 2. Four original frames from the sequence.

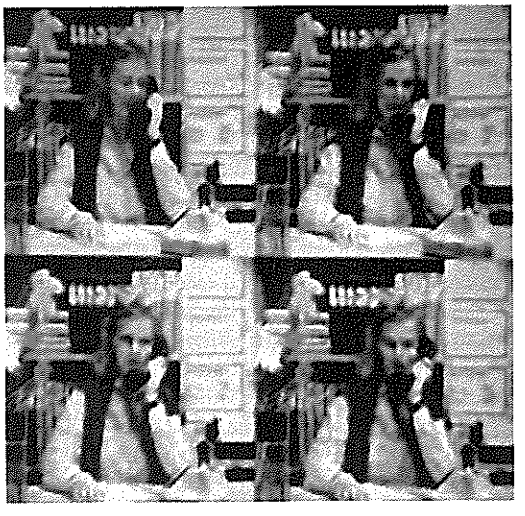


Figure 3. Four frames from the reconstructed sequence.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

REGION-ORIENTED CODING OF MOVING VIDEO - COMPATIBLE QUALITY IMPROVEMENT BY OBJECT-MASK GENERATION

Christoph Stiller, Wolfgang Guse and Michael Gilge
Institute for Communication Engineering
Aachen University of Technology
Melatener Str. 23, D-5100 Aachen, West-Germany

ABSTRACT

In most video codecs block-oriented displacement compensation is applied. Of course regions in video scenes do not move blockwise, but with arbitrarily shaped borders. Therefore block-oriented displacement compensation does not exploit interframe correlation efficiently compared with object-oriented displacement compensation, which adapts the shape of moving regions to the shape of the objects in the scene. The decoder presented in this paper uses this effect to improve the image quality of a block-oriented codec without requiring any alteration of the coder. In particular no contour data needs to be transmitted. The decoder algorithm is based on object-segmentation which serves mainly to copy stable background information from a background storage into the image.

Introduction and System Overview

Block-oriented displacement compensation has been standardized by the CCITT H.261 norm [1] and Reference Model 5 [2] for p·64 kbit/s coding of moving video. Vectors corresponding to blocks which contain moving regions (object) as well as static regions (background) cannot describe the motion of both regions correctly. Typically, those vectors describe the motion of the object. Hence in blocks including both, background and object, the block-oriented displacement compensation moves the background along with the objects. The resulting prediction error in such regions has large amplitudes. Consequently basic patterns of the DCT remain visible when a rough quantizised DCT is applied for update coding, although these regions (background) are often known from previous frames.

The decoder introduced in this paper improves the picture quality of a block-oriented codec, without requiring any changes of the coder. The block-diagram is depicted in figure 1.

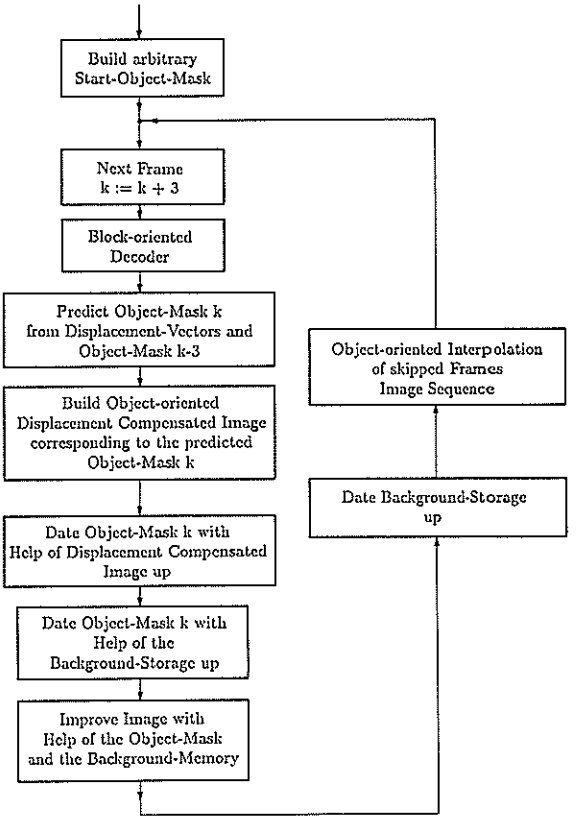


Figure 1: Block-Diagram of the Decoder

The decoder is based on an object-mask, which is a binary image marking object and background regions. This corresponds to a typical video-phone situation where speakers (objects) are moving in front of structured static background. In an arbitrary sit-

uation, where several independently moving regions overlap another, the introduced decoder improves image quality at edges between moving regions and static background. At edges between independently moving objects the image of the block-oriented decoder is exactly reconstructed. The object-mask is not completely new calculated for each frame but follows moving objects continuously.

An improved image is constructed using the object-mask to distinguish between objects and background.

- Background (static and decovered) regions which are known in the background storage are copied from there. Background regions which are unknown in the background storage are copied from the block-oriented reconstructed image.
- Moving objects are copied from the block-oriented recontructed image and spatially lowpass filtered causing a subjectively improved image.

Additionally the object-mask serves to detect background regions and refresh the background storage from the block-oriented reconstructed image there.

Generation and Update of the Object-Mask

The main feature of the decoder is the adaptation of the object-mask to the shape and position of objects. This adaptation is realized in three steps:

1. Displacement compensation of the object-mask
   The decoder starts with an arbitrary object-mask. A prediction of the new object-mask is generated by displacing the object-mask of the previous frame according to the transmitted displacement-vectors. Since vectors corresponding to blocks which mainly contain background area and only include a small part of the object more probably describe the motion of the background than the motion of the object, those vectors are (for prediction of the object-mask) replaced by the vector of that neighbouring block which includes most object pixels. This way knowledge about shape and position of objects in the last frame is conserved for the prediction of the actual object-mask.

2. Adapt object-mask according to the block-oriented reconstructed image
   Regions assigned wrong in the object-mask will lead to an erroneous object-oriented displacement compensation. Therefore in regions, where object-oriented displacement compensation fails (large difference to block-oriented image), it is tested whether a change of that region in the object-mask improves the object-oriented displacement compensation. In that case, that region is changed accordingly in the object-mask. This way, the object-mask is improved in mean-square-error sense of the displacement compensated image. The block-oriented reconstructed image serves as approximation of the original image.

3. Adapt object-mask according to the information of the background storage
   If background area is accidently included into the object-mask, in these regions the displacement of the previous frame leads to similar results as the displacement of the background-memory when the same displacement-vectors are applied. These regions are detected and the object-mask is set to "background" if the substitution of the displacement compensated image with the undisplaced background does not cause large differences to the block-oriented image.

Object-oriented Displacement Compensation and Interpolation of skipped Frames

The object-mask enables an improved displacement compensation, since objects can be displaced independently from the background. The displacement compensated image is constructed distinguishing between three regions:

- Static background is copied from the previous block-oriented reconstructed image.
- Decovered background regions which are known in the background storage are copied from there. The regions which are unknown in the background storage are displaced blockwise from the previous block-oriented reconstructed image and are therefore identical to the actual block-oriented reconstructed image.
- Moving objects are displaced with the vectors which were used for displacement of the object mask (previous section).

In a similar manner the interpolation of skipped frames is distinctively improved by prediction of an object-mask corresponding to the skipped frame and object-oriented displacement compensation with this object-mask.

Simulation Results

The decoder was tested using the sequences "Swing", "Miss America" and "Trevor" for the block-oriented coder introduced in [3]. Since "Miss America" has no structured background, moving the background along with the object does not cause any additional errors in the displaced image and hence the algorithm does not lead to improved decoding results.

Figure 2.c shows the first calculated object-mask for a frame from "Swing" (figure 2.a) when an object-mask without any objects (figure 2.b) is used as arbitrarily start object-mask. It can be seen that the first object-mask is already a good approximation of the real object-shape. Since knowledge about the object-shape and position in the last frame is available for the decoder in later frames, the object-mask is further improved.

Figure 3.a shows a block-oriented decoded image and figure 3.b it's improvement using the object-mask, while figures 4.a and 4.b show the decoder's improvement for a skipped frame which is motion adaptive interpolated from the transmitted frames and the displacement-vectors.

Conclusions

Block-oriented displacement compensation neglects interframe-correlation in some regions. Therefore block-oriented displacement compensation leads to unreliable prediction compared with object-oriented displacement compensation. As shown by simulations, object-mask generation is a tool to improve the quality of block-oriented codecs without requiring any changes of the coder. This allows for quality improvements compatible to the H.261 norm [1] and Reference Model 5 [2].

Since it is possible to calculate a good object-mask in the decoder even if no contour data is coded explicitly, object-segmentation allows to exploit interframe-correlation more efficiently, without necessarily requiring additional contour data to be transmitted. The object-mask can even save data because displacement information can be estimated and coded more efficiently with knowlegde about the contour of moving objects. Therefore an object-oriented displacement compensation at the coder can be expected to improve image quality further. For syncronization between coder and decoder, the same object-mask is calculated at coder and decoder and an update for the object-oriented displaced image is coded [4]. A generalization of the binary object-mask introduced in this paper leads to an object-mask that distinguishes between any differently moving objects, even if they have common edges [5].

Acknowledgements

The authors would like to thank ANT-Nachrichtentechnik GmbH for good cooperation and for sponsoring the project.

References

[1] CCITT Standard Recommodation H.261, Part II, Section 2.1: "Specification for p*64 kbit/s flexible Hardware".

[2] CCITT, Working Party XV/4: "Description of Reference Model 5 (RM 5)", Specialists Group on Coding for Visual Telephony, Doc. No. 339, March 1988.

[3] Grotz, Mayer, Suessmayer: "A 64 kbit/s Videophone Codec with forward Analysis and Control", Signal Processing: Image Communications 1 (1989), pp. 103-115, Elsevier Science Publishers B. V.

[4] Stiller: "Verfahren zur Bildung eines Praediktionsbildes", Application for a Patent at the German Patent Office, No. BK 89/69, August 1989, (in German).

[5] Guse, Gilge, Stiller: "Region Oriented Coding of Moving Video - Motion Compensation by Segment Matching", Proc. 5th European Signal Processing Conf. Eusipco 90, Barcelona Spain, Sept 90 this Volume.

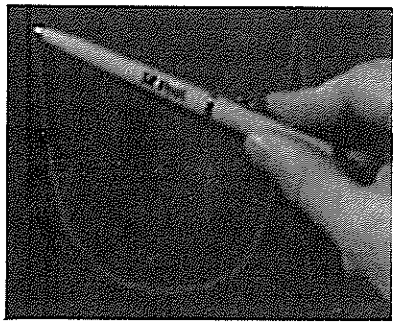


Figure 2.a: Original Frame

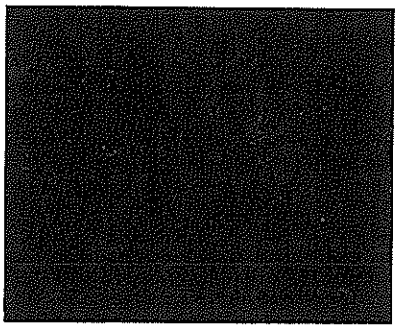


Figure 2.b: Start-Object-Mask not including any information

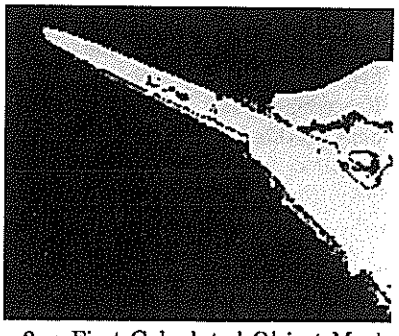


Figure 2.c: First Calculated Object-Mask

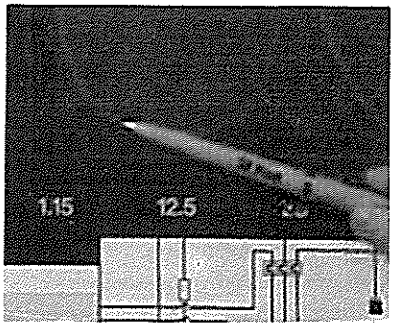


Figure 3.a: Block-oriented decoded Frame

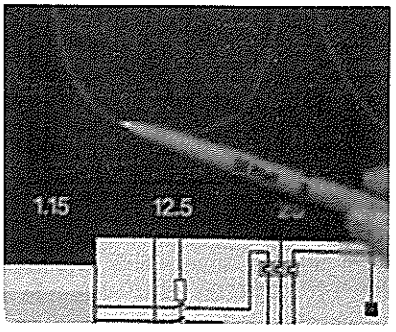


Figure 3.b: Object-oriented decoded Frame

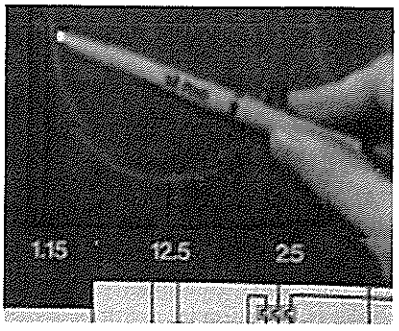


Figure 4.a: Block-oriented interpolated Frame

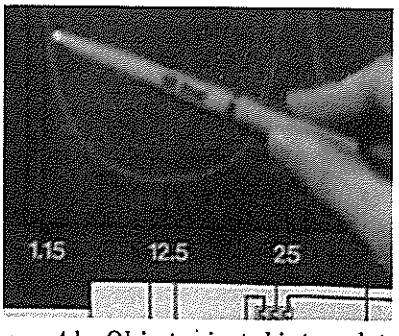


Figure 4.b: Object-oriented interpolated Frame

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

AN ATM ADAPTED VIDEO CODING ALGORITHM USING KNOWLEDGE BASED TECHNIQUES

F.Pereira (IST/JNICT - Portugal)
L.Masera (CSELT - Italy)

Centro Studi e Laboratori Telecomunicazioni
Via Guglielmo Reiss Romoli, 274
10148 Torino - ITALIA

The recent interest of video coding people on ATM environments is strictly related with their flexiblity to deal with signals which have time varying requirements. The variable bitrate coding quality improvements are greater for lower bitrates what recommends this kind of techniques for videotelephone transmissions. The final subjective quality may still be improved taking into account the different subjective impact of the various videotelephone image areas - face, body and background. ATM transmissions are particularly suited for this coding which tries to optimize the temporal and spatial distribution of the available bitrate.

1. INTRODUCTION

The development in the last years of multimedia/multisignal networks is strictly related with asynchronous transmissions. In fact the future Broadband ISDN (B-ISDN) will support the Asynchronous Transfer Mode (ATM) for a large range of services among which assume particular relevance the video signals. These signals range from typical communication signals, as videotelephone, to typical distribution signals, as normal or high definition TV. The interest of video coding people on this kind of environments is strictly related with their flexibility to deal with signals that require a time varying bandwidth, as done typically video signals. This characteristic justifies the intense study on variable bitrate coding techniques that allow to reach noticeable improvements on the image quality; these quality improvements are greater for the lower bitrates what specially recommends this kind of techniques for videotelephone transmissions.

To increase the overall image quality it is important to distribute the available bitrate following the image activity variations and taking into account the different subjective impact of some image characteristics. This fact is particularly important for videotelephone sequences where relevant improvements can be obtained by means of background/foreground segmentation and identification of areas with different subjective impact. Exploiting the knowledge that this kind of images represent head and shoulder of a speaker shooted by a TV camera, it is possible to identify the face, body and background areas. The face area catches the major attention of the interlocutor and, if specially processed in order to obtain a very good quality, it can improve the subjective quality perceived by the user, even if the global Signal to Noise Ratio is not improved. The use of knowledge based techniques seems particularly suited for ATM transmissions where bitrate variations can easily be supported.

This paper presents the performance of an ATM adapted video coding algorithm [1] using knowledge based techniques in order to improve the overall subjective quality of videotelephone coded sequences.

2. THE KNOWLEDGE BASED TECHNIQUES

The knowledge based techniques have here the only aim of finding the three major subjective areas; it is worth noting that the present application does not require an high accuracy for the image segmentation as the classification will be used on the macroblock level. This fact justifies the use of a simple algorithm to roughly segment the videotelephone image limiting the computational effort.

The knowledge based processing begins with the application of the Sobel gradient to the present frame (intraframe processing) and the computation of the frame differences between the present and the previous frame (interframe processing). In order to cut-off the background from both the resulting processed frames, two thresholds are automatically

selected by analysing the triangular areas at the top-right and at the top-left of these frames; this procedure assumes that only background is present in these areas.

By merging the information coming from the intraframe and the interframe processing a rough sketch of the speaker's head and shoulder shape is produced. Afterthat some noise is eliminated by means of a non-linear filtering and a more precise area enveloping the speaker is determined. At this point the face area is detected by means of a simple geometrical criteria based on its dimension and position within the area covered by the speaker. As final result each macroblock is classified as belonging to one of the three subjective areas - face, body and background.

3. THE ATM ADAPTED CODING ALGORITHM

In the field of video coding, special importance is given to the extension of the existing algorithms, designed for synchronous environments, to ATM transmissions. One of these algorithms is the Reference Model (RM) studied in the CCITT - Specialists Group on Coding for Visual Telephony (SGXV/WPXV/1), as future standard for synchronous transmissions at p*64 kbit/s (CCITT H.261 Recommendation) [2]. This algorithm is essentially characterized by the interframe coding based on the hybrid scheme, hierarchical subdivision of an image into GOB (Group of Blocks) and Macroblocks, prediction of the incoming block values by means of motion compensation and filtering, Discrete Cosine Transform applied to the prediction-error blocks and a buffer control basically acting on the quantization step.

One of the CCITT H.261 ATM extensions is the Modified Reference Model [1] which design is strictly related to the parameters adopted to characterize the variable bitrate (VBR) flow. These parameters are negotiated at the call setup and are here the average bitrate - computed over the whole sequence - and the peak bitrate - computed over a frame. In order to respect the negotiated statistical parameters two control strategies were implemented - the average and peak bitrate controls.

The average bitrate control parameter, called "excess", is defined as the difference between the number of bits produced after the beginning of the transmission and the number of bits that could be transmitted during the same time on a fixed channel working at the average bitrate initially negotiated. This parameter permits to respect the negotiated average bitrate acting essentially on the quantization step throught the Quantization Step Control. The peak bitrate control is used to avoid that the negotiated maximum number of bits per frame is exceeded.

3.1. The Quantization Step Strategies

The introduction of the knowledge based results in the ATM adapted coding algorithm is particularly easy in this coding scheme since it uses a block based hierarchical spatial partition of the image. This structure allows to classify each block (or macroblock) as face or body or background depending on the knowledge based informations. Since the coding algorithm does a uniform quantization of the DCT coefficients, the different processing of the various macroblock categories may efficiently be made by means of the quantization step (QS). The macroblocks classified with higher subjective impact have higher privileges and viceversa.

The possibility of dynamically allocate the available bits, optimizing the final subjective quality, is an interesting challenge; the quality gains are particularly dependent on the relative dimensions of the subjective areas and on the kind of image of each area, specially the areas that must be penalized to improve the quality in the higher subjective impact areas .

The present study considers three quantization step strategies:

Strategy 0 : No QS privileges

The non consideration of areas with different subjective impact leads to a coding where the more stationary areas (specially the background) have higher Signal to Noise Ratio and subjective quality. The average quantization step is similar for all the subjective areas but noticeable impairments may result, specially in the face area.

Strategy 1 : The face and the background areas have a fixed QS *(defining parameters : QSface and QSback)*

Since the face is the most important subjective area and the background is the less important one, this strategy attributes to the face a low QS and to the background a very high QS. This solution allows to guarantee a very uniform quality over the face and background areas directly related to their subjective relevance. In this strategy the body absorbes the

activity variations with a resulting quality that depends on the face and background activities, on QSface and QSback and on the bitrate.

Strategy 2 : The face and the body areas have fixed QS decrements applied to the QSs resulting from the Quantization Step Control and the background has a fixed QS *(defining parameters : Decface, Decbody and QSback)*

In this strategy the face and body subjective areas absorbe the image activity variations even if with different intensity; the background has a fixed QS related to its relative subjective relevance. The face and body QSs are related to the overall coding situation (trade-off bitrate-quality) having values that depend directly on the excess value for each frame; note however that the face and body QSs are computed simultaneously in the first face or body macroblock and once computed are maintained constant over all the frame; in a first approach these two values differ by (Decface-Decbody) which imposes a quality gap between the two areas. In order to avoid noticeable fluctuations, the amplitudes of the face and body QS variations between two frames are limited. This strategy has the capability of reaching an equilibrium situation between the negotiated statistical bitstream characteristics (specially the average bitrate) and the QS privileges what is not always possible for strategy 1.

3.2. The Adaptation Mechanism

As face, body and background activity characteristics may be very different for diverse sequences, it is difficult to initially find the strategy defining parameters that lead to the ideal equilibrium between the subjective quality of the defined subjective areas. This fact justifies the necessity of introducing an automatic mechanism that can dynamically adapt the strategy defining parameters to the specific image requirements. The adaptation mechanism has as basic rule the higher relevance of the face subjective quality which it is the first to be priviliged and the lower relevance of the background subjective quality which it is the first to be penalized. The adjustment of the strategy defining parameters is made every frame using as control parameters the Luminance Signal to Noise Ratios and the Quantization Steps of the previous frame and the excess. The strategy defining parameters have a maximum amplitude variation of 2; the face and body maximum QS variation amplitudes of strategy 2 are always valid. The minimum and maximum global QS values that are 4 and 64, respectively, are always respected.

This adaptation mechanism allows to reach not only the ideal quality trade-off between the subjective areas but also the trade-off between the statistical bitstream constraints and the QS strategy, initially not guaranted for strategy 1.

4. RESULTS AND CONCLUSIONS

The presented quantization step strategies have been studied for low bitrates (48, 64 and 128 Kbit/s) using the standard videotelephone sequences 'Claire' (frames 1-79), 'Miss America' (frames 80-128) and 'Trevor' (frames 129-157); note that 'Trevor' is

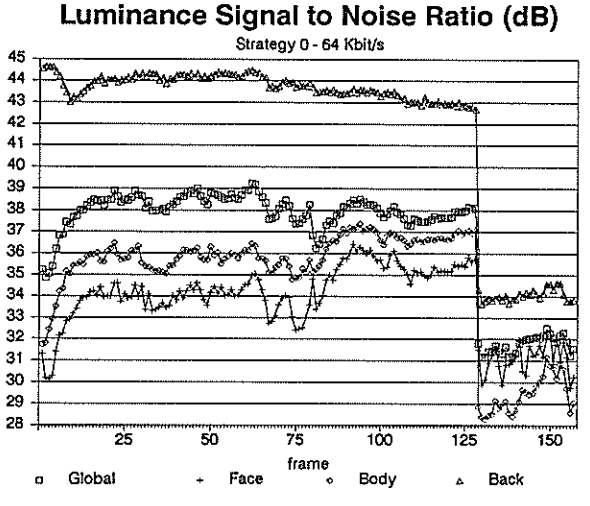


Figure 1

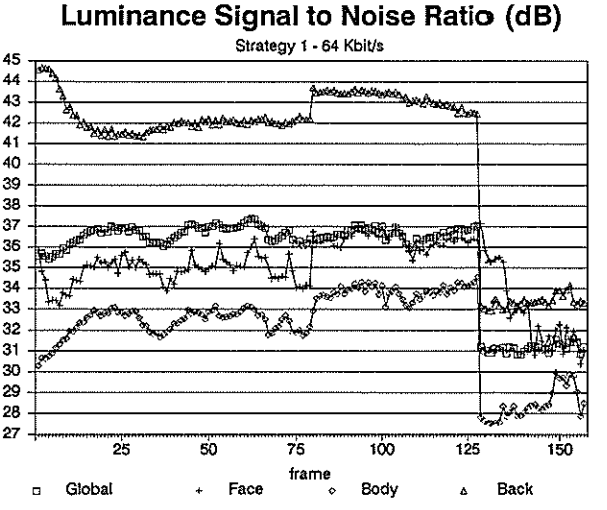


Figure 2

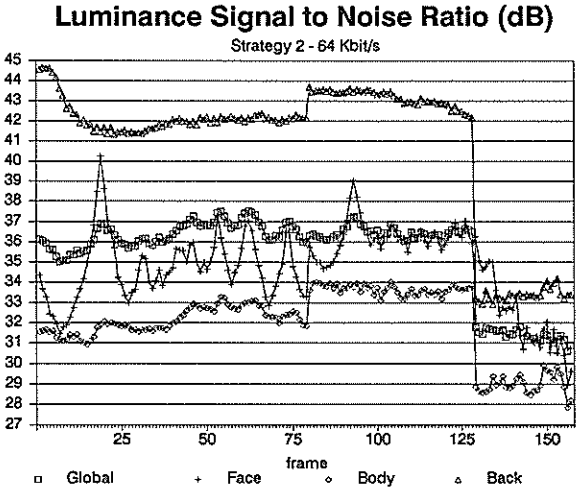


Figure 3

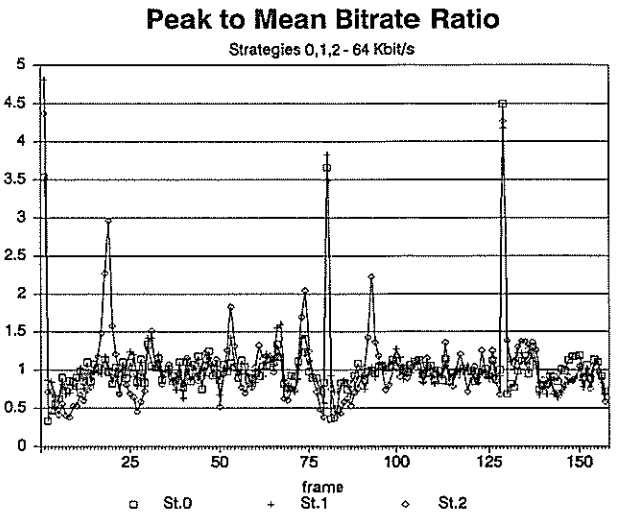


Figure 4

more difficult to code than the other subsequences due to its faster and wider movements. In figures 1 to 4 are presented the following results :

- *Luminance Signal to Noise Ratio (LSNR) for the face, body, background and global for the presented strategies*
- *Peak to Mean Bitrate Ratio (frame level) for the presented strategies*

The analysis of the obtained results suggests the following conclusions :

- *The quality gains are very dependent on the relative subjective area dimensions and on the kind of image of each area. For the present sequences, for example, the background has little to give and almost all the face quality gains result from the body quality detriment (note that for 'Trevor' the situation is a little different); this fact justifies the final filtering of the body area, at least for the lower bitrates.*
- *The adaptatiom mechanism is essential since the strategy defining parameters must be dynamically adapted to the sequence requirements. In this case the adaptation mechanism has been adjusted to give a LSNR gap of 2-3 dB between the face and the body.*
- *The final subjective impact of strategies 1 and 2 is similar and is noticeably greater than that of strategy 0.*
- *The global results of strategies 1 and 2 are similar but strategy 2 is characterized by quality fluctuations related to the critical moments due to the direct relation between the excess and the face quantization step.*
- *The variation of the subjective quality gains with the bitrate depend critically on the image characteristics. Generally speaking, the subjective quality gains decrease with the bitrate since for higher bitrates the quality impairments of strategy 0 are less noticeable; however it is possible to have an initial subjective quality gains increase since for very low bitrates the quality gains are almost inexistent if the body and background strategy 0 QSs are already close to the maximum QS.*
- *The maximum PMBR values are similar for all the strategies.*
- *The presented algorithm is completely compatible with the standard CCITT algorithm since the H.261 frame structure permits quantization step changes on a macroblock level. This fact is very important since the knowledge based techniques introduction is completely transparent to all the other CCITT H.261 users.*

REFERENCES

[1] Pereira,F., and Quaglia,M.:"Extension of the CCITT Visual Communication Coding Algorithm for Operation in ATM Networks", Image Communication Journal, vol.1, n.2, April 1990

[2] CCITT SG XV, Draft Recommendation H.261, Tokyo Meeting, 1989

'SIGNAL PROCESSING V: Theories and Applications'
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ANALYSIS OF PEL-RECURSIVE WIENER-BASED ESTIMATION ALGORITHMS FOR GENERAL 2D MOTION

L. Böröczky, K. Fazekas, Dept. of Microwave Telecom.,
T. Szabados, Dept. of Mathematics,
Technical University of Budapest
XI. Goldmann tér 3. BUDAPEST, HUNGARY, H-1111

ABSTRACT. In this paper pel-recursive Wiener-based estimation algorithms are analyzed for general 2D motion. The first Wiener-based version of motion estimation was proposed by Biemond et al. and later they extended their algorithm to general 2D motion as well. In previous papers we analyzed their assumptions, introduced more realistic models, now we extend it to the case of general 2D motion. The proof for the convergence of Biemond's algorithm given by us, is valid for the present case too. Numerical simulations are carried out for comparing the performances of the original Biemond's method and our methods. These simulations are useful for finding optimal parameters of the proposed algorithms as well.

1. INTRODUCTION

There is a growing interest in transmission of images of moving objects at low bit rates (videophone, videoconference, etc.) Motion estimation is one of the key commponents of low bit rate video coding. All major coding methods (predictive, transform, and interpolative) can be improved by motion estimation. Because of the computational complexity, mainly motion estimation algorithms based on translational models were applied. A pel-recursive motion estimation algorithm for this purpose was introduced by Netravali and Robbins [1]. A Wiener-based version of it was proposed by Biemond et al. [2], which was shown to perform well. Their paper was concerned with pure 2D translational motion of one rigid object. In [4] and [5] we analyzed their algorithm: more realistic models were developed, compared and under suitable conditions we proved the convergence of their method.
Later, Biemond et al. extended their algorithm to general 2D motion, namely rotation and translation, of a rigid object [3]. In the present paper we extend our previously proposed methods for this general case.

2. PRINCIPLES

If a rigid object moves with a 2D motion during the time interval [t,t-1], then

$$I(z,t) = I(z_{tr},t-1), \qquad (1)$$

where I(z,t) denotes the intensity value within a frame at a point z=(x,y) in the image of the moving object at time t,

$$z_{tr} = T_{-\varphi}\, z - d, \quad T_\varphi = \begin{bmatrix}\cos\varphi & -\sin\varphi\\ \sin\varphi & \cos\varphi\end{bmatrix}, \quad d=\begin{bmatrix}d_x\\ d_y\end{bmatrix}, \qquad (2)$$

φ is the angle of the rotation and d is the displacement vector.
Define a function called the transformed frame difference:

$$\Delta = I(z,t) - I(z_{tr}(p^i),t-1),$$

$$p^i = (d_x^i, d_y^i, \varphi)^T, \qquad (3)$$

where p^i is the ith estimation of the true motion vector. This error, which has to be minimized, converges to the true motion vector. Therefore an iterative algorithm is performed:

$$p^{i+1} = p^i + u^i, \qquad (4)$$

where u^i is the estimated update vector.
For a Wiener-based estimation of the update term one supposes that

$$\Delta = Gu + v, \qquad (5)$$

where G is the matrix of the first derivatives of the intensities at points of the so-called causal window and v is the linearization error. For finding u a linear regression model is applied [3]:

$$\hat{u} = L\Delta + m, \qquad (6)$$

where one wants to find the matrix L and vector m such that

$$E(\,\|u-\hat{u}\|\,)^2 \qquad (7)$$

be minimum (E denotes expectation). Then the parameters of the linear estimator are:

$$L = (R_u G^T + R_{uv})\left[GR_u G^T + R_v + GR_{uv} + (GR_{uv})^T\right]^{-1} \qquad (8)$$

$$m = E(u) - L\,E(\Delta),\ E(\Delta) = G\,E(u) + E(v), \qquad (9)$$

where $R_u = E(uu^T)$, $R_v = E(vv^T)$ and $R_{uv} = E(uv^T)$.
Several approaches are possible for finding the parameters L and m of the linear estimator.

Biemond et. al. [3] supposed that the update u and the linearization error v have zero expectations, diagonal covariance matrices and they are also orthogonal to each other. In [4] and [5] we stated that these assumptions do not hold in practice. Therefore, in case of pure translational motion, we introduced more realistic but at the same time more complicated algorithms. Since good results were obtained, it seems to be worth extending those methods for the ones described in [4] and [5]. We mention that these methods include 1D search [5] to eliminate numerical instability.

3. METHODS

In the sequel the original method proposed by Biemond et al. [3] is called Method 1.
In the case of Method 2, the assumptions are:
-u is a random vector with $E(u) = 0$, $R_u = \sigma_u^2 \cdot I_2$,
-v is a random vector with $E(v) = 0$, and arbitrary R_v,
$-R_{uv}$ is arbitrary.
Thus

$$\hat{u} = L\Delta \qquad (10)$$

where L is given by (8).

Method 3 is a modification of Method 2. Since both our theoretical computations and numerical simulations showed that the linearization error v has a non-zero mean in general, we also determined $E(\Delta) = E(v)$ during the simulation. Thus we obtained

$$m = -L \cdot E(\Delta) \qquad (11)$$

and an update vector

$$\hat{u} = L(\Delta - E(\Delta)) \qquad (12)$$

where L is given by (8).

Method 4 is a combination of Method 1 and 3. At each pel a $\hat{u}_0$ is computed first, then it is used a a mean of the randomly generated u while making statistics for R_v and R_{uv}, finally a modified update is determined by Method 3:

$$\hat{u} = L(\Delta - E(\Delta)) + \hat{u}_0. \qquad (13)$$

We proved the convergence of Biemond's algorithm in case of pure translational motion under suitable conditions [5]. That proof can also be extended for the present case.
For, the only difference between the cases of translational motion and general 2D motion using Method 1 is that the number of parameters is increased by one in the latter. Consequently, the proof in [5] is valid for the latter case with no significant change.

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

We carried out simulational experiments to compare the speed of convergence of the different algorithms and to show the influence of some variables (e.g. size and shape of support, number of iterations per pel) on the performance of the proposed algorithms.
The first experiments were performed on the synthetic-rotated image data in order to present clearly the convergence of the rotation parameter estimation. Fig. 1. shows one frame from the image sequence "Trevor White", that was used in the simulation experiment as an actual frame. Fig.2. illustrates its transformed version (previous frame), where the rotation is -6° degree (0,104 rad) around the image center. For the convenience of comparison, all investigated algorithms were simulated under the same conditions:
-The iterations were carried out along the image rows separately and the estimation of parameters obtained in the previous pixel was considered as the initial estimate in the current pixel location.

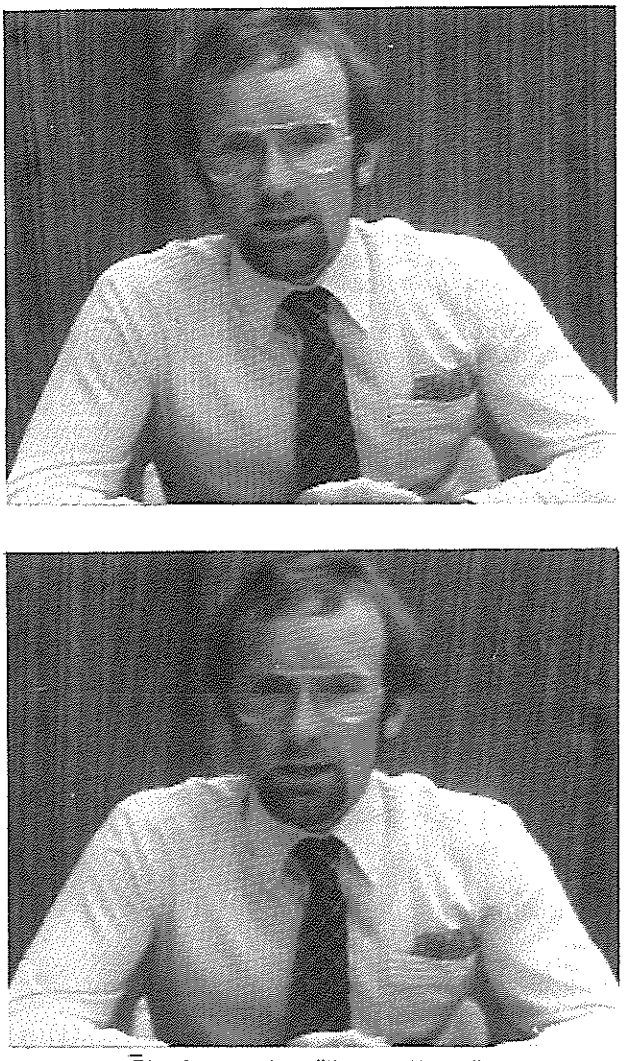


Fig.2. Rotated "Trevor White"

-For comparison of different methods one iteration per pel was performed and a causal support with N = 13 pixels was used (see Fig.4).
-A bilinear interpolation was applied to calculate the TFD (transformed frame difference) and the derivates in the transformed locations of the previous frame.
-In Methods 2, 3 and 4 the covariance matrices R_u, R_v and R_{uv} were statistically calculated from the previously obtained $u = (u_x, u_y, u_t)^T$ and TDF vectors in the pixels of the support (in Fig.4 the pixels with "x" sign). Thus R_v and R_{uv} are approximated using R_u, R_z and R_{uz} according to the following expressions:

$$R_v = E(zz^T) + GR_uG^T - GE(uz^T) - E(zu^T)G^T, \quad (11)$$

$$R_{uv} = E(uz^T) - R_uG^T \quad (12)$$

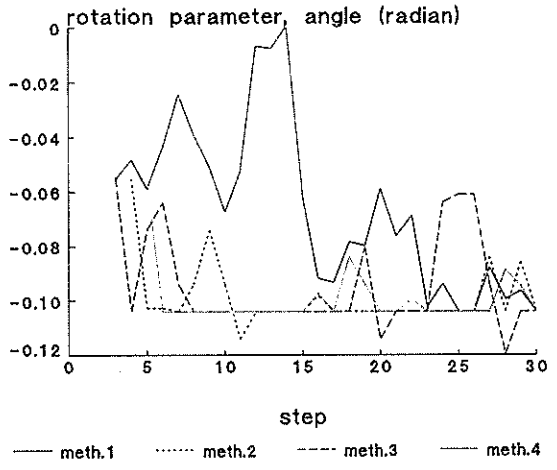


Fig.3. Comparison of methods

Fig.3. illustrates typical behaviours of the different algorithms. The proposed algorithms (Methods 2, 3 and 4) perform better than Method 1. The experiments showed no significant difference among the behaviours of the proposed methods, therefore we chose the simplest one (Method 2), to investigate how the different variables influence the speed and stability of the convergence. One of the most important variable is the size and shape of the support, used in the estimation procedure, especially in Methods 2, 3 and 4, where the covariance matrices are computed using also the information of the pixels in the support. Fig.4 shows four different supports, which were applied to the experiments.

xxx	xxxxx	xxxx	xxxxxxx
xxx	xxxxx	xxxx	xxxxxxx
xxo	xxo	xxxx	xxxxxxx
		xxxo	xxxo
N=9	N=13	N=16	N=25

Fig.4 Different causal supports

Fig.5. is a convergence plot of rotation parameter for Method 2, for various support sizes and shapes. As expected, a larger window size gives a better performance. The support with N=25 pixels gives the fastest convergence, but this size can be too large to exclude the image boundaries if the estimator with a change detector is applied to a real image sequence. The support with N=13 performs considerably better than the support with N=9, thus N=13 support can be recommended to use in the proposed methods. We also performed experiments with multiple iteration per pixel for Method 2. As Fig.6. shows, an increasing number of iteration improve the speed and stability of convergence. It should be noticed that from a mathematical point of view it is obvious to increase the number of iteration per pel in order to get better convergence, but the number of iteration is a choice constrained by timing (real time implementation) aspects, too.
To illustrate, how the proposed algorithms capable to estimate simultaneously the displacement and rotation parameters, we carried out experiment to a pair of frame from the "Trevor White" sequence. The rotation was -6^0 degree around the image center and dx=1, dy=1. Fig.7. shows the convergence plot of estimated parameters.
In this case the convergence is somewhat slower, but there is a stable convergence for every parameter.
From the initial experiments it can be concluded, that it is worth extending our previously proposed algorithms to estimate a general 2D motion. It is also turned out, that their performance can be improved using appropriate variables (support, initial estimate, number of iterations per pel). Therefore, our further attention will be focused on tuning these variables.

ACKNOWLEDGEMENT

The authors would like to express their thanks to Professor Jan Biemond and his staff at the Information Theory Group of the Delft University of Technology, for their help in performing computer simulations and for helpful discussions.

REFERENCES

[1] A.N. Netravali and J.D. Robbins, Motion-compensated television coding: Part 1, The Bell Syst. Techn. Journ., vol. BSTJ-58,no. 3, 1979, pp 629-668.
[2] J. Biemond, L. Looijenga, D.E. Boekee and R.H.J.H. Plompen, A pel-recursive Wiener-based displacement estimation algorithm, Signal Processing, vol. 13, 1987, pp 399-412.

[3] J. Biemond, J.N. Driessen, A.M. Geurtz and D.E. Boekee, A pel-recursive Wiener-based algorithm for simultaneous estimation of rotation and translation In: Visual Comm. Image Proc. III, Nov. 1988, Cambridge MA.
[4] T. Szabados, L. Böröczky, K. Fazekas, Analysis of a pel-recursive Wiener-based motion estimation algorithm, Proc. of ISSSE '89, Sept. 1989, Erlangen, pp 318-320.
[5] L. Böröczky, K. Fazekas, T. Szabados, Convergence analysis of a pel-recursive Wiener-based motion estimation algorithm, Time-Varying Image Processing And Moving Object Recognition, 2 Proc. of the 3rd International Workshop, Florence, Italy, 1989. ELSEVIER, 1990 Amsterdam, pp 38-45.

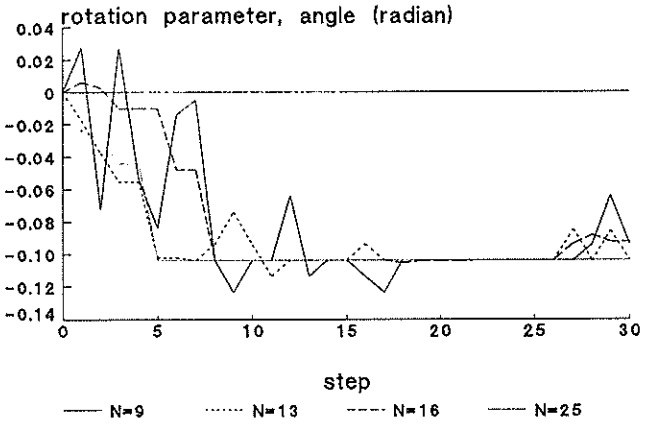


Fig.5. Comparison of supports

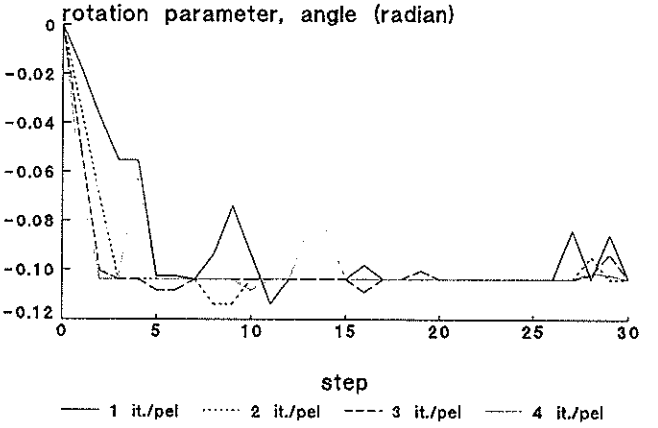


Fig.6. Comparison of number of iterations

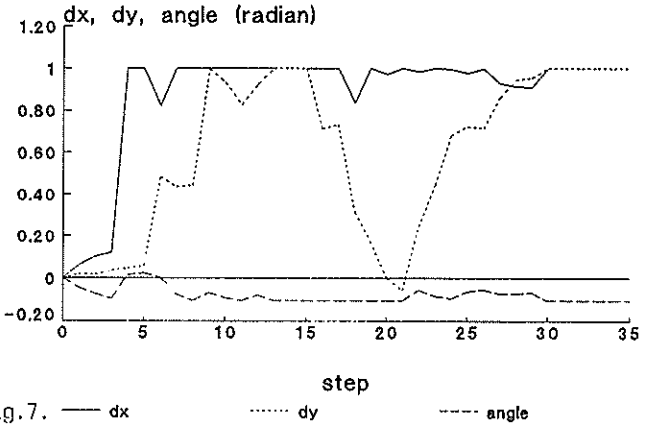


Fig.7.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A MODIFIED 2D-LOGARITHMIC SEARCH PROCEDURE FOR A MOTION COMPENSATED AND PRESEGMENTED PREDICTIVE CODING

Vincenzo DEL RE and Giovanni ZARONE

Dipartimento di Ingegneria Elettronica, Università di Napoli
via Claudio, 21 I-80125 Napoli, Italy

Among the "interframe" image coding techniques, movement compensated predictive coding has been deeply investigated, thanks to its properties of simplicity and high compression level. In this context, the present paper proposes and implements some modifications to the well known "2-D logarithmic search procedure" [1,2] in order to achieve some performance improvements. The suggested modifications, in fact, increase the movement estimation reliability and decrease the prediction error entropy, as the presented results demonstrate.

Moreover a typical artifact of block matching algorithms -due to the partition of images in fixed area cells- is overcome by performing a presegmentation of each picture in moved and fixed areas, and by assigning the estimated movement only to the moved pels of investigated blocks. Such a strategy improves the quality of the processed images.

Simulations utilizing real sequences were carried on according to the proposed logic and some result are presented. They seem very promising, since the proposed modifications of the 2-D logarithmic search procedure permit to achieve a rather homogeneous field of displacement vectors, while the presegmentation allow to decrease the bit rate and to improve the quality of the reconstructed image, at the cost of a small increase of computational complexity.

Finally, the separate quantization of prediction errors in moved and fixed regions is considered in the Appendix.

1. INTRODUCTION

The movement compensation technique based on block matching algorithms is often combined with 2-D logarithmic search procedure (2DLSP) in order to minimize the necessary operations. According to the assumption of the last method, the adopted distortion function -almost always the MAD- is convex within search area. Actually such an assumption very frequent is not appropriate, since several minima, which are equal or approximately equal, are found at different displacements. If this happens, the algorithm for researching the DMD (direction of minimum distorsion) can make mistakes giving rise to a not homogeneous field of displacement vectors. Therefore, in the present paper, the assumption of convexity of distorsion function is removed and modifications to Jain and Jain [2] algorithm are introduced.

2. THE MODIFIED 2DLSP.

Thanks to the assumption of convexity of the distorsion function, Jain & Jain's algorithm does not foresee that the same distorsion value can be obtained at two or more different displacements, nor when five neither nine (the final ones!) search area pels are tested. Actually, the above hypothesis very frequently is not appropriate, since several minima, wich are equal or approximately equal, are found at different displacements. This section deals with such an occurrance.

The proposed version of the algorithm uses MAD [3] as distorsion function, for the well known advantages, in terms of semplicity and computational time. This choice is strenghtened by Srinivasan & Rao's studies [4]. The presented algorithm looks for two points corresponding to two MAD minima. The pel corresponding to the lowest value of MAD function is selected to carry on the research only if the minimum is "well marked", that is to say that the difference between this value and the other selected one is greater than a given threshold T. Otherwise the situation looks like very embarassing, especially in the intermediate steps of the search procedure. In this case, in fact, the choice of one point or another may lead to a bad estimation of displacement vectors. In such a situation the proposed procedure gives up the 2D-Logarithmic Search Procedure, assuming the current block displacement equal to the average of displacements pertaining to the two nearest already investigated blocks (the upper and the left one). On the other hand, if the "not-well marked" MAD minimum is obtained only in the final step of the procedure (when all nine pixel of the ultimately reached 3×3 area are tested), two ways can be followed: either 2DLSP is still given up, as done in the intermediate steps, (A=0); or a possible error on displacement estimation (± 2 pels maximum) must be accepted using 2DLSP till the end (A=1). The choice between these last two alternatives has been made upon experimental considerations, which demonstrated that the second way has to be preferred (in fact it looks unfair to renounce the 2DLSP when it is almost come to its "happy end"). As mentioned above, experiments have been performed to determine the most appropriate value for T and A parameters. Tab. I presents the result obtained utilising

		A=0			A=1		
	T	CPUT	PE	DV	CPUT	PE	DV
Miss A	0	1032	0.80	.12	1007	0.78	.10
	.001	1030	0.80	.12	1006	0.78	.10
	.01	1023	0.84	.17	1002	0.80	.10
	.1	943	0.94	.06	948	0.91	.07
	1	675	1.13	.01	667	1.11	.02
Trevor	0	1027	0.90	.13	1022	0.72	.10
	.001	1027	0.91	.12	1019	0.73	.10
	.01	1018	1.01	.13	1005	0.97	.10
	.1	975	1.09	.08	969	1.05	.08
	1	640	1.13	.05	637	1.17	.04

Table I - Dependence of coder performance on T and A parameters.
[CPUT] = μs/pel; [PE] = [DV] = bits/pel.

"Miss America" and "Trevor" sequences, in terms of CPU time (CPUT), entropy of the prediction errors (PE) and of displacement vectors (DV). The simulations were carried out using FORTRAN programs developed on a Digital Microvax II system. Tab. I shows that

a) when T is great, the bit rate of displacement vectors decreases fast, whether if A equals 0 or 1. In fact, "severity" of MAD minimum search algorithm grows with T, so that more frequently 2DLSP is given up and displacement estimation from the adiacent blocks is obtained. So, a more homogeneous displacement vector field follows, performing a decrement of vector bit rate at the cost of an increment of prediction error bit rate. The reduction of CPU times as T increases can be explained in the same manner.

b) The A parameter has a weak influence on the studied quantities. Nevertheless, CPU time and prediction error bit rate decrease when A equals 1.

	Miss A			Trevor		
Algor.	HD	VD	PE	HD	VD	PE
Unmod.	2.41	2.61	3.42	2.72	2.90	3.20
Modif.	2.16	2.43	3.33	2.40	2.65	3.18

Table II - Entropy of horizontal (HD) and vertical (VD) displacements (in bits/component) and of prediction errors (PE, in bits/pel).

Finally, a reasonable choice appears to be T=0.01, A=1. The proposed approach, with these values for T and A parameters, attains a very homogeneous field of displacement vectors, at the cost of almost 15% increment of CPU time, with respect to the "classical" Jain & Jain's algorithm. This is corroborated by both the entropy decrement of prediction error and entropy decrement of horizontal and vertical displacement vectors (Tab. II). Figures 1 and 2 plot

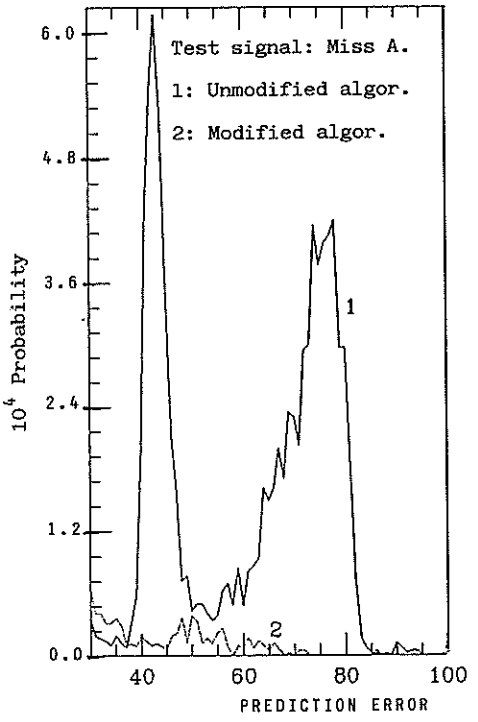


Fig. 1

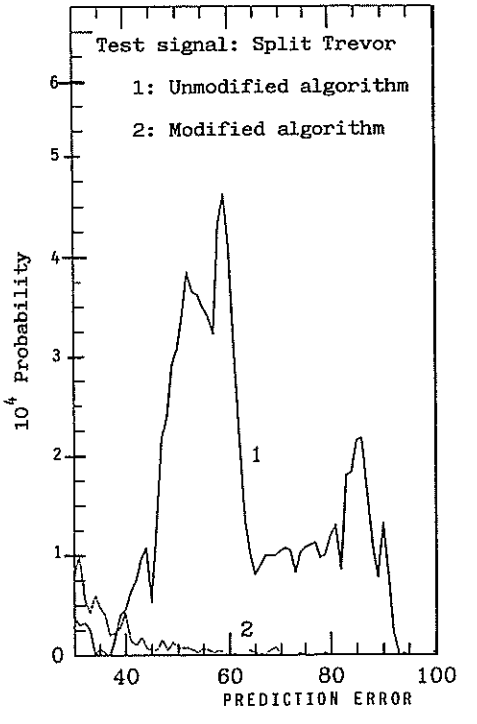


Fig. 2

the estimate of probability of prediction errors in the range (30,200), for both the conventional and the modified algorithm, using "Miss America" and "Trevor", respectively. The result show, with the reference to the modified version, a decreasing of probability of the higher errors.

3. PREDICTIVE CODING USING PRESEGMENTATION

In the context of predictive coding, a further decrease in bit rate can be achieved by singling out "fixed" pels and avoiding transmitting both displacement vectors and prediction errors pertaining to these pels. This approach, which performs a presegmentation in fixed and moved pels of each picture to be processed, is commonly adopted by pel recursive algorithms, but does not result being applied to block matching algorithms.

Once the moving area is singled out (by Netravali e Robbins [5] algorithm), the displacement is evaluated only for the moved blocks (i.e. the blocks containing at least one moved pixel). Then the block estimated displacement is assigned only to the moved pixel of the current block. Besides the saving in bit rate accomplished by avoiding transmitting both prediction errors and displacement vectors, this approach provides a better tracking of moving object shapes, obviating to a great extent a typical artifact of block matching tecniques, due to a fixed subdivision of each picture in previously sized blocks. By using the described presegmentation and (according to psyco-visual criteria) considering acceptable a prediction error less then or equal to 2, three different areas in the current frame can be identified: a fixed zone; a moved zone with "small" prediction errors (where, therefore, the application of movement compensation reveales itself profitable); a moved zone with "large" prediction errors (where, on the contrary, the application of the movement compensation turns out unprofitable). For the fixed area both displacement vectors and prediction errors are not transmitted. However, the addresses of pels belonging to the three different regions need to be transmitted, so that the receiver can correctly perform the reconstruction of the image by utilizing an adequate strategy. Table III shows the bit rates pertinent to prediction errors and to displacement vectors for the two quoted sequences. The comparison has been performed between the proposed method -wich requires the transmission of a three level map- and a motion compensated conditional replenishment method, wich avoid transmitting prediction errors below a given threshold (equal to 2) and needs the transmission of a two level map as a side information.

	Miss A		Trevor	
Preseg.	PE	DV	PE	DV
No	0.94	0.10	0.97	0.10
Yes	0.52	0.03	0.83	0.02

Table III - Coding cost (in bits/pel) of prediction errors (PE) and displacement vectors (DV).

We remark the convenience to apply the proposed method. The necessity of a three level may, in fact, should not rapresent an insurmountable difficulty: current studies show that it can be very efficiently coded by adopting a bidimensional run-lenght technique, obtained by scanning the map along a space filling curve [6].

4. CONCLUDING REMARKS

In this paper some modifications to 2DLSP (to improve motion estimation) and a presegmentation of each picture of the sequence (to make more appropriate the utilization of this estimation) are proposed and tested. The preliminary results seem very encouraging, although an increased computational difficulty follows. Another drawback -due to the presegmentation technique and to the subdivision of the moving area in a zone with "small" prediction errors and another zone with "large" prediction errors- is the necessity of sending to the receiver a three level addresses map. The cost of this side information, however, can be made very low by adopting "ad hoc" strategies.

APPENDIX: THE PREDICTION ERROR QUANTIZATION

Quantizers used in DPCM coding can be designed either on pure statistical basis or by using psyco-visual approaches, which take into account the still partial knowledge about the human visual mechanism.

For the "interframe" coders - as those ones used in the present work - the situation is somewhat more complex, provided that quantization errors more frequently occur in the "moving area" of the image, and their visibility depends on spatial and temporal variations of the scene.

At present, very reliable models are not available for the "interframe" case, neither about the prediction error distribution, nor about the quantization error visibility. Hence, "trials and errors" approaches are often required, and the designer experience plays a primary role.

Results hitherto presented, for the proposed motion compensated predictive coding strategy using presegmen-

	IN	OUT	L2		L3			
S1	0-1	0						
	2-4	3	IN	OUT	IN	OUT	IN	OUT
	0	0	3-6	4	3-6	4	71-82	77
S2	1-2	1	7-13	9	7-12	10	83-94	87
	3-4	3	14-22	16	13-21	17	95-107	102
			23-29	24	22-29	26	108-120	115
	3-6	4	30-38	33	30-38	35	121-133	128
	7-13	9	39-49	45	39-49	44	134-146	141
L1	14-20	16	50-69	60	50-59	55	147-159	154
	21-38	27	70-255	85	60-70	65	160-172	167
	39-255	50					173-255	179

Table A1 - Characteristics of S1, S2, L1, L2 and L3 quantizers.

		L1	L2	L3	L4
Miss A	PEE (bits/pel)	1.88	1.90	1.92	3.87
	SNR (dB)	73.4	80.5	84.2	∞
Trevor	PEE (bits/pel)	2.52	2.57	2.59	4.84
	SNR (dB)	97.9	105.1	118.7	∞

Table A2 - Region L performance.

		*	S1	S2	S3
Miss A	PEE (bits/pel)	0.00	1.33	1.97	2.69
	SNR (dB)	74.0	87.1	93.3	∞
Trevor	PEE (bits/pel)	0.00	0.88	1.66	2.13
	SNR (dB)	95.9	104.6	117.3	∞

Table A3 - Region S performance.
* indicates "no prediction error transmission"

tation, have been obtained avoiding to transmit prediction errors pertaining to the fixed area and to that part of the moving area where movement compensation was "successful" (see part 3), while the prediction errors obtained in the remaining part of the moving area (where movement compensation failed) were transmitted without any further quantization.

When the quantization problem is considered, the fixed area and the moving area with "small" prediction errors have to be grouped and handled - from the quantization point of view - in the same way. In fact, such moving zones very probably correspond to slow movement objects, that the observer can track. So, he recovers in part the sensitivity to disturbs that he normally presents at the fixed zones [7], and the above zones need to be transmitted with an appropriate fidelity. For clarity, call

Region S: the fixed area and the moving area where prediction errors less or equal to 2 (in absolute value) are obtained;

Region L: the moving area where prediction errors greater than 2 (in absolute value) are obtained.

Different quantizers, with different restitution level numbers were tested, either for region S or for region L. The used quantizers were obtained by trials and errors either "ex novo" or starting from already known ones. They are listed hereafter in a synthetic way, reporting the positive part of the quantizing laws, provided their simmetry (Tab. A1). Note that an overlap between the input ranges of quantizers pertaining to Region S and Region L exists, due to the presegmentation algorithm. According to our implementation of this algorithm, a pel for which a prediction error equal to 3 or 4 may be judged fixed or moved depending on how near pels were classified.

In order to objectively evaluate the reconstructed image quality, the SNR parameter was used, separately for Region S and Region L. In fact, it would not be significant to use a unique SNR for the whole image, provided that the observer sensitivity to analogous subjective errors in the two different regions is different.

Tables A2 and A3 present the prediction error entropy (PEE) and SNR pertaining, respectively, to Region L and S for the above introduced quantizers (test sequences: Miss America and Trevor). By S3 and L4 quantizing laws no further degradation is introduced respect to that one inherently present in the original digital signal. Strictly, the quantization affects also the displacement vector entropy and the map entropy, but our study demonstrates that this effect is negligible.

REFERENCES

[1] A.N.Netravali, B.G.Haskell, "Digital Pictures", Sect. 5.2.3.g, Plenum Press, New York, London, 1988.

[2] J.R.Jain, A.R.Jain, "Displacement Measurement and its applications in Interframe Image Coding", IEEE Trans. on Comm., pp. 1799-1808, December 1981.

[3] H.G.Mussmann, P.Pirsch, H.J.Grallert, "Advances in Picture Coding", PIEEE, pp. 523-548, April 1985.

[4] R.Srinavasan, K.R.Rao, "Predictive Coding based on efficient motion estimation" ICC 1984 Proc., pp. 521-526, May 1984.

[5] A.N.Netravali, J.D.Robbins, "Motion compensated TV coding", (Part.I), BSTJ, pp. 631-670, March 1979.

[6] G.Poggi, S.Stinchi, "Exploitation of Peano space-filling curve for coding addressing information", submitted for admission to Int. Workshop on 64 kbit/s coding of moving video, Rotterdam (NL), 1990.

[7] S.C.Kwatra, C.M.Lin, W.A.Whyte, "An adaptive algorithm for motion compensated interframe color image coding", IEEE Trans. on Comm., pp. 747-753, July 1987.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

SIMULATION OF A TELECONFERENCE CODEC FOR ISDN

S.Sallent A.Artero J.ZAMORA

Department of Applied Mathematics and Telematics
ETSETB-UPC, Apartado 30002, Barcelona 08034, Spain.

This paper describes an efficient spatio/temporal adaptive technique to the video-conference system that uses a new implementation of interframe prediction, conditional picture element replenishment and several subsampling strategies. The multimode coder/decoder is based on a new segmentation algorithm using a quadtree scheme. All the frames are divided in active and non-active blocks. This results in active blocks being associated by a operating mode depending of the displacement vector.

1. INTRODUCTION

Image sequence compression is essential for transmission and storage applications such as video-conferencing, video-phone, and digital transmission of cable TV.

The signals used in videoconference are generated by scanning a scene several times a second even though there may not be any change in the scene from one frame to the next. As a result of this the videoconference sequences are characterized by large areas of the scene change very little or not at all between successive frames. Thus, in each frame only a small amount of new information is required to specify these areas to the receiver. There is a high degree of correlation between adjacent picture elements in one frame or between elements in different fields or frames. Each frame of the sequence can be segmented into moving and stationary areas (head and Shoulders model).

This paper describes an efficient adaptive technique applied to the video-conference system that uses a new implementation of multimode codec. The modes are defined by several strategies, as motion compensation, interframe and intraframe prediction, temporal interpolation, 3D subsampling and quantization.

The codec is based on preserving the significant features of an image sequence, extract them by segmentation before coding. The procedure is based by applying adaptive thresholding segmentation to the difference between consecutive frames. This technique only transmits the information necessary to describe the intensity of elements that changed between frames. The simulated codec uses two frame memories, one at the receiver and other at the transmitter.

This technique has been applied to a simulated teleconference system and the bit rate assigned is very low, 192 kilobits per second.

2. SEGMENTATION ALGORITHM

Here a segmentation adaptive procedure is presented. The problem is to obtain a compact representation of the sequence information. The goal is to approximate the frames of the segmented sequence by moving and stationary areas (head and shoulders model).

In order to accomplish this goal, a regular decomposition quadtree method is used to segment the frames of the sequence into homogeneous regions of different block sizes. The block size ranging from 256 x 256 to 8 X 8 pixels. The root node corresponds to the entire frame and each son of a node represents a quadrant of the region represented by that node.

The segmentation process is controled according to two thresholds T_l and T_b. The number of pixels whose absolute value differences of the luminance between pixels of two consecutive frames is greater than a certain threshold T_l, are labeled as actives. The second threshold labels the block as active if the number of active pixels exceeds T_b.

The algorithm can be expressed as

$$\text{if} \quad |S_i(l,m,n) - S_{i-1}(l-1,m,n)| > T_l$$
$$\Rightarrow S_i(l,m,n) \quad \text{is active}$$
$$\text{and } A_i = A_i + 1 \ ;$$

$$\text{then one block } S_i \text{ is active if}$$
$$A_i > T_b \ ; \qquad (1)$$

where i and i-1 are the indices of two consecutive temporal blocks, and l,m,n are the temporal and spatial coordinates respectly.

Each one of the blocks labeled as active will be recursively segmented into four quadrants, unless the size of the block reaches the minimum size (8 x 8). In this way the quadtree is created and each block is classified into two categories active and non active.

This method leads to the representation of the original frame by a set of differents squares which are used to segment and label the original sequence. Figure 1 shows one frame of the original test sequence and the associated segmented frame.

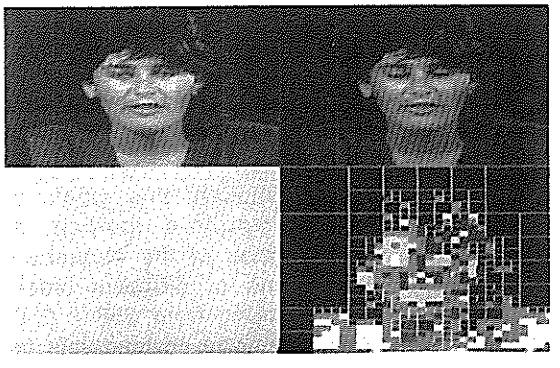


Figure 1. a) Two original consecutive frames. b) Frame difference of both. c) Application of the segmentation algoritm.

This segmentation algorithm is adaptive in the sense that exist three predefined areas where each one of them has their proper theresholds. Figure 2 shows the three adaptive zones.

The bits required to represent the quadtree depends on the information content in the frame difference as well as the thresholds used to recursively subdivide each block into four quadrants. To achieve a very low bit rate this thresholds are pre-set to T_l = 4 (over 256 grey levels) for all frame, being $T_b = T_b/0.6$ in zone 1, $T_b = T_b$ in zone 2 and $T_b = T_b/1.4$ in zone 3.

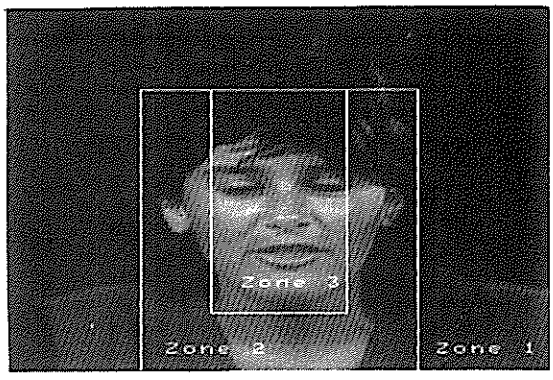
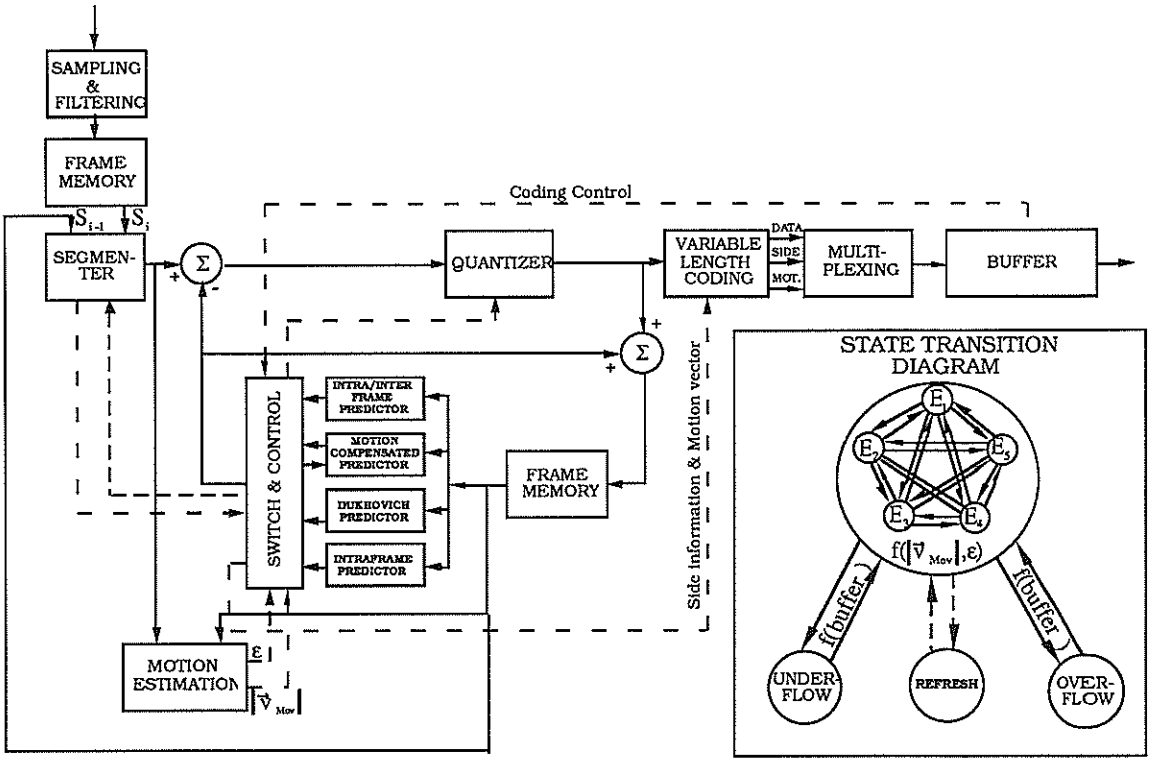


Figure 2. All the frames of the sequence have been splitted in the three activity zones according to temporal and spatial statistics.

The videoconference sequences can be modeled as a number of objects with different slow motion (foreground) superimposed on a background. The proposed segmentation algorithm assumes that each active block contains a single object moving on a fixed background. The active blocks are then fed to the movement detector which assigns the value of displacement vector and the proper distortion error to each active block. The elemental displacement can be specified by a translation if rigid body motion is assumed.

In each active block the translation vector $\vec{v}$ is found applying displacement estimators, and the distortion error ε is calculated using MAD (Module Absolute Distance) as a distance measure. Hence, overhead information must be transmitted, due to the position of the active block within the quadtree, and two parameters, the translation vector associated to active block.

The overhead information is a three percent over the total amount information and It is sending through the channel in multiplexed form.



ENCODER

3. MULTIMODE CODER

To transmit sequences of videoconference over a fixed-rate channel like as B-ISDN a buffer is required to smoth the information rate. The size of the buffer depends on the activity in the sequence signal and the maximum delay permitted. According to this, we introduce a spatio/temporal multimode codec based on the segmenter described above and controled by two decision parameters. The operation of multimode coder is illustrated in the figure 3. The coder has five modes of operation indicated by E_1,.....,E_5, each mode represents a specific choice of coder parameters (inter/intraframe prediction, movement-compensated prediction, non lineal and lineal predictions, sampling lattices and quantizer levels). The state transition rules are based on the module of the displacement vector $|\vec{\nabla}|$, the distortion error ε and the occupancy of the buffer. The states are controled by the switcher and are shown in table 1, together with typical set thresholds, and the principal coder parameters.

Figure 3 . Block diagram of the multimode coder and decoder. State transition diagram for 190 kilobits/s multimode coder.

The five modes define a main operation mode. Also a secundary operation mode controled by the buffer ocupancy is defined. The secondary mode has two submodes, the over and underflow which control the segmenter and the quantizer. For example in the blocks where there is little activity (underflow mode) a fine quantizer can be used, and in blocks with more activity the quantization can be coarser (overflow mode). Also a refresh sub-mode, which sends a fixed number of pixels of the original frame is implemented. Their position is calculated randomly within a frame.

4. EXPERIMENTAL RESULTS

$\|\bar{V}_{Mov}\|$	ε	TYPE OF CODIFICATION	E_K
{0}	[0,EM)	INTRA-INTERFRAME PREDICTOR	E_1
$[M_1, M_2)$	[0,E1)	REPEATED COMP. BLOCK	E_2
$[M_1, M_2)$ $[M_2, M_3)$	[E1,EM) [0,EM)	MOTION - COMPENSATION	E_3
$[0, M_3)$	$[EM, \infty)$	DUKHOVITCH	E_4
$[M_3, \infty)$	$[0, \infty)$	INTRAFRAME PREDICTOR	E_5

Table 1. State decision table based on parameters M1, M2, M3, EM and E1 being M1=1, M2=2, M3=8, EM=7.5 and E1=1 where EM is the average normalized error.

Figure 4. a) 5 and 6 "Miss America" coded frames with a compression ratio of 80, b) Two frames of the "Walter" sequence coded at compression ratio of 50.

A monochromatic "Walter Cronkite" sequence, which contains 16 frames of 256 x 256 pixels quantized uniformly to 8 bit per pixel, and "Miss America" (24 frames) were used as the test sequence sets.

Figure 4 a) shows the processed frames, 5 and 6 of the "Miss America" sequence, depending on the thresholds values shown above, with a compression ratio of 80. Figure 4 b) shows the processed frames 3 and 4 of the "Walter" sequence with a compression ratio of 50. Different compression ratio and sequence quality could be obtained varying the thresholds values.

Figure 5 presents the signal to noise ratio and the compression ratio, for the first twenty four coded frames of "Miss America" with an average speed of 206 Kilobits per second. In the transmitter a decimation process reduces the transmission speed at ten frames per second. In the receiver an interpolation process recovers the missing frames.

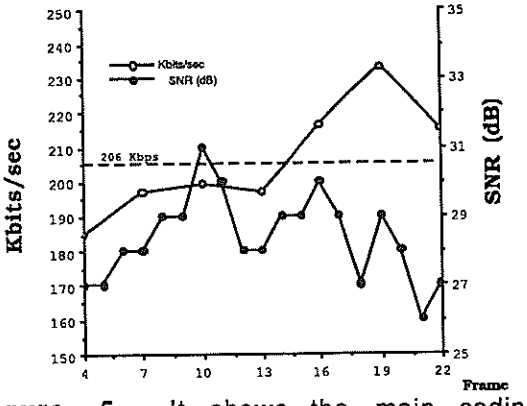


Figure 5. It shows the main coding parameters for the "Miss America" sequence.

5. CONCLUSIONS

In this paper we have proposed a new multimode coder and have shown its potencial for encoding image sequences of videoconference at lower bit rate. It is shown that the scheme presents low block effect, offers low computational complexity and can be implemented in parallel. Computer results are presented at 64 x n kilobits per second with excellent visual quality.

6. REFERENCES

[1] "Advanced Picture Coding", H.G. Husmann and P. Pisch, Proc. of IEEE, Vol. 73, Nº4, pp. 523-548, April 1985.

[2] "Motion Compensated Television Coding: Part I", A.N. Netravali and J.D. Robbins, The Bell Syst. Techn. Journ., Vol. BSTJ-58, Nº3, pp. 631-670, March 1979.

[3] "An Interpolative Spatial Domain Technique for Coding Image Sequences", P.J. Cordell, R.J. Clarke, Proc. of ICASSP-89, Vol. Nº4, pp 1917-1920, May 1989, Glasgow.

[4] "A three-dimensional spatial non-linear predictor for television", I.J. Dukhovich, J.B. O'Neal, IEEE Trans. On Com. COM-26, 578 (1978).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ON A HYBRID PREDICTIVE-INTERPOLATIVE SCHEME FOR REDUCING PROCESSING SPEED IN DPCM TV CODECS

RICARDO L. DE QUEIROZ
JOÃO B. T. YABU-UTI
Depto de Comunicações - Faculdade de Engenharia Elétrica
Universidade Estadual de Campinas
CP 6101 Campinas - SP 13081 BRAZIL
Fax : + 55-192-394717 Phone : + 55-192-397750

In this paper, DPCM was applied to Pyramid Coding strategy forming a simple and feasible hybrid extra/interpolative scheme. The Pyramid DPCM had the primary intention of increasing the interval between samples in DPCM. Furthermore, it has been shown that lower bit rates than regular DPCM are required for high quality resulting images.

1) INTRODUCTION

In DPCM TV CODEC's, the sampling rate lies near 10MHz, leaving an interval between samples in the order of 100ns and not allowing much time for all the needed processing. Due to the high degree of complexity of the CODEC's which have been lately proposed, their implementation could become a very difficult task. In the proposed scheme, adding parallel computation, this interval would increase to 200ns and 400ns, maintaining quality and reducing bit-rate. It also allows frequency differenciated coding, like Sub-Band Coding [1], multiresolution, progressive transmission as well as DPCM simplicity. Due to this fact, we do not expect great savings, but considerable performance improvement when compared with regular DPCM under the same conditions. That led us to a comparative behavior throughout this paper.

2) PYRAMID DPCM

The pyramid technique for progressive coding of images can be found in [2] and [3]. Let the original image be an array of NxN pixels (with N an integer power of 2) represented by ordering its lines in the sequence x(n). Let the sequence x'(n) be obtained by prefiltering and decimating x(n), with further interpolation of the decimated signal. Here, it is considered a 2-to-1 decimation and band restriction factor. In the pyramid construction process, we have made use of the following notation :

$x_0(n) = x(n)$

$x_{k+1}(n) = x_k(n)$ decimated

$x'_{k+1}(n) \Rightarrow x_k(n)$ by interpolation

$L_k(n) = x_k(n) - x'_k(n) \qquad k=0,1,\ldots,M-1. \qquad (E1)$

To recover x(n) we only need to reverse (E1) with the knowledge of

$x_M(n)=L_M(n);\ L_0(n);L_1(n);\ldots;L_{M-1}(n)$

In this Pyramid, there is an overhead of recorded nodes [3], which amounts roughly to doubling the number of original image pixels. However, if we had made use of a Reduced-Pyramid [3], these numbers would be equal. This pyramid scheme could be achieved without prefiltering before decimating. With proper interpolative filtering, half the samples in x'(n) would be coincident with those in x(n) and, therefore, the difference between them does not need to be coded.

In order to make the Pyramid feasible, we will restrict our attention to 3-level Reduced-Pyramid (M=2). Note that L2(n) and L1(n) would contain 1/4 of the nodes and L0(n) 1/2 of them. L2 is also composed by 1/4 of the original samples of x(n). If we code L2 with DPCM, its samples would be coded by taking the differences between the samples and extrapolative predictions of them [1]. Since L1 and L0 are formed by differences between interpolator's output and original samples, the whole system could be viewed as a hybrid inter/extrapolative prediction scheme for DPCM (or predictive/interpolative). The keypoint in this approach is the reduction of entropy of these differences due to the improved performance of interpolative prediction when compared with extrapolative one. Furthermore L2's DPCM will work with a 4 times longer sampling interval.

In Figure 1 the steps towards the DPCM pyramid are pursued, as follows :

i) Decimate, by 2 and 4, x(n) in order to find $x_1(n)$ e $x_2(n)$.
ii) Code $x_2(n)$ with a regular DPCM, therefore L2(n) is formed by the differences between $x_2(n)$ samples and their predictions.

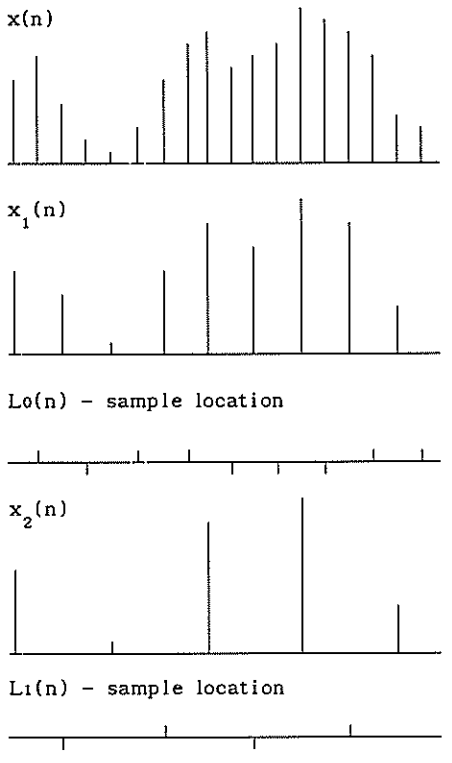


FIGURE 1 - EXAMPLE.

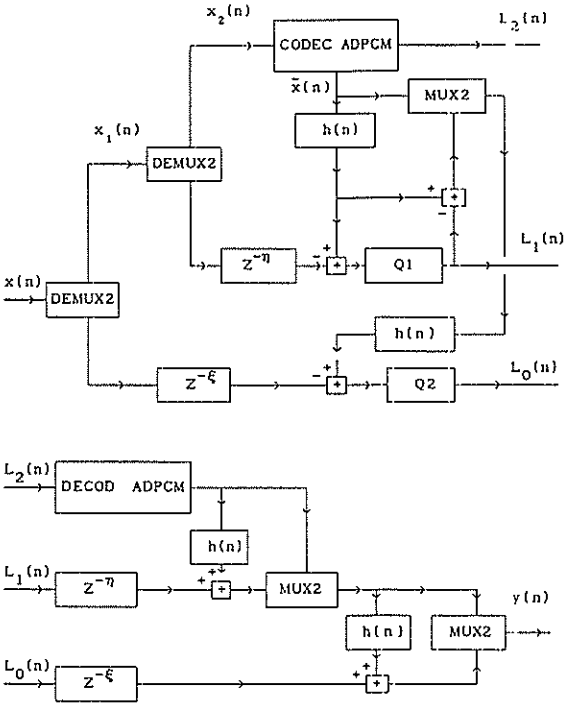


FIGURE 2 - PYRAMID DPCM TRANSMITTER AND RECEIVER

iii) Interpolate $x_1(n)$ and let Lo(n) be formed by the errors of this interpolation and x(n). Do not code the errors corresponding to samples represented in $x_1(n)$.

iv) Interpolate $x_2(n)$ and let L1(n) be formed by the errors of this interpolation and $x_1(n)$. Do not code the errors corresponding to samples represented in $x_2(n)$.

In Figure 2, the Pyramid DPCM structure is presented. Note that, in DPCM CODEC, locally decoded values of its input [1] are used. Note, also, that this strategy is also applied to the construction of lower levels. In this figure, the filters are the interpolators and DEMUX2 and MUX2 are devices that divide and reconstruct, respectively, their input samples (even-n samples for one branch and odd-n for the other).

Being α,β,γ the mean bit rates for L2's DPCM, L1, Lo, the global bit rate produced by the Pyramid DPCM is given by :

$$R = (\alpha + \beta + 2\gamma) \frac{1}{4} \qquad (E2)$$

3) EXTRA/INTERPOLATIVE PREDICTORS

Let $\hat{x}(n)$ be the predicted value of x(n) in DPCM. With $\mathbb{A}$ and $\mathbb{X}$ as filter and input vectors, we have:

$$\hat{x}(n) = \mathbb{A}^t \mathbb{X} \qquad (E3)$$

$\mathbb{X}^t$=[x(n-1)] ; $\mathbb{A}$=[1] ;

$\mathbb{X}^t$=[x(n-1) x(n-1-L) x(n-L)] ; $\mathbb{A}$=[1 -0.7 0.7] ;

$\mathbb{X}^t$=as in [4] ; $\mathbb{A}$=[1 -1/2 1/2 -1/2 1/2] ;

$\mathbb{X}^t$=[x(n-1) x(n-2) x(n-3)] ; $\mathbb{A}$= FLS adaptive ;

Where L is the sample length of one line in a field and these sets $\mathbb{X}$-$\mathbb{A}$ represent Past Sample, Intrafield, Interfield [4], and FLS adaptive [5] prediction approaches, respectively.

The interpolator here used is a discrete extension of the Cubic Convolution Kernel [6] [7] due to its extreme simplicity (even considering its 7 coefficients), adequated polyphase structure [8] and good performance. According to the sampling rate of an upper level, we have a FIR filter interpolator given by its impulse response as

$h(0)=1 \quad h(\pm2)=0 \quad h(\pm1)=9/16 \quad h(\pm3)=-1/16$

$$x_{k-1}(2n+1) = (x_k(n)+x_k(n+1))\frac{9}{16} - (x_k(n-1)+x_k(n+2))\frac{1}{16} + L_{k-1}(n) \qquad (E4)$$

for k=1 and 2, n=0,1...,$N/2^k$ (into one line).

4) REGULAR x PYRAMID DPCM

By comparing the standard DPCM with the Pyramid DPCM, we may say that the main advantages of the later are :

a)Better data compression, due to (i) the prediction gain of interpolation over extrapolation and (ii) to the fact that is possible to adopt three distinct quantization procedures, one for each level, optimizing coding and achieving improved subjective performance in comparison to the regular scheme.

b)Enlargement of processing interval between samples. Those intervals are 4 times longer in L_2 and L_1 and 2 times longer for L_0. (See Figure 1).

c)Facility to extend to a Multiresolution approach by conditional progressive transmission

The main disadvantages rise from the needs for apropriated logic to multiplex those levels and for the addiction of parallel computation to conventional DPCM. For comparisons, in a first step we evaluated the error entropies. Let p(k) be the probability of X=k ($k \in \mathbb{K}$); the entropies, here considered, are given by :

$$H[X] = -\sum_{i\in\mathbb{K}} p(i)\log_2 p(i) \qquad (E5)$$

$$H0 = H[x(n)] \qquad (E6a)$$

$$H1 = H[x(n)-x(n-1)] \qquad (E6b)$$

$$H4 = H[x(n)-x(n-4)] \qquad (E6c)$$

$$H_{L1} = H[L_1(n)] \qquad (E6d)$$

$$H_{L0} = H[L_0(n)] \qquad (E6e)$$

Now, in order to compare the full-image and L_2 DPCM coding for past-sample prediction, we must compare H1 and H4. However, the entropy gain (G) must also take into account H_{L0} and H_{L1}. In table I, the results of tests over 3 images are presented leading to a mean gain around 0.2 b/pel. If prefiltering is permitted, as in TABLE II, the mean gain rises to 0.5 b/pel. This prefiltering will improve interpolation, eliminating aliasing, but it will slightly corrupt the samples in L_2.

$$G = H1 - \frac{H4+H_{L1}+2H_{L0}}{4} \qquad (E7)$$

The prefilter used was an optimal [9] 15-tap FIR filter. Similar results were found with a 23-tap Hanning-weighted FIR filter. Repeating the process for Intrafield, Interfield and Intraline FLS-adaptive prediction approaches, it was verified that the mean gain decays with predictor's improvement. From 0.4 b/pel (Intrafield) to 0.2 b/pel (Interfield and FLS) with prefiltered inputs. However, with unfiltered inputs, the mean gain decays from 0.15 (Intrafield) b/pel to 0.07 b/pel (Interfield).

For coding simulations, we used exactly the same DPCM for L_2 and for the full-image. This includes adaptive two-dimensional prediction and fixed 31 level scalar quantizer. For the prediction equations we used intrafield prediction updated by a 2D LMS algorithm [10] with $\mu=0.1/255^2$.

$$\mathbb{A}(n+1) = \mathbb{A}(n) + \mu\,(x(n) - \hat{x}(n))\,\mathbb{X} \qquad (E8)$$

$L_0(n)$ was coded with a fixed 7- or 9-level quantizer, and L_1 with a 11-level fixed quantizer. Since we are invoking scalar quantization with Huffman coding, the lower bound for each quantizer mean bit rate is $1+\varepsilon$ b/pel. Figures 3–5 show comparative details extracted from reconstructed images. In these, the upper left quarter (UL) is extracted from original image; upper right (UR): (L_0:7 levels); Bottom Left (BL) : (L_0:9 levels); Bottom Right (BR): Standard DPCM. Those results are sumarized in TABLE III, indicating overall SNR and global mean bit-rate. In the comparative images, the high-quality reconstruction of images can be directly inferred, since the processed images are practically undistinguishable from the original, but Pyramid Coding required lower bit-rate.

5) CONCLUSION

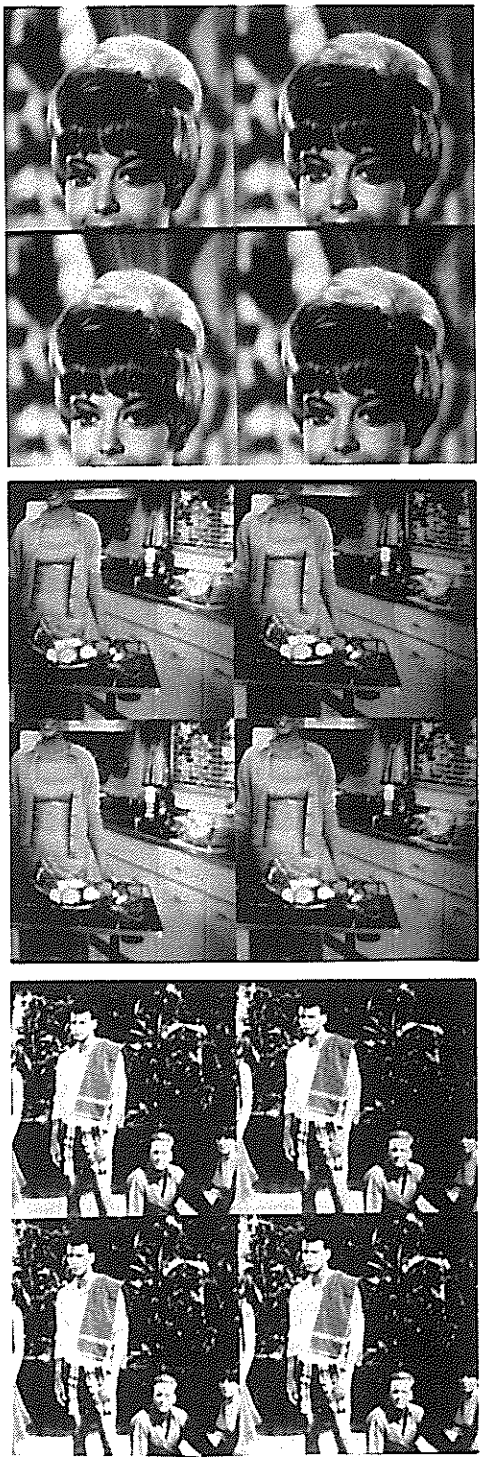
We tried to propose Pyramid DPCM as an alternative to conventional DPCM in high sampling rate coding environments. One point that must be strongly enphasized is that the results here achieved are too far from optimum. They are relevant when compared with standard DPCM under the same conditions, situation into which Pyramid scheme reveals to be an attractive alternative combining Sub-Band, Pyramid and DPCM coding for achieving a superior performance.

TABLE I

IMAGE	H0	H1	H4	H_{L0}	H_{L1}	G
BEACH	7.33	4.61	5.97	3.49	4.79	0.18
ZELDA	6.97	3.40	4.93	2.16	2.81	0.38
KITCH	6.86	3.56	4.90	2.59	3.56	0.15

TABLE II

IMAGE	H4	H_{L0}	H_{L1}	G
BEACH	5.84	2.91	4.45	0.40
ZELDA	4.86	1.50	2.42	0.67
KITCH	4.78	1.98	3.19	0.42



FIGURES 3,4,5 - COMPARATIVE IMAGES

TABLE III:Bit-Rate(b/pel);SNR(dB)

IMAGE	UR PYR 1	BL PYR 2	BR DPCM
ZELDA	1.45/40	1.85/43	2.4/46
KITCH	1.8/45	2.2/49	2.5/54
BEACH	2.5/36	2.9/38	3.2/39

REFERENCES

[1] N.S.Jayant, P.Noll, *Digital Coding of Waveforms,* Englewood Cliffs, NJ, Prentice-Hall, 1984.

[2] P.J.Burt, E.H.Adelson, *The Laplacian Pyramid as Compact Image Code*, IEEE Trans Commun., COM-31, pp 532-540, April 1983.

[3] L.Wang,M.Goldberg,*Reduced- difference Pyramid,* Optical Engineering, Vol 28 #7, July 1989.

[4] CCIR Study Groups : Doc.CMTT/45E - July 1987.

[5] M.Bellanger, *Adaptive Digital Filters and Signal Analysis,* Marcel Dekker Inc., NY,1987.

[6] R.G.Keys, *Cubic Convolution Interpolation For Digital Image Processing,* IEEE Trans on ASSP, ASSP-29, pp 746-749, June 1983.

[7] R.L.Queiroz, J.B.Yabu-uti, *Técnicas de Projeto de Filtros FIR para Dizimação e Interpolação de Sinais Discretos*,RT-178, Contract 208/89, Tech. Report UNICAMP/ TELEBRÁS, July 1989.

[8] R.E.Crochiere, L.R.Rabiner, *Multirate Digital Signal Processing.* Englewood Cliffs, NJ, Prentice-Hall, 1983.

[9] H.S.Malvar, D.H.Stealin, *Optimal Pre- and Postfillters for Decimation and Interpolation of Random Signals*, IEEE Trans. on Comm., pp 67-73, January 1988.

[10] M.Hadhoud, D.Thomas, *Two-D Adaptive LMS Algorithm,* IEEE Trans.on Circuits and Systems, CAS-35, May 1988.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Performance Evaluation of Hierarchical Coding Schemes for HDTV

Frank Bosveld, Reginald L. Lagendijk and Jan Biemond.

Delft University of Technology, Department of Electrical Engineering
Information Theory Group, P.O. Box 5031, 2600 GA Delft, The Netherlands

The Broadband Integrated Services Digital Network (BISDN) based on lightwave technology is supposed to become the all purpose exchange area communications network of the future. All digital services are integrated with applications ranging from facsimile, videophone, teleconferencing to Extended Quality TV (EQTV, signals according to the CCIR rec. 601) and High Definition TV (HDTV) distribution. This paper evaluates two progressive hierarchical subband coding schemes for HDTV, the so-called Refinement system and the Selection system. These schemes code the HDTV signal for the 135 Mb/s channel of BISDN, where the EQTV signal is coded for a part of this channel, for example 45 Mb/s. The effect of progressively coding the HDTV signal is shown for various sequences by experimentally obtained rate distortion functions and by allocating different bandwidths to the EQTV signal. Further, several system parameters are examined such as the optimal QMF filter and the structure of the optimal signal decomposition. Also the performance of the subband decomposition is compared to the performance of a DCT decomposition of the HDTV signal.

1 Introduction

During the last few years, there has been an increasing interest in bandwidth reduction of digital video sequences for the Broadband Integrated Services Digital Network (BISDN) [1,2,6,7,9,10]. This future network - based on lightwave technology, using ATM and intelligent networking - will become the all purpose exchange area network. BISDN will integrate both interactive and distribution services. The first services will consist, for instance, of person to person communications (i.e. voice, videophone and data), person (or group) to group communications (i.e. videoconference) and videosurveillance at bit rates below 2 Mb/s. The distribution services include TV distribution, i.e. signals according to CCIR recommendation 601 - sometimes refered to as Extended Quality TV (EQTV) - and High Definition TV (HDTV) distribution at bit rates of respectively 45 and 135 Mb/s.

Because the uncompressed bit rates of the forementioned video distribution services are approximately 216 Mb/s and 1.2 Gb/s, a datacompression/reduction is necessary. One possible way to achieve this is to compress and transmit both signals separately on the BISDN network. However, additional bandwidth for the EQTV signal is needed. This can be avoided by coding the EQTV signal as a subset of the HDTV signal, which is possible because the EQTV frequency band lies within the HDTV frequency band. As a result, the HDTV signal is coded in 135 Mb/s with the EQTV signal coded in a subset of 45 Mb/s.

In such a *hierarchical coding scheme* the HDTV signal consists of the EQTV signal along with high frequency HDTV information, called the HDTVcomplement (HDTVc), enhancement or surplus signal. The EQTV signal is obtained from the HDTV signal by two dimensional lowpass filtering and a 2:1 subsampling together with an adjustment of the aspect ratio from 16:9 to 4:3 [1,2,9]. However, through the hierarchical relation between the signals, the compressed EQTV signal at 45 Mb/s causes severe coding artifacts in the reconstructed HDTV signal at 135 Mb/s. Therefore, the EQTV signal needs to be coded progressively, which means that also some refinement information about the coded EQTV signal is transmitted in the coded HDTVc signal. An advantage of the above mentioned division is that it adapts well to ATM networks. Namely, the more important EQTV signal packets can be transmitted with a higher priority than the less important HDTVc signal packets. This guarantees in the case of cell loss, resulting from network congestion or cell misdelivery, a reasonable (low pass) reconstruction of the HDTV signal.

Bellisio [1] reviews three papers on progressive hierarchical coding. The first scheme [7] uses a Laplacian pyramid to obtain the EQTV and HDTVc signals and is attractive because its implicit feedback of the EQTV coding errors into the HDTVc signal. The second scheme [6] is based on a QMF subband decomposition into 4 subbands. The lowest subband defines the EQTV signal. This scheme is not progressive but the possibility of progressive coding is mentioned [1,2]. The HDTVc signal consists of the remaining 3 high frequency subbands together with some possible refinement information. Finally, the third scheme [2] uses a block-oriented $n \times n$ DCT. The EQTV signal is defined in the frequency domain as the $\frac{n}{2} \times \frac{n}{2}$ lower frequency coefficients. The EQTV signal is progressively coded by refining the EQTV coefficients. Together with the higher frequency coefficients these are the components of the HDTVc signal.

Biemond *et al.* [9] give a theoretical analysis of the concept of progressive coding based on the rate distortion theory. Further two intraframe progressive subband coding schemes are presented, called the Refinement system and the Selection system. Progressive coding is achieved in the Refinement system by using difference subbands, which contain the quantization errors of the EQTV subbands. These difference subbands are inserted into the HDTVc signal. The Selection system codes the EQTV signal progressively by optimally selecting coded EQTV subbands.

In this paper we evaluate the Refinement system and the Selection system for several bit rates and sequences. In Section 2 the two coding schemes are described. The (progressive) coding of the HDTV and the EQTV signal for both systems is evaluated in more detail in Section 3. In Section 4, we investigate the various system parameters, such as the optimal SBC splitting filter and the optimal signal decomposition.

2 Two hierarchical coding schemes

Refinement system

The basic principle of the Refinement system is shown in Figure 1. The HDTV signal, which has a 16:9 aspect ratio, is divided into 28 subbands using a tree-structured subband de-

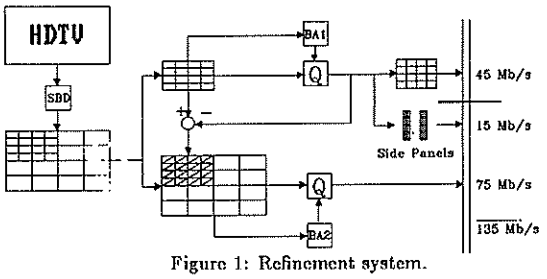


Figure 1: Refinement system.

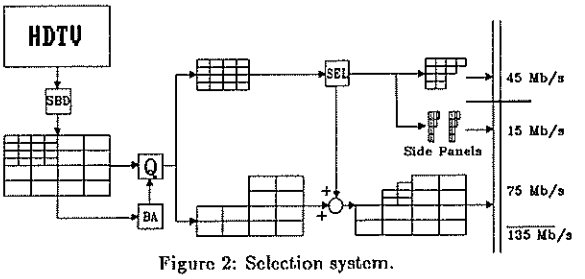


Figure 2: Selection system.

composition (SBD). Because the EQTV signal is defined as the filtered 2:1 decimated HDTV signal, the lower 16 subbands (upperleft corner) define the EQTV signal (excluding side panels). The remaining 12 subbands are the high-frequency subbands of the HDTV signal. The Refinement system has a two-stage coding structure. First the 16 EQTV subbands are quantized (Q) under supervision of a bit allocation (BA1), that optimally distributes the available bits over the subbands through selecting quantizers from a set of possible Max-Lloyd quantizers. These quantizers are optimized for the Generalized Gaussian (GG) PDF with a shape parameter c=0.5, which reasonably describes the PDFs of the subbands [4]. As described in [4] the lowest subband of the EQTV signal is DPCM coded, while the other subbands are PCM coded. Because the EQTV subbands have an aspect ratio of 16:9 instead of 4:3, the bit allocation codes these subbands at a bit rate of $\frac{16}{12} \times 45 = 60$ Mb/s. This results in homogeneously coded EQTV subbands which guarantees a homogeneous reconstruction of the HDTV signal. The extra 15 Mb/s are necessary for coding the side panels, which are stripped of the EQTV signal before transmission.

Subsequently the coding errors of the EQTV signal are inserted into the HDTVcomplement signal by using difference subbands, containing the differences between the original and quantized subband values. Besides these difference subbands, the HDTVc signal also contains the 12 high frequency HDTV subbands. New Max-Lloyd quantizers are derived for the difference subbands [8] and are used in the second coding stage. Next, the second bit allocation (BA2) distributes the available bits for the HDTVc signal by optimally selecting the quantizers for each of the subbands. The HDTVc signal is now coded in 75 Mb/s. Together with the side panels (15 Mb/s) and the EQTV signal (45 Mb/s), the entire HDTV signal is coded in 135 Mb/s.

Selection system

The second system proposed in [9] is the Selection system, which is designed with two goals. First, the HDTV signal has to be coded optimally and secondly the system must have a low complexity. The basic idea of the Selection system is shown in Figure 2. The HDTV signal is again divided into 28 subbands which are all quantized (Q) under supervision of *one* bit allocation (BA). This guarantees an optimally coded HDTV signal that fits into the 135 Mb/s channel. The EQTV signal has to be constructed from the 16 coded EQTV subbands, each having an aspect ratio of 16:9. In general the bit rate of these 16 subbands exceeds 60 Mb/s (i.e. 45 Mb/s EQTV signal plus 15 Mb/s side panels). This, of course, is precisely the reason why progressive coding is required. We therefore can transmit only a restricted number of EQTV subbands in the 45 Mb/s channel, while the remaining EQTV subbands are transmitted in the 90 Mb/s channel (SEL).

The MSE distortion in the reconstructed EQTV signal is the summation of all MSE distortions in each subband [4,5]. If the i^{th} subband is not selected, the MSE distortion in the reconstructed EQTV signal equals the variance σ_i^2 of the subband. If subband i is selected, using r_i bits of the available bit rate, the distortion equals the quantization error variance $\sigma_{q,i}^2$ of the subband. The relative decrease in MSE is thus given by measure m_i [4]

$$m_i = \frac{\sigma_i^2 - \sigma_{q,i}^2}{r_i}. \qquad (1)$$

We select successively those subbands with the largest values of m_i until the sum of the bit rates of the selected subbands exceeds 60 Mb/s. The EQTV subbands which are not selected, are transmitted in the HDTVc signal. Before transmission of the EQTV signal, the side panels are stripped off (15 Mb/s). As a result the EQTV signal is coded in 45 Mb/s and the entire HDTV signal in 135 Mb/s.

Necessity of progressive coding

Biemond *et al.* [9] make use of the rate distortion theory to show that the two-stage coding structure of the Refinement system always provides an optimally coded EQTV signal. For certain bit rates also an optimally coded HDTV signal can be obtained. Figure 3 shows the bit rates of the two signals and is divided into two parts through bound (b). Below bound (b) the HDTV signal is coded optimally, if and only if the quantization errors of the EQTV signal are requantized (progressive coding). Above bound (b) the coding of the HDTV signal is always suboptimal because relatively too many bits are used to code the EQTV signal. As a result the EQTV quantization errors are not requantized anymore. We observe that the available bit rates for the EQTV and HDTV signals in BISDN, i.e. point (a), are beneath the bound and thus both signals can be coded optimally with the Refinement system. For the Selection system the same figure applies, but in this case the HDTV signal is always coded optimally. It is the EQTV signal which is only coded optimally at point (d). Here the bandwidth, allocated to the EQTV subbands through the HDTV coding stage, just equals the EQTV channel capacity. At lower bit rates the HDTV bit allocation allocates more bits to the EQTV signal than the EQTV channel capacity while at higher bit rates the opposite holds. In the next section we will take a closer look at these coding effects.

3 Signal coding

In this section we evaluate the (progressive) coding of the EQTV and HDTV signal using experimentally retrieved rate distortion functions. The rate distortion functions are obtain by coding a cut out of 288 × 480 pixels of the following three YUV sequences; 'Teeny', 'Costgirls' and 'Car'. The coding is intrafield, except for the non-interlaced 'Teeny' sequence where the coding is intraframe. Unless stated otherwise, all experiments are done using a 28 subband signal decomposition and a QMF splitting filter of 32 taps.

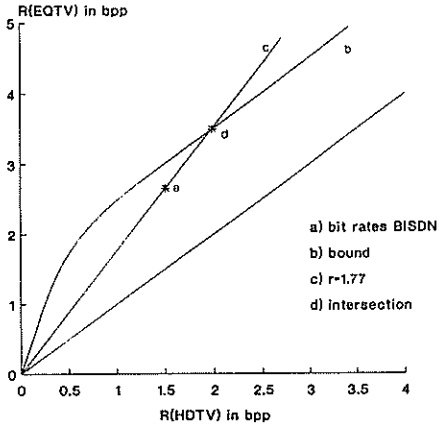


Figure 3: Bit rate diagram.

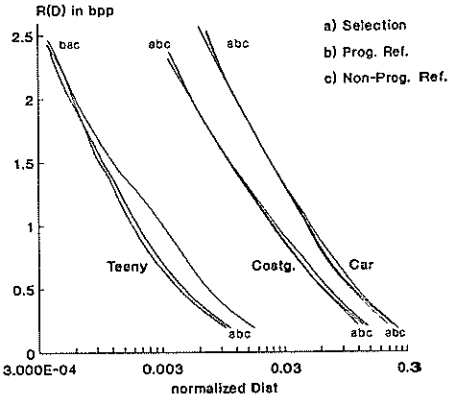


Figure 4: HDTV rate distortion curves.

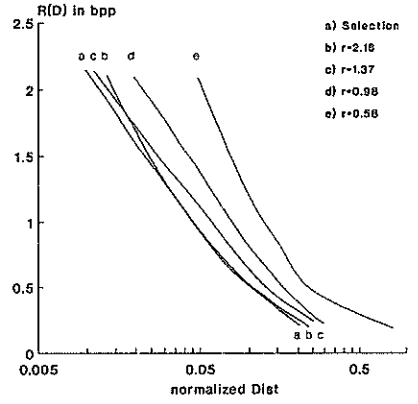


Figure 5: Non-progressive HDTV RDCs for various r.

Progressively coding HDTV

From Section 2 it follows that there are three possible ways of coding HDTV, namely HDTV coding with a) the Selection system, b) the progressive Refinement system and c) the non-progressive Refinement system, i.e. no feedback of the EQTV quantization errors. For the three test images, each rate distortion function is plotted in Figure 4, resulting in three trios of RDCs. The most left trio of RDCs is obtained from 'Teeny'. As can be seen, the performance of the HDTV coding of the Selection system (a) is superior to the other two RDCs. A moderate reduction in distortion (approx. 2.3 dB) is achieved by the RDC of the progressive Refinement system (b) with respect to the RDC of the non-progressive Refinement system (c), especially at low bit rates. At high bit rates, all curves closely approach each other, which confirms the theory that at higher bit rates the EQTV quantization errors are no longer requantized. The trios of RDCs of 'Costgirls' and 'Car' have larger distortions than the RDCs of 'Teeny' for identical bit rates. The above described effects, however, are also present in these trios but they are very modest. The maximum difference in distortion between the RDCs of the progressive and non-progressive Refinement system is about 0.6 dB. This results from the fact that the autocorrelation is smaller than the autocorrelation of Teeny ($\rho_h, \rho_v = 0.98, 0.99$ versus $0.90, 0.74$), which implies that the bits are distributed more uniformly over the subbands. Hence, less bits need to be allocated to the EQTV subbands, making the necessity for progressive coding decrease. For the 'Teeny' sequence the progressive Refinement system actually outperforms the HDTV coding of the Selection system because the performance of the latter is restricted through a maximum number of quantizers levels (128).

For the 'Car' image, we evaluated the effects of allocating different bandwidths to the EQTV signal on the reconstruction quality of the HDTV signal. Since the performance of the progressive Refinement system is independent of the EQTV bandwidth, only the performance of the non-progressive Refinement system is shown. In Figure 5 the EQTV signal is coded for respectively 15/25/35 and 55 Mb/s while the HDTV signal is coded at 135 Mb/s. The RDCs are obtained by varying the bit rates, keeping the bit rate ratio $r = \frac{R_{EQTV}}{R_{HDTV}}$ fixed at respectively 0.58/0.98/1.37 and 2.16. For comparison purposes also the optimal HDTV coding by the Selection system (a) is shown. By allocating smaller bandwidths to the EQTV signal, i.e. we walk across the RDCs, the distortion in the non-progressively coded HDTV signal increases. The behaviour of each RDC separately, can be explained with Figure 3 where line (b) has an r=1.77. The non-progressive RDC resulting from this line, as shown in Figure 4(c), is suboptimal at low bit rates, equals the optimal HDTV curve at point (d) and becomes again suboptimal at higher bit rates. The rate distortion curve in Figure 5, obtained with r=2.16, shows this behaviour more explicitly, while the RDC, obtained with r=1.37, intends to touch the optimal RDC (a). The RDCs, obtained with $r < 1$, clearly do not intend to intersect the optimal curve, which is affirmed in Figure 3 where in this case the bound and the r-lines diverge.

EQTV coding

The Refinement system always codes the EQTV signal optimally, i.e. with one bit allocation and one coding stage. The coded EQTV signal of the Selection system, however, is coded suboptimal because only a restricted number of subbands is used in the reconstruction. Therefore, the RDC of the Selection system must have larger distortions at identical bit rates than the RDC of the Refinement system. The experimentally obtained rate distortion functions of the three test images are plotted in Figure 6. As expected, the RDC of the Selection system (b) is situated at the right hand side of the RDC of the Refinement system (a). At high bit rates the RDCs equal each other because all subbands are selected. However, the coding of the Selection system is still suboptimal because not all available bits are used. As opposed to the HDTV RDCs in Figure 4, the difference in distortion between the EQTV RDCs (a) and (b) are nearly independent of the ACF of the test images. This is due to the fact that the quantization errors of the selected subbands in the EQTV reconstruction of the Selection system are dominated by the distortion due to the non-selected subbands.

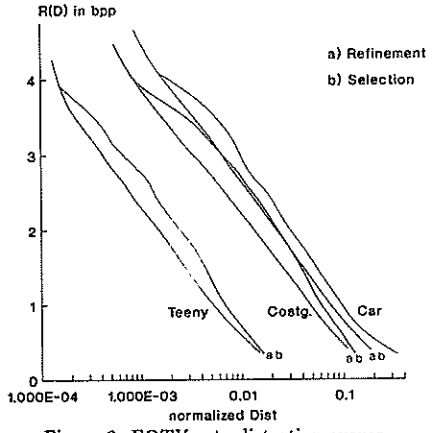


Figure 6: EQTV rate distortion curves.

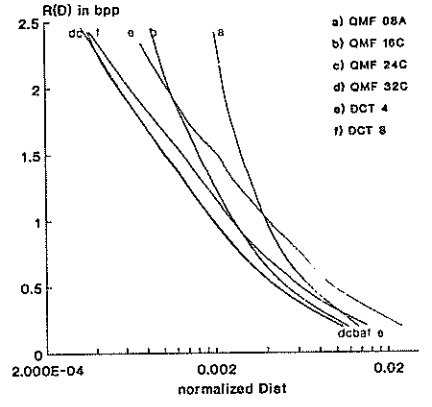


Figure 7: HDTV RDCs for various QMF filters.

4 System parameters

Optimal splitting filter

The following experiment investigates the optimal number of taps of the splitting filter. The splitting filters used are proposed in [3] and vary from 8 to 32 taps. Due to space limitations only the RDCs of the coded HDTV signal of the Refinement system obtained from 'Teeny' are plotted in Figure 7. The RDCs of the filters 08A and 16C are clipped at a certain distortion, which is the result of their large QMF error as defined in [4]. The RDCs of the other QMF filters are all within a distortion range of 1 dB, with the best performance for the 32C filter. Also shown are the RDCs of two DCT decompositions with block sizes 4 and 8. After grouping the similar frequency coefficients of all blocks, these decompositions lead to a 16 and 64 band splitted signal respectively, which is treated similarly as the subband splitted signal. The DCT decomposition with block size 8 has the best performance of the two DCT RDCs, but is still worse than the RDCs of the QMF filters. To reach the same distortion as the best subband splitted signal, approximately 10% additional bits are needed. Further, the resulting *blocking* artifacts in the reconstructed signals of the DCT coded signals are more annoying than the *low pass* artifacts of the subband coded signals. For the other test images *all* subband RDCs are within a range of 1 dB.

Signal decomposition

In the final experiment we investigate the optimal signal decomposition, which involves both the coding of the HDTV signal and the coding of the EQTV signal. Figure 8 shows the RDCs of

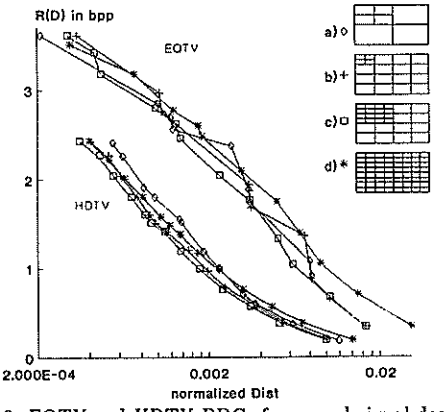


Figure 8: EQTV and HDTV RDCs for several signal decompositions.

the EQTV and HDTV signals obtained by the Selection system on 'Teeny' for several subband decompositions, which are also shown in Figure 8 (a),(b),(c). Also the 64 band decomposition of the 8 x 8 block DCT decomposition is shown (d). These decompositions have respectively 7/19/28 and 64 (sub)bands for the HDTV signal and 4/7/16 and 16 (sub)bands for the EQTV signal. The performance of the HDTV RDCs of the subband decompositions increases if the number of subbands becomes larger. This is due to the high ACF of 'Teeny' because for the other test images this performance gain is very modest. The RDC of the HDTV signal, obtained with the DCT decomposition, is inferior to the subband RDCs at low bit rates, but for higher bit rates the distance between the RDCs is reduced (1.4dB-0.6dB). The RDCs of the coding of the EQTV signal of the Selection system show the performance of the selection mechanism. The 7 subband decomposition result in a jumpy RDC which is very bit rate critical. This is due to the fact that only 4 subbands can be selected. More continuous RDCs are obtained with the 19 and 28 subband decompositions. The 64 band DCT decomposition provides a continuous RDC, but has larger distortions at identical bit rates than the subband RDCs.

References

[1] J.A. Bellisio and K.H. Tzou, "HDTV and the emerging broadband ISDN network", Proc. SPIE Visual Communications and Image processing Conf. '88, pp.772-786.

[2] K.H. Tzou and T. Chen, "Compatible HDTV coding for broadband ISDN", Globecom '88 Florida, Dec 1988, pp.743-749.

[3] Johnston,J.D., "A filter family designed for use in quadrature mirror filter banks", Proceedings IEEE conf. on ASSP, pp. 291-294, April 1980.

[4] P.H. Westerink, "Subband coding of images", Phd. thesis, Delft, Oct.1989.

[5] J.W. Woods and S.D. O'Neil, "Subband coding of images", IEEE trans. on ASSP. Vol ASSP-34, no.5, Oct. 1986.

[6] D.J. LeGall, H. Gaggioni and C.T. Chen, "Transmission of HDTV signals under 140 Mbits/s using a Sub-band Decomposition and Discrete Cosine Transform Coding", Signal processing of HDTV, Elsevier Science Publishers, pp 287-293, 1988.

[7] T.C. Chen, K.H. Tzou and P.E. Fleisher, "A hierarchical HDTV coding system using a DPCM-PCM approach", Proc. SPIE Visual Communications and Image processing Conf. '88, pp. 804-811.

[8] F. Bosveld, "Hierarchical subband coding of HDTV in BISDN", Msc thesis, Delft, Nov. 1989 (In dutch).

[9] J. Biemond, F. Bosveld and R.L. Lagendijk, "Hierarchical subband coding of HDTV in BISDN", ICASSP 90, April 1990, Albuquerque, New Mexico.

[10] D. Cai and X. Lin, "HDTV signal coding in broadband ISDN", Proceedings PCS, March 1990.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ANTIALIASING MEDIAN-TYPE FILTERS FOR IMAGE DECIMATION AND PROCESSING†

Irek DEFÉE and Yrjö NEUVO

Signal Processing Laboratory,
Department of Electrical Engineering,
Tampere University of Technology,
P. O. Box 527, SF–33101 Tampere, Finland.

A class of median-type filters which remove certain high-frequency periodic structures from signals and images is presented. These filters are obtained by taking combinations of median filters which have opposite phase shift properties for square-wave periodic signals.

1. INTRODUCTION

Median and median-type filters are widely used in edge preserving, noise-smoothing signal filtering. In image processing, where detail preservation is very important, multilevel and linear-median filters were successfully applied [1]. In many applications, such as TV signal processing, image coding and compression, video telephony etc., elimination of fine periodic structures is desirable since they are a cause of objectionable aliasing artefacts [2], [3]. Besides, there is a growing popularity of multiresolution techniques in pattern recognition and image analysis [4]. In these techniques, images are repeatedly decimated and the resulting data structure is called an image pyramid. Before the decimation is performed, lowpass filtering is necessary to prevent aliasing. Linear lowpass filters can be easily designed for the elimination of any signal components above specified frequency, but this is achieved at the cost of blurring edges which is objectionable. The design of linear filters for image decimation has been the subject of detailed investigations in order to find a compromise between good attenuation of aliasing artefacts, small blurring and small window size [5].

In this paper, we shall investigate designs and properties of median-type filters, which apart of edge preservation, have also capability for the removal of high frequency periodic signal structures. Periodic signal structures constitute roots of standard median filters and cannot be removed by them. These periodic patterns give rise to aliasing when signals are decimated. Filters proposed in this paper eliminate high frequency components and preserve sharp image edges. They may be thus better suitable for further processing tasks related to image decimation and processing.

The concept of an edge-preserving median filtering with periodic pattern removing ability is based on the observation that standard median filters operating on fine periodic signal structures produce outputs which are either out-of-phase or in-phase with an input signal, depending on the filter length. For the removal of the periodic structures we thus use a combination of median filters with two different lengths, which cancel periodic patterns in their output and preserve other features well. In the application of this concept to image processing, the filters must be designed to cope with periodic signal structures of different orientations.

We have designed and tested filter configurations based both on median and median-type filters.The filter properties were checked using numerical simulations on noisy test signals and on real noisy images. Several filter structures were selected having good ability for periodic pattern removal, good image detail preservation and noise attenuation. Generally, best results were obtained with multilevel median-type filters. Due to their simple structure, these filters can have potential applications in multiresolution image analysis where muliple image decimation is required, and in image coding problems.

2. PRELIMINARIES

2.1. Linear Decimation of Signals

When a signal with a highest frequency limited to $F_s/2$ is going to be decimated by a factor of M, its frequency content has to be limited to a value $F_s/2M$ since otherwise decimated spectra will overlap, giving rise to *aliasing* errors. To avoid, this, the signal must be lowpass filtered before the decimation, in order to remove all frequencies $F_s/2M \leq f \leq F_s/2$. In this process, an original signal $x(n)$ is convolved with the impulse response $h(n)$ of a linear filter, and the output is decimated as follows

†This research was supported in part by the Academy of Finland.

$$y(m) = \sum_{k=-\infty}^{+\infty} h(k)x(Mm-k) \qquad (1)$$

The process described by (1) introduces changes to the original signal, which sometimes cannot be accepted. In particular, edges present in the original signal are blurred by the linear weighted summation in (1). It must be stressed that single step-like, isolated edges don't give rise to aliasing since they are too localized, but these edges are inevitably blurred by (1). Edge blurring is especially undesirable in image processing since it significantly degrades preceptual quality of images. Edge information is also important in segmentation and pattern recognition problems, which often rely on decimation in order to reduce the computational load.

Good filters for image decimation should have small size, good out-of -band attenuation and low edge blurring. The design of FIR filters for image decimation has been systematically studied in [5]. The filters obtained there represent optimized linear solution to the decimation problem, taking into account all the conflicting requirements mentioned above.

2.1. Edge Preserving Lowpass Median-Based Filters

It is a well-known fact that standard 1-D median filter of length N=2k+1 removes signal elements which have length $\leq k$. Assume now that an input signal is a square wave of period 2 followed by a step edge and with added Gaussian noise, as shown in Fig. 1 a).

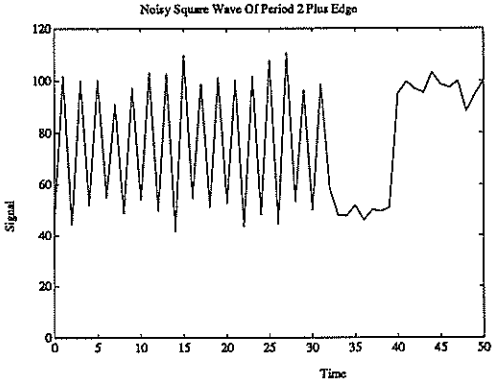


Figure 1 a)

In Fig. 1 b), the result of filtering of this signal by a 3-point median filter is shown, and in Fig. 1 c) the result of filtering by a 5-point median. Since the input signal is noiseless, the output signals remain unchanged (they are root signals of the filters). After closer inspection of the periodic parts of the median filter outputs we can see that the output of the 3-point median filter is shifted by half of the period and the output of the 5-point median is not shifted. By taking an average of both filter outputs, periodic part of the signal from Fig. 1 a) can be completely smoothed out, while the edge part is still preserved. This is shown in Fig. 1 d).

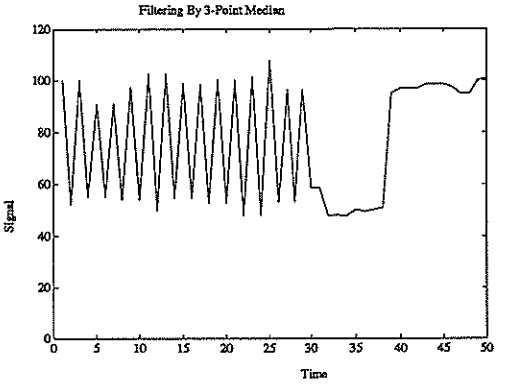


Figure 1 b)

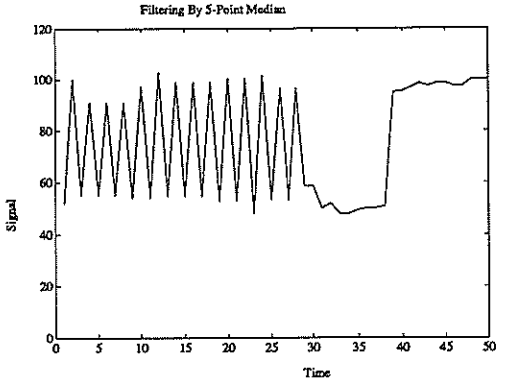


Figure 1 c)

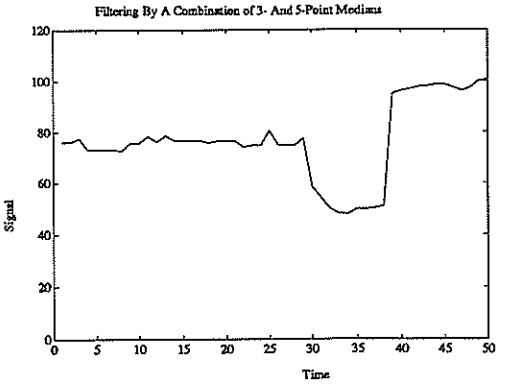


Figure 1 d)

The operation of taking an average of two median filter outputs which are shifted in phase is basic to our approach for edge-preserving periodic signal filtering

presented in this paper and it is shown schematically in Fig. 2. Since this structure can remove high-frequency components from a signal while preserving edges, it should be well-adapted for the decimation filters.

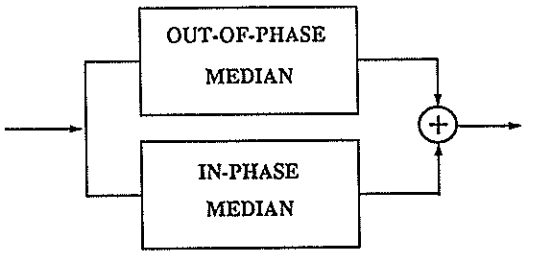


Figure 2

3. MEDIAN-BASED LOWPASS FILTER DESIGN

3.1. Basic Relations

We assume for simplicity that high frequency signal components which have to be removed from the decimated signal, are modeled by square waves with period 2, 4, ..., M-1. Decimation by M means removing from the signal all the square waves with periods $\leq M-1$. To achieve this, we use the general structure from Fig. 2 with median-type filters selected to produce phase shift required for the cancellation of the square waves.

We consider first the problem of the selection of median filters for the removal of a square wave signal with period 2K. One of the median-type filters in Fig. 2 should produce phase-shifted signal and the other one should preserve the signal unchanged. The following property is readily seen:

Property 1.

Minimal length of a median filter producing output signal shifted by half of a square wave with period 2K is equal to 2K+1. Median filters of length shorter than 2K+1 produce outputs which are not shifted with respect to the input square wave.

By this property, taking a median filter of length 2K+1 and a filter of length $\leq 2K-1$ in Fig. 2, ensures that a square wave with period 2K will be removed. This does not, however, guarantees that square waves with shorter periods will be removed from the signal too. A direct method for the removal of all square waves with any period less than 2K would be to filter the signal by a set of filters designed for the removal of one of the periods 2, 4, 8,..., 2K+1. Another possibility is to find one median structure which has the phase shift property for all periods. The following property is useful for finding such median filters:

Property 2.

The length L_s of a median filter producing shifted output signal for a square wave periodic input signal of period 2K is given by

$$(2s+1)2K+1 \leq L_s \leq 2(s+1)2K-1, s=0,1,\ldots \quad (2)$$

Property 3.

The length L_s of a median filter which preserves the phase of the input square wave signal of period 2K is given by

$$2s2K+1 \leq L_s \leq (2s+1)2K-1, s=0,1,\ldots \quad (3)$$

To check if a median filter shifts the phase of an input square wave, we need to check if there exists a number s that the (2) or (3) is satisfied. For example, median filter of length 3 inverts the phase of square waves of period 2 and median filter of length 7 inverts the phase of input square waves with period 2, 4, and 6. Thus, the median filter of length 7 inverts the phase of *all* lower-period sequences. However, in general finding a filter having such property is a nontrivial task. A median-type filter preserving the phase of square waves with any period 2, 4, 6, ... can be obtained by cascading 3-point median filters in a way shown in Fig. 3. This filter can be used in combination with phase-shifitng filters of any length.

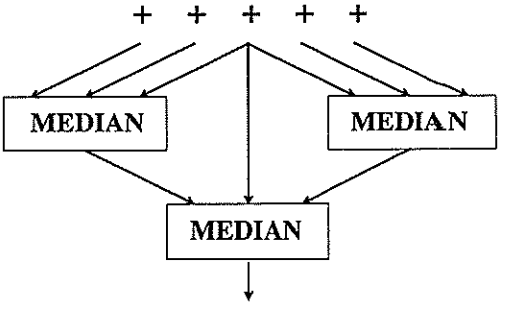


Figure 3

4. 2-D MEDIAN-TYPE LOW PASS FILTERS

The 2-D case is more difficult since the periodic square waves can distributed along horizontal, vertical and diagonal directions (Fig. 4). We shall now give several designs for filters with the capability for the removal of all the three types of periodic patterns from Fig. 4. First possibility is to use a cascade of 1-D filters described above and operating in vertical and horizontal directions. Such filters are easy to devise and realize for

the small decimation ratios (...2,3,4,) required usually in image processing. Another possiblity is to design genuine 2-D filters. A class of shaped and/or mulitlevel median-type filters can be used to obtain 2-D analogs of the phase shifting property for output signals. In the case of 2-D square waves of period 2 the design can be based on median-type filters operating in a 3×3 mask. One design is based on a 3×3 standard median filter and a cross-shaped vertical/horizontal multilevel median filter. Another design, which has better sensitivity to image details, is obtained by taking an average of the outputs of the cross- and diamond shaped multilevel median filters (Fig. 5).

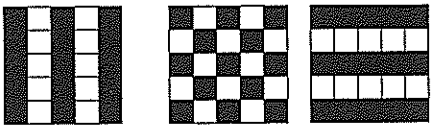


Figure 4

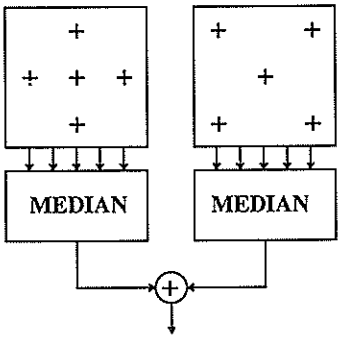


Figure 5

We have found that such combinations of a cross- and diamond-shaped median filters have many similar properties to the combinations of the 1-D median filters. From the point of image filtering/decimation where the window mask is required to be small, we found a combination of 7×7 cross- and diamond-shaped median filters to be especially suitable for practical applications. Experiments were performed on noisy images (Fig. 6) to check the properties of the 2-D lowpass 2-D median-type filters. Result of processing by a median-type lowpass filter in a 7×7 window is shown in Fig. 7.

6. CONCLUSION

Further research is required for the investigation of properties of the median-type filters with increased low-pass filtering.

REFERENCES

[1] A. Nieminen, P. Heinonen and Y. Neuvo, "A New Class of Detail-Preserving Filters for Image Processing," *IEEE Trans. Pattern Anal. Mach. Intelligence,* vol. PAMI-9, pp. 74–90, No.1, January 1987.

[2] P. Burt and E. Adelson, "The Laplacian Pyramid As a Compact Image Code," *IEEE Trans. Commun.*, vol. COM-33, pp. 532–540, No. 4, January 1983.

[3] M. Kunt, A. Ikonomopoulos and M. Kocher, "Second Generation Image Coding Techniques," *Proc. IEEE,* vol. 73, pp. 549–574, No. 4, April 1985.

[4] P. Meer, E.S. Baugher and A. Rosenfeld, "Frequency Domain Analysis and Synthesis of Image Pyramid Generating Kernels," *IEEE Trans. Pattern Anal. Mach. Intelligence,* vol. PAMI-9, pp. 512–522, No.4, July 1987.

Figure 6

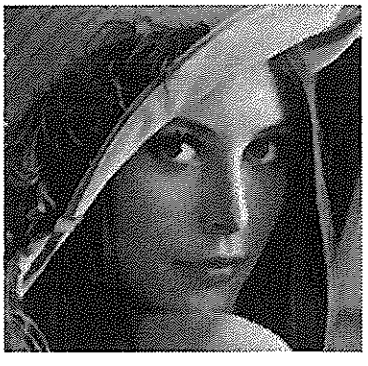


Figure 7

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

MARGINAL ORDER STATISTICS IN COLOR IMAGE PROCESSING

I. PITAS
Department of Electrical Engineering
University of Thessaloniki
Thessaloniki 54006, GREECE

ABSTRACT

This paper discusses the extension of order statistic filtering to color image processing. It gives the probability distributions of the so-called marginal order statistics and their use in multivariate location estimation. Especially the use of the marginal median in color image filtering is treated in detail.

1.INTRODUCTION

In any color coordinate system, each color is represented by a vector in a three dimensional space, e.g. by the vector $[R\ G\ B]^T$ in the CIE spectral primary system RGB [1,2]. Therefore, any color image can be considered as a two-dimensional vector sequence, or in other words, as a two-dimensional three-channel sequence.In the RGB and XYZ systems color chromaticity can be represented as a vector (r,g) or (x,y) in the two-dimensional space [1,2]. In cases of constant brightness, the chromaticity of a color image can be considered as a two-dimensional two-channel sequence. Therefore color image processing is essentially multichannel signal processing [3-5].

In the following, the use of ordering and order statistics [6] in multichannel signal and image processing will be investigated, with particular application in color image processing. Order statistics filtering and especially median filtering have found extensive applications in BW image filtering [6]. Therefore, there is a natural interest in extending them to multichannel signals. An introduction to the ordering of multivariate random variables is included in section 2. A particular type of ordering called marginal ordering will be used in the rest of this paper. The probability distributions of the marginal order statistics is also investigated in section 2. The use of the marginal order statistics as multivariate location estimators and their influence function will be described in section 3. Simulation examples of the use of the order statistics and especially of the median in color image processing will be described in section 4. Conclusions are drawn in section 5.

2. MARGINAL ORDERING IN MULTIVARIATE DATA

Let us denote by $\mathbf{X}$ a p-dimensional random variable, i.e. a p-dimensional vector of random variables $\mathbf{X}=[X_1,\ldots,X_p]^T$. We shall denote by $f(\mathbf{X})$, $F(\mathbf{X})$ the probability density function and the cumulative density function of this p-dimensional random variable respectively [7]. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n random samples from a p-dimensional distribution having mean $\mathbf{m}$ and dispersion matrix $\mathbf{\Sigma}$. The notion of data ordering can not be extended in a straightforward way in the case of multivariate data. An excellent treatment of the ordering of multivariate data was given by Barnett [8]. It is shown that there exist several ways to order multivariate data. There exist no unambiguous, universally agreeable total ordering of the n multivariate samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. The following so-called sub-ordering principles are discussed in the literature [8]: marginal ordering, reduced (aggregate) ordering, partial ordering and conditional (sequential) ordering.

In marginal ordering, the multivariate samples are ordered along each one of the p-dimensions:

$$\begin{aligned} &x_{1(1)} \leq x_{1(2)} \leq \cdots \leq x_{1(n)} \\ &x_{2(1)} \leq x_{2(2)} \leq \cdots \leq x_{2(n)} \\ &x_{p(1)} \leq x_{p(2)} \leq \cdots \leq x_{p(n)} \end{aligned} \tag{1}$$

i.e. ordering is performed in each channel of the multichannel signal. $x_{1(1)}, x_{2(1)}, \ldots, x_{p(1)}$ are the minimal elements in each dimension. $x_{1(n)}, x_{2(n)}, \ldots, x_{p(n)}$ are the maximal elements in each dimension. $\mathbf{x}_{(\nu+1)}=[x_{1(\nu+1)}, \ldots, x_{p(\nu+1)}]$ is the (marginal) median of the multivariate data for $n=2\nu+1$. The i-th marginal order statistic is the vector $\mathbf{x}_{(i)}=[x_{1(i)},\ldots,x_{p(i)}]^T$. Needless to say that the median or any i-th marginal order statistic may not correspond to any of the samples $\mathbf{x}_1,\ldots,\mathbf{x}_n$. In contrast, in the one-dimensional case there exists an one-to-one correspondence between the samples $x_1,\ldots,x_n$ and the order statistics $x_{(1)},\ldots,x_{(n)}$. In the following we shall concentrate on the probability distributions of the marginal order statistics.

The study of the probability distribution of marginal order statistics started relatively early [6] and the probability density function of the marginal median was investigated. In the following the cumulative distribution function and the probability distribution function of the marginal order statistics will be described. The two-dimensional case will be described first for simplicity reasons.

Let us denote by $F_{(r1,r2)}(x_1,x_2)$ the cdf:

$$F_{(r_1,r_2)}(x_1x_2) = P\{X_{1(r_1)} \leq x_1,\ X_{2(r_2)} \leq x_2\} \quad (2)$$

of the marginal order statistic $X_{1(r1)},X_{2(r2)}$ when n data samples are available. Let us also denote by $F_i(x_1,x_2)$, $i=0,\ldots,3$ the probability masses on the four regions of the plane defined by (x_1,x_2):

$$F_0(x_1,x_2) = P\{X_1 \leq x_1,\ X_2 \leq x_2\} = F(x_1,x_2)$$

$$F_1(x_1,x_2) = P\{X_1 > x_1,\ X_2 \leq x_2\}$$

$$F_2(x_1,x_2) = P\{X_1 \leq x_1,\ X_2 > x_2\} \quad (3)$$

$$F_3(x_1,x_2) = P\{X_1 > x_1,\ X_2 > x_2\}$$

$F(x_1,x_2)$ is the joint cdf of the random vector $\mathbf{X}=[X_1,X_2]^T$. The cdf $F_{(r1,r2)}(x_1,x_2)$ is given by:

$$F_{(r_1,r_2)}(x_1,x_2) = \sum_{i_1=r_1}^{n}\sum_{i_2=r_2}^{n} P\{i_1 \text{ of } X_{1i} \leq x_1,\ i_2 \text{ of } X_{2i} \leq x_2\} =$$

$$\sum_{i_1=r_1}^{n}\sum_{i_2=r_2}^{n}\sum_{n_0=\max(0,i_1+i_2-n)}^{\min(i_1,i_2)} \frac{n!}{n_0!(i_1-n_0)!(i_2-n_0)!(n-i_1-i_2+n_0)!}$$

$$\cdot F_0^{n_0}(x_1,x_2)F_1^{i_2-n_0}(x_1,x_2)F_2^{i_1-n_0}(x_1,x_2)F_3^{n-i_1-i_2+n_0}(x_1,x_2) \quad (4)$$

(4) is relatively complicated and does not give analytic expressions of $F_{(r1,r2)}(x_1,x_2)$ for arbitrary cdf $F(x_1,x_2)$. However it can be easily computed numerically. The probability density function $f_{(r1,r2)}(x_1,x_2)$ of the marginal order statistics can be easily found from (4):

$$f_{(r_1,r_2)}(x_1,x_2) = \frac{\partial^2 F_{(r_1,r_2)}(x_1,x_2)}{\partial x_1\,\partial x_2} \quad (5)$$

and can also be calculated by numerical differentiation.

The cdf of the three dimensional marginal order statistics can be found in a similar way. The three dimensional space is divided in eight subspaces by a point (x_1,x_2,x_3). These subspaces and the corresponding probability masses are shown in Figure 1. The cdf is given by:

$$F_{(r_1,r_2,r_3)} = \sum_{i_1=r_1}^{n}\sum_{i_2=r_2}^{n}\sum_{i_3=r_3}^{n}\sum_{n_0}\cdots\sum_{n_{2^3-1}} \frac{n!}{\prod_{i=0}^{2^3-1} n_i!} \prod_{i=0}^{2^3-1} F_1^{n_i}(x_1,x_2,x_3) \quad (6)$$

subject to the constraints

$$\sum_{i=0}^{7} n_i = n$$

$$n_0+n_2+n_4+n_6 = i_1$$

$$n_0+n_1+n_4+n_5 = i_2 \quad (7)$$

$$n_0+n_1+n_2+n_3 = i_3$$

The cdf (3), (7) can be used to derive the variance of the output of a marginal order statistics filter.

3. MARGINAL ORDER STATISTICS AS ESTIMATORS OF THE MULTIDIMENSIONAL LOCATION

One dimensional order statistics (especially the median) and their linear combinations, e.g. the a-trimmed mean and the L-filter [6] have extensively been used as estimators of the one-dimensional location. In fact the so-called L-estimators of location and scale are based on order statistics [9].

The definitions of L estimators can be easily extended to the p-dimensional case by using marginal order statistics. The following estimator will be called p-dimensional marginal L-estimator:

$$(8)\quad T_n = \sum_{i_1=1}^{n} \cdots \sum_{i_p=1}^{n} A_{i_1,..,i_p}\, \mathbf{x}_{(i_1,..,i_p)}$$

where $\mathbf{x}_{(i1,..,ip)}=[x_{1(i1)},...,x_{p(ip)}]^T$ are the marginal order statistics and $A_{i1,...,ip}$ are pxp matrices. The performance of the marginal L estimator depends on the choice of the matrices $A_{i1,...,ip}$. The marginal median, the marginal maximum $\mathbf{x}_{(n)}$ the marginal minimum $\mathbf{x}_{(1)}$, the p-dimensional marginal a-trimmed mean [7] and the arithmetic mean $\mathbf{x}$ are special cases of (8). The robustness of the L estimators in the presence of outlying data points impulses can be found by using the p-dimensional influence function [9]. Let F be the p-dimensional probability distribution of the data and T be the functional of an L estimator at this distribution. The influence function IF(x,T,F) measures the change of T caused by an additional observation at the point x:

$$(9)\quad IF(x,T,F) \triangleq \lim_{t\to 0} \frac{T[(1-t)F + t\Delta_x] - T[F]}{t}$$

If this change is bounded, the estimator has good robustness properties and an outlying observation cannot destroy its performance. Thus the robustness of an estimator can be measured in terms of its gross error sensitivity [9]:

$$(10)\quad \gamma^*(T,F) \triangleq \sup_{x} \{ || IF(x,T,F) || \}$$

where $||.||$ denotes the Euclidean norm. If it is finite, the estimator is called B-robust. It can be proven that, under some conditions [9], the estimator T is asymptotically normal and that its covariance matrix is given by:

$$(11)\quad V(T,F) = \int IF(x,T,F)\, IF(x,T,F)^T dF(x)$$

The marginal median is a B-robust estimator because its gross error sensitivity is finite [10]:

$$(12)\quad \gamma^*(T,F) = \frac{1}{2}\sqrt{\frac{1}{f^2_{x_1}(F^{-1}_{x_1}(\frac{1}{2}))} + \frac{1}{f^2_{x_2}(F^{-1}_{x_2}(\frac{1}{2}))}}$$

4. SIMULATION EXAMPLES

Two sets of experiments have been performed to assess the performance of the marginal order statistics in general and of the marginal median in particular. The first set of experiments deals with its ability to filter multichannel impulsive noise of the form:

$$(13)\quad s = \begin{cases} d & \text{with probability } 1-p \\ n & \text{with probability } p \end{cases}$$

where n is noise whose distribution function is much different from the distribution of d. The signal distribution is Gaussian of the form:

$$(14)\quad f(x,y) = \frac{300}{2\pi} \exp\{-\frac{1}{2}[200(x-m_1)^2 + 500(x-m_1)(y-m_2) + 500(y-m_2)^2]\}$$

with $m_1=m_2=0.33$. Such a signal corresponds to noisy white light, in the XYZ chromaticity diagram. The impulses have the same form (14) with different means $m_1=0.6$, $m_2=0.3$. Such a signal corresponds to red light in the XYZ chromaticity diagram. The impulse probability in the noisy signal of Figure 2 is 0.2. The cartesian data plot $(x_1(n), x_2(n))$ n=1,..N in Figure 2a reveals the existence of impulses. The two noisy signals $x_1(n)$, $x_2(n)$ n=1,..N cannot be shown due to lack of space. The results of the moving average are hopelessly bad and are not displayed here. The results of the median filter for n=5 are shown in Figures 2b in the cartesian plot $(x_{1(3)}(n), x_{2(3)}(n))$ n=1,..N. As it is clearly seen, several impulses exist after median filtering, especially along the x axis. This example illustrates that impulsive noise removal is much more complex in multichannel signal processing and perhaps more sophisticated techniques must be used.

The second set of experiments has been performed on color images corrupted by additive white noise having uniform and Gaussian distributions. It has been proven that the moving average filter outperfoms marginal median filter, especially for short-tailed distributions (uniform noise) [10]. Simulation examples are not presented here due to lack of color printing.

5. CONCLUSIONS

The theory of marginal order statistics filtering has been presented in this paper with applications in color image filtering. It has been found by simulation that the moving average filter behaves better than the median in short-tailed noise distributions. The marginal median performs better than the moving average filter in long-tailed and impulsive distributions. It was also shown that, in general, multichannel impulsive noise removal is more difficult than in the single channel case.

REFERENCES

[1] A.K.Jain Fundamentals of Digital Image Processing, Prentice Hall 1989.

[2] G. Wyszecki, W.S. Stiles Color Science, Wiley, 1967.

[3] B.R.Hunt "Karhunen-Loeve multispectral image restoration, part I: theory", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-32, No.3, pp.592-599, June 1984.

[4] N.P. Galatsanos, R.T. Chin "Digital restoration of multichannel images", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-37, No.3, pp.415-421, March 1989.

[5] G. Aggelopoulos, I. Pitas "Least-squares multichannel filters in color image restoration", Proc. European Conference on Circuit Theory and Design ECCTD89, Brighton, England, September 1989.

[6] I. Pitas, A.N. Venetsanopoulos Nonlinear digital filters: principles and applications, Kluwer Academic, 1990.

[7] G.A.F.Seber Multivariate Observations, J.Wiley, 1984.

[8] V.Barnett "The ordering of multivariate data", Journal of the Statistical Society of America A,vol.139, pt.3, pp.318-354, 1976.

[9] F.Hampel, E.Ronchetti, P.Rousseeuw, W.Stahel Robust statistics, John Wiley, 1986.

[10] I. Pitas "Marginal order statistics in color image filtering", Optical Engineering, May 1990.

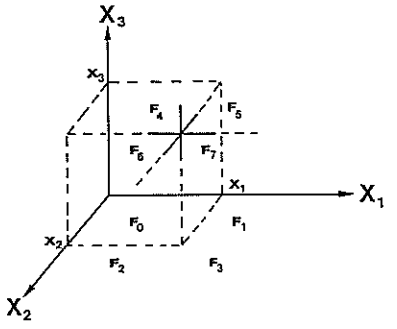


Figure 1: Regions used in the derivation of the cdf of the three-dimensional marginal order statistics.

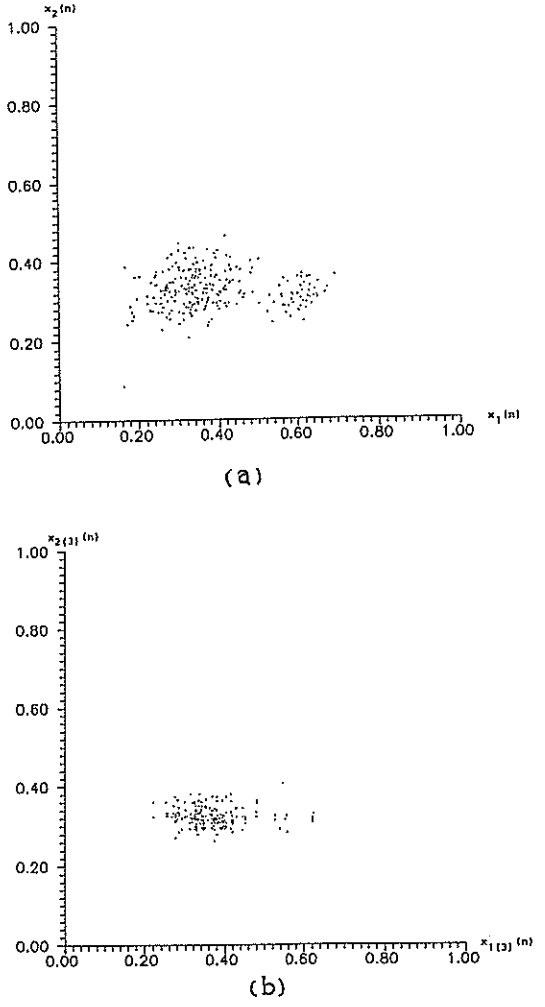


Figure 2: (a) Cartesian plot of the samples of the two-channel signal (b) Cartesian plot of the data points of the two-channel marginal median filter output.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

SYMMETRICAL RECURSIVE MEDIAN FILTERS APPLICATION TO NOISE REDUCTION AND EDGE DETECTION

Ph.BOLON , A.RAJI, P.LAMBERT, M.MOUHOUB

Laboratoire d'Automatique et MicroInformatique Industrielle
LAMII - Université de Savoie - BP 806 - F.74016 ANNECY CEDEX (France)
(CNRS-GDR 134 : Signal and Image Processing)

ABSTRACT

Median filters (MF) have the property to smooth noise and to preserve abrupt changes of intensity in an image. Recursive median filters (RMF) have the same sensitivity to abrupt changes. Their performances are improved as far as noise reduction is concerned. Hence, they can be used in order to build up gradient operators. However , some asymmetry is introduced by recursivity. In this paper, symmetrical recursive median filters (SRMF) are presented. Some of their statistical properties are discussed and compared to the ones of asymmetrical recursive median filters.
Then gradient operators using SRMF are introduced and their performances compared with the ones of a classical Canny-Deriche operator.

keywords : image processing, order filters, median filters, noise reduction, edge detection

I - INTRODUCTION

In the context of industrial vision, one has to localize and to recognize objects. In the case of noisy images, it is necessary to develop operators that reduce noise and preserve variations of intensity caused by the edges of objects. These operators should not produce any change in edge location or transition amplitude, since these parameters can be used by pattern recognition methods. Moreover, it is interesting to find local operators, i.e. operating on a small number of pixels. Firstly the implementation is easier. Secondly, it is possible to point out interesting structures that are small or close together.
Median filters (MF) [1] have the property to smooth noise and to preserve abrupt changes of intensity. However, if the noise distribution is bounded, or has a low kurtosis parameter, linear filters have better performances than median filters [2][3].
Some local operators based on median filtering were proposed for noise reduction [4] and edge detection [5][6].

Recursive median filters (RMF) have the same sensitivity to abrupt changes [7]. Their statistical properties were studied in the case of an impulse noise [8], or a continuous one with unimodal and symmetrical distribution. [9]. Their ability to reduce the noise is better than the one of ordinary 'transversal' median filters, and approaches the one of optimal order filters [9].
Therefore it is possible to build up a gradient operator based on recursive median filtering and differentiation (RMG operator). Its main characteristics are : good detection and good localization of edges, and simplicity of implementation. Since recursivity introduces asymmetry, we propose to use symmetrical recursive median filters (SRMF) for noise reduction and gradient operators based on symmetrical recursive median filtering (SRMG) for edge detection.

In this paper, performances of MF, RMF and SRMF are compared by studying their output variance.
Gradient operators are commonly characterized by means of a detection criterion and a localization criterion [10]. We then present some characteristic curves (probability of detection vs probability of false alarm). By studying the spatial distribution of the probability of detecting an edge, it can be shown that SRMG operator produces a better localization effect than a Deriche operator.

II - GRADIENT OPERATORS

In the case of noisy images, gradient operators are generally based on separable lowpass filtered derivatives [10].

Let X_{ij} denote the intensity at pixel : row i, column j.
The row gradient component Gx is obtained by :

$$Gx = L_C * L_R * \delta'_R \qquad \text{Eq.1}$$

with L : lowpass linear 1D filter
δ' : derivation operator
* : convolution product
Index C (resp. R) denotes column (resp. row) processing.

Component Gy is obtained by exchanging indexes C and R.

Choosing the lowpass filter bandwidth results from a trade-off between detection and localization of edges. Noise reduction (and good detection) is obtained with a lower bandwidth. However, if several edges are present in a region, spatial averaging produces a shift in their estimated location. Considering a detection-localization criterion, an optimal lowpass derivation linear filter can be found in the case of gaussian white noises [10].
Because of its noise reduction and edge preservation properties, we propose to use recursive median filtering instead of linear lowpass filtering (see Fig.1). Derivatives are then estimated by linear filtering with impulse response {+1, -1}.

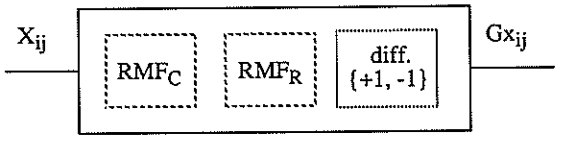


Fig.1 Gradient operator (row component)

Let X be the input image and $Y=RMF_C(X)$ and $Z = RMF_R(Y)$. Let N=2n+1 be the filter size, and 'med' the median operator. We have :

$$Z_{ij} = \text{med}(Z_{i\,j-n},...,Z_{i\,j-1},Y_{ij},...,Y_{i\,j+n}) \qquad \text{Eq.2}$$

It should be noticed that, unlike the linear case, the different processing steps cannot be commuted.
Deterministic and statistical properties can be studied in the case of a monodimensional signal.

deterministic properties : Signals composed of constant regions, joined together by monotonic transitions, are roots for median and recursive median filters [7].Therefore, step edges will produce unshifted peaks in the gradient signal. If we consider the effect of the presence of structures onto the gradient signal, the relevant parameter is not their amplitude, but rather their size. Structures having size lower than N/2 will have no effect on the gradient signal. The other ones will produce unshifted peaks, whose amplitude is equal to the edge amplitude, provided that the minimum distance between two of them is greater than n.

statistical properties : They can be analyzed by means of a detection criterion or a localization criterion. Fig.2 displays the probability of detection of a step edge (contrast h=3) at the right location (Pd) vs the probability to detect an edge in a stationary area (false alarm : Pf). It can be seen that

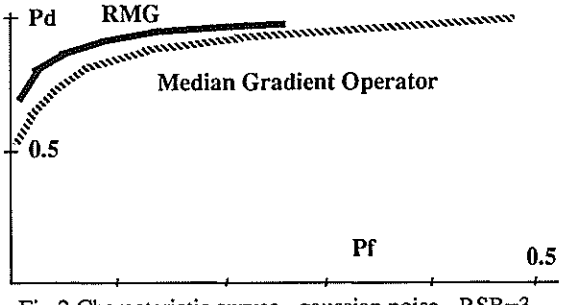


Fig.2 Characteristic curves - gaussian noise - RSB=3

recursivity improves the performances as far as detection is concerned. As for the localization criterion, Fig.6 shows that one can obtain a better edge localization than with a Canny-Deriche operator [10].

implementation : Fast algorithms for median filtering are available [11]. They can be adapted to recursive filtering [12]. The computation time does not depend on the filter size.

Recursive Median Gradient (RMG) operators have good performances. They can be implemented by means of fast algorithms. However, because of recursivity, they introduce some asymmetry in the processing. It can be seen on figure 6, that the spatial distribution of the probability of detection is not symmetrical with respect to the edge location. Symmetrical recursive median filters can be used in order to cope with this problem.

III - SYMMETRICAL RECURSIVE MEDIAN FILTERS

Since we are using separable filters, we can consider the case of 1D signals.
Let X_k be the input signal. Let N=2n+1 be the filter size. Let 'med' denote the median operator.
We have :

$$YF_k = \text{med}\,(YF_{k-n},..YF_{k-1},X_k,...,X_{k+n}) \quad k=kmin..kmax$$
$$YB_k = \text{med}\,(X_{k-n},..,X_k,YB_{k+1},..,YB_{k+n}) \quad k=kmax..kmin$$

and the output of the symmetrical recursive median filter (SRMF) is defined by :

$$Y_k = (YF_k + YB_k)/2 \qquad \text{Eq.3}$$

In this section, the differences between RMF and SRMF operators are discussed, from the deterministic and statistical point of view.

deterministic properties : Since signals composed of constant regions, joined together by monotonic transitions, are roots for recursive median filters, they are roots for SRMF operators. RMF and SRMF produce different results in the case of signals composed of constant regions separated by oscillation areas. Let s_k be such a signal, we have :

$s_k = 0$ k=kmin..k0 $s_k=h$ k=k1..kmax
s_k oscillating k=k0+1 .. k1-1

Let yf_k (resp. yb_k) be the output of the forward (resp. backward) filter, and y_k the SRMF output. We then have :

$yf_k=0$ k=kmin..k1-1 $yf_k=h$ k=k1..kmax
$yb_k=0$ k=kmin..k0 $yb_k=h$ k=k0+1..kmax

and

$y_k=0$ k=kmin..k0, y $y_k=h/2$ k=k0+1..k1-1
$y_k=h$ k=k1..kmax

statistical properties : In this section, we study the output variance of SRMF operators, in the case of a unit variance white input noise. Fig.3-4 point out the influence of filter size N. These variances were experimentally computed, us-

ing a pseudo-random number generator. The number of samples used is Ns=10 000. Taking into account the correlation between successive filter outputs, it can be seen that the standard deviation of the variance estimates is less than 0.015. Curves displayed on Fig.3 represent the output variance of MF, RMF, and SRMF, in the case of a gaussian noise. The output variance of the optimal linear filter having size N (average filter) is 1/N. We can remark that, as soon as N ≥5, SRMF has its output variance less than 1/N.
As far as noise reduction is concerned, median filters are not suited to uniformly distributed input noises [3]. Nevertheless, it can be seen on Fig.4 that SRMF perform better than linear filters as soon as filter size N is greater than 9.

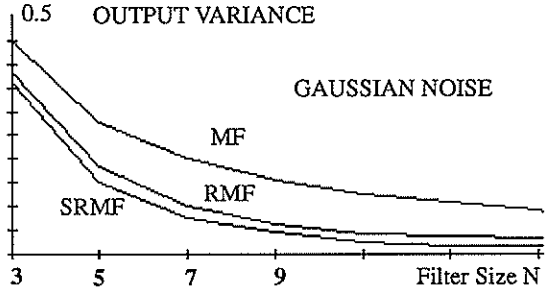


Fig.3 Output variance. Unit variance input noise

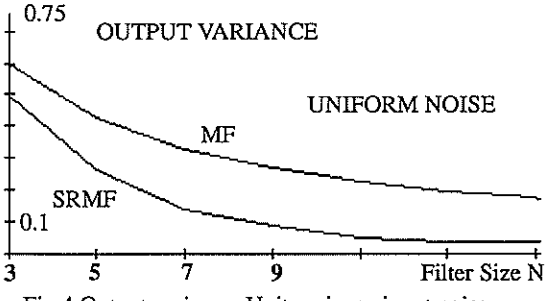


Fig.4 Output variance. Unit variance input noise

SRMF operators have the same sensitivity to abrupt changes as ordinary median filters. However, their noise reduction effect is more important. They can be useful tools in a context of industrial vision.

IV - SYMMETRICAL RECURSIVE MEDIAN GRADIENT OPERATORS

These gradient operators (so-called SRMG) use symmetrical recursive median filters for noise reduction. The derivation is obtained by linear filtering as in section II.

Deterministic properties result from the ones of SRMF operators. Like with RMG operators, structures having size lower than N/2 will have no effect on the gradient signal.

Statistical properties can be studied by determining the characteristic curves (Pd vs Pf) as in section II. We consider a step edge, with contrast h, starting at pixel no. 0. A white gaussian noise, having standard deviation σ, is added to the signal. Let Gx be the gradient component signal. If Gx(0) is greater than a given threshold t, then an edge is detected. The probability of false alarm (Pf) is the probability that an edge is detected in a stationary region. We consider signal Gx, and not its absolute value. This is the reason why, in diagrams, parameter Pf is less than 0.5 . Characteristic curves can parametrized by means of a signal-to-noise ration which is defined as : SNR=h/σ .

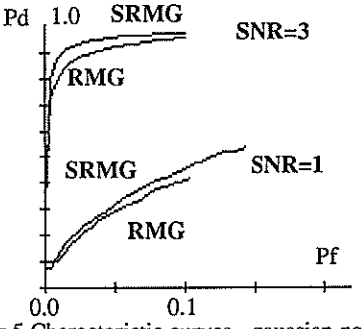


Fig.5 Characteristic curves - gaussian noise

Characteristic curves displayed on Fig.5 were obtained with a monodimensional SRMG operator having size N=5. Because of the noise reduction effect of SRMF operators, performances are improved compared with the ones of RMG operators. It should be noticed that because of streaking effects of median filters [13] and recursive median filters [12], parameter Pf cannot reach its bound 0.5 .
The localization properties can be studied by determining the spatial distribution of probability Pd, for fixed threshold t. Figure 6 represents the spatial distribution of the probability of detection, in the case of a step edge starting at pixel number 0. A white gaussian noise is added to the signal. We compare an RMG operator of size N=5, and a Canny-Deriche operator [10] having the same Pd and Pf values.

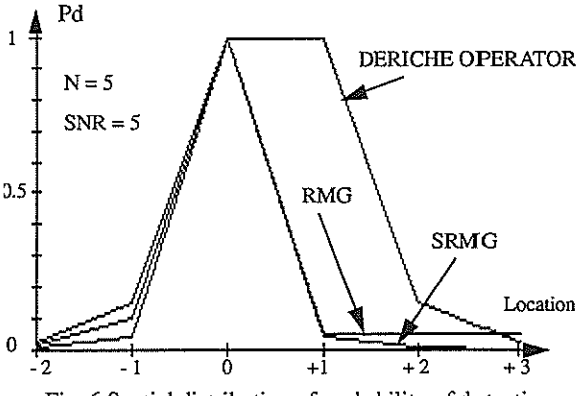


Fig. 6 Spatial distribution of probability of detection

It can be seen that the SRMG operator produces a symmetrical distribution with a better localization effect than the Canny-Deriche one. This phenomenon can be observed on photographs presented on figure 7.

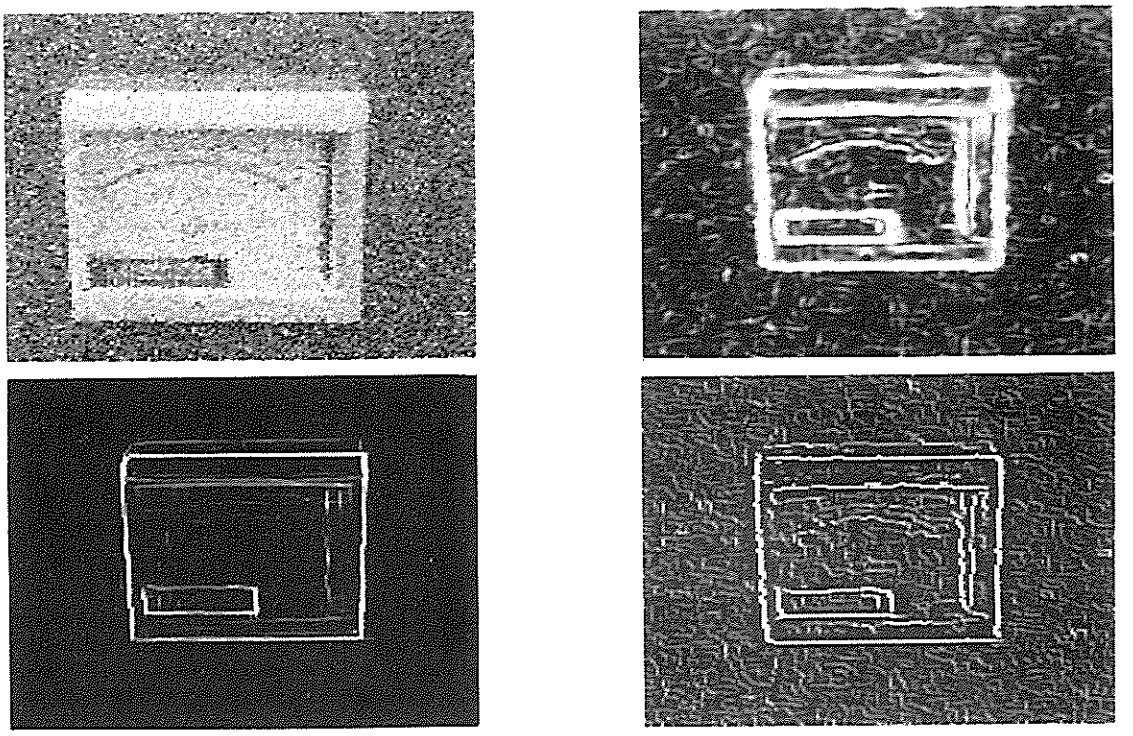


Fig.7
a/ original image, white double-exponential noise, SNR=2
b/ gradient image - SRMG operator - size N=9
c/ gradient image - Deriche operator - α=0.8
d/ edge image - local maxima of c

a	c
b	d

V - CONCLUSION

In this paper, we present symmetrical recursive median filters. Given a filter size N, these filters have a better noise reduction effect than ordinary median filters and recursive median filters. Even in the case of a uniformly distributed noise, they have better performances than the optimal linear filter, as soon as filter size N is greater than nine.

The filter size is related to the number of pixels that are processed at each period. Small filter sizes simplify filter realizations. It is also related to the maximum size of the structures, in the input signal, that are not modified.

Because of these properties, it is possible to use these filters in order to build up gradient operators. These operators have the property to be robust with respect to noise, and to detect unshifted edges.

VI - REFERENCES

[1]N.C.Gallagher, G.L.Wise, 'A theoretical analysis of the properties of median filters', IEEE Trans.ASSP, vol 29 no.6, 1981

[2] H.A.David, 'Order statistics', Wiley Interscience, 1981

[3] A.C.Bovik, T.S.Huang, D.C.Munson, 'A generalization of median filtering using linear combination of order statistics', IEEE Trans.ASSP vol.31 no.6, 1983

[4] P.Heinonen, Y.Neuvo, 'FIR-median hybrid filters with predictive FIR substructures', IEEE Trans.ASSP vol.36 no.6, 1988

[5] A.C.Bovik, D.C.Munson, 'Edge detection using median comparisons', CVGIP vol.33, 1986

[6] I.Defée, P.Heinonen, Y.Neuvo, 'Linear median hybrid edge detectors with symmetrical masks', Proc.Eusipco88, Signal Processing IV theories and applications, Lacoume et al. Ed., Elsevier Sc. Publishers, 1988

[7] M.P.Mc Loughlin, G.R.Arce :"Deterministic properties of recursive separable median filters",IEEE Tr. ASSP,vol.35 no.1, 1981

[8] G.R.Arce, R.J.Crinon, 'Median filters : analysis for 2D recursively filtered signals', Proc.ICASSP84, San Diego CA, USA, March 1984

[9] Ph.Bolon, M.Mouhoub, 'Recursive separable median filters. Application to noisy image processing', Proc. Eusipco88, Signal Processing IV theories and applications, Lacoume et al. Ed., Elsevier Sc. Publishers, 1988

[10] R.Deriche, 'Détection optimale de contours avec mise en oeuvre récursive', Proc. 11th Conf GRETSI, Nice, France, June 1987

[11] T.S.Huang, G.J.Yang, G.Y.Tang, 'A fast 2D median filtering algorithm', IEEE Trans. ASSP vol.27, no.1, 1979

[12] M.Mouhoub, 'Filtres d'ordre, filtre median recursif: analyse et évaluation des performances en traitement d'image', PhD thesis, INSA Lyon, Apr.89

[13] A.C.Bovik, 'Streaking in median filtered images', IEEE Trans. ASSP vol.35 no.4, 1987

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

ADAPTIVE ORDER FILTERS : APPLICATION TO EDGE ENHANCEMENT OF NOISY IMAGES

Ph.BOLON , J.L.FRUTTAZ

Laboratoire d'Automatique et MicroInformatique Industrielle
LAMII - Université de Savoie - BP 806 - F.74016 ANNECY CEDEX (France)
(CNRS-GDR 134 : Signal and Image Processing)

ABSTRACT

This paper presents an adaptive filter, based on ranked order statistics, that can produce 'edge enhancement' and noise reduction in an image. This filter is a modified version of CS-filtering proposed by Lee and Fam (1987). It is composed of a selecting stage and a filtering stage that are order filters. The choice of the selecting filter coefficients is discussed. Some statistical properties are presented in the case of noises having continuous distributions (uniform, gaussian, exponential). In the case of noisy images, performances of Adaptive Order Filter turn out to be better than the ones of CS-filters. Experimental results obtained on actual images from the CNRS database, and from industrial scenes are presented.

keywords : image processing, adaptive filtering, order filters, noise reduction, edge detection

I - INTRODUCTION

We are working in the context of noisy images representing objects having blurred edges. In the case of images obtained by means of a moving camera, this blurring effect can be caused by unfocusing or by a shift of the camera during the digitization period. Because of the physical process, noisy blurred edges can also be obtained in the case of angiography or ultrasonics echography images.

Considering a digitized image, blurred edges produce a transition in intensity ranging over several pixels. "Edge Enhancement" consists in reducing the width of the transition, without modifying its height (i.e. contrast between ground and object intensity). Using deconvolution techniques, such as Wiener filters, requires *a priori* knowledge about, or identification of, the blurring process. Moreover, in order to make the implementation easier, we are looking for local operators, i.e. operating with a small number of pixels.

Median filters [1] have the property to smooth noise and to preserve monotonic transitions. Adaptive methods based on order statistics were proposed by Restrepo and Bovik [2], Zamperoni [3], Lin and Wilson [4], for noise reduction, and by Zamperoni [5] for texture segmentation. Lee and Fam [6] proposed to use CS-filters in order to cope with the edge enhancement problem. However, they studied their properties in the deterministic case, or with a low level noise having bounded distribution only.

In this paper we present a generalized structure of Adaptive Order Filters (AOF), composed of a selecting stage and of a filtering stage. This structure includes CS filters. We study some statistical properties of AOF, comparing them with the ones of CS filters, in the case of noises having continuous distributions. Because of their structure, AOF turn out to be more robust with respect to noise than CS filters.

In this paper we present the structure of Adaptive Order Filters. Then we discuss the choice of the coefficient of the selecting order filter and study some of its statistical properties. We present experimental results obtained on actual images from the CNRS database, and from industrial scenes.

II - ADAPTIVE ORDER FILTERS

We consider the case of a 1D noisy blurred edge described by Eq.1. Let X_k be the intensity at pixel n^o k. We have :

$$X_k = s_k + \sigma B_k \qquad \text{eq.1}$$

with $s_k = 0 \quad k \leq k0$

$= h \quad k \geq k0 + w$

$s_{k+1} > s_k \quad k = k0 \,..\, k0 + w$ (transition area)

B_k is a zero-mean unit-variance white noise, having a continuous distribution function F, and a probability density function (pdf) f that is assumed to be even.
h denotes the edge contrast, and w the width of the transition area. In this transition area, the blurred edge is composed of a convex part, followed by a concave one.

We call edge location the pixel at which sequence $\{s_k\}$ presents an inflection point. The objective of filtering is to smooth noise and to reduce the size of the transition area and to preserve the edge location.
Adaptive Order Filters can be described by the following block diagram :

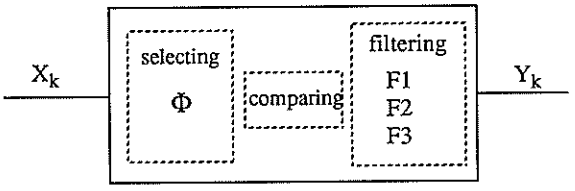


Fig.1 Block diagram of Adaptive Order Filters

Let N=2n+1 be the filter size, and $X_{(i)}$ the i^{th} order statistics in the set $\{Xk\} = \{X_{k-n},...,X_{k+n}\}$.
Φ, F1, F2, F3 are order filters. The output Yk = AOF {Xk} is obtained by means of the following algorithm :

```
if Φ{Xk} > t1 then
      Yk = F1{Xk}
else if t2 ≤ Φ{Xk} ≤ t1 then
        Yk = F2{Xk}
      else if Φ{Xk} < t2 then
        Yk = F3{Xk}
          endif
      endif
endif                    (t1 and t2 are real numbers)
```

Within the transition area, the output of selecting filter Φ should be different in the concave part and in the convex part. In the convex (resp. concave) part, the filter output could be the minimum (resp. maximun) value in the filter window. At the inflection point, the input signal should be preserved.

CS filters [6] are a special case of AOF. We have :

$$\Phi\{Xk\} = \sum f_i X_{(i)} \quad f_i = 1/N \quad i \neq n+1$$

$$f_{n+1} = 1/N - 1 \qquad \text{eq.2}$$

$$t1 = t2 = 0$$

$$F1\{Xk\} = F2\{Xk\} = X_{(1+j)}$$

$$F3\{Xk\} = X_{(N-j)} \qquad \text{eq.3}$$

From signal s_k we can define convex and concave areas. Regions that are neither convex nor concave are said to be quasi-linear. It should be noticed that, because of the discretization, convexity (concavity) is defined with respect to a filter size N.

Choosing filter coefficients:
- selecting stage . Let f_i be the filter coefficients. The output of selecting filter Φ should be zero in a constant region. Hence we have:

$$\sum f_i = 0 \qquad \text{eq.4}$$

Let $\{x_k\}$ be a convex sequence. and $\{z_k\}$ the symmetrical concave one. Because of eq.4, the selecting filter output is not modified if a constant is added to the input sequence. Therefore we can assume that $z_k=-x_k$ without loss of generality. Let u1 (resp.u2) be the output when the input is $\{x_k\}$ (resp $\{z_k\}$). We should have u1=-u2. Since $z_{(i)}=-x_{(N+1-i)}$, this implies that :

$$f_i=f_{N+1-i} \qquad \text{eq.5}$$

Since the convexity of an area is determined by the position of the secant, and since monotonic signals are roots of median filters, we propose to choose the following coefficients:

$$f_1 = f_N = 1/2 \ , \ f_{n+1} = -1 \qquad \text{eq.6}$$

It should be noticed that such a selecting filter is more sensitive to the presence of an edge than the one proposed by Lee and Fam [7].

- filtering stage. Filter F2 generally operates on stationary areas of the input signal. We then can choose it as the optimal order filter [8][9]. Filter F1 (resp.F3) is obtained by emphasizing lower (resp higher) statistics. If CS filters are used with noisy images, the only way to avoid noise amplification is to increase the value of parameter j of eq.3. This decreases the edge enhancement effect. With AOF, the noise reduction effect is less dependent on the edge enhancement effect, since thresholds t1 and t2 can be tuned. In the case of a noise having a symmetrical distribution, one can choose t2=-t1.

III - STATISTICAL PROPERTIES

We consider the case of symmetrical noise distribution. Since edges can be described by increasing or decreasing sequences, we can choose t1 = -t2 = t >0. Statistical properties of AOF strongly depend on the selecting stage.
Let U be the ouput of filter Φ. Assume that we have |U|>t in a stationary region (false alarm). Since filters F1 and F3 tend to emphasize extreme values in the filter window, a noise amplification can occur. This situation can be controlled by means of the choice of the value of threshold t. Let P_f be the probaility of false alarm, and F_U be the cdf of random variable U. We have ,

$$P_f = F_U(-t) + 1-F_U(t) = 2.F_U(-t) \qquad \text{eq.6}$$

Probability of false alarm : Let g be the joint probability density function of $X_{(1)}$, $X_{(n+1)}$, $X_{(N)}$. We have [9]

$$g(x,y,z) = C \, . \, f(x) \, f(y) \, f(z) \, (F(y)-F(x))^{n-1} (F(z)-F(y))^{n-1}$$

$$\text{if } x \leq y \leq z$$

$$= 0 \text{ else} \qquad \text{eq.7}$$

with $C = N! \, / \, \{(n-1)! \, (n-1)!\}$

If we change to the co-ordinate system (u,v,w) with
u = x/2 -y +z/2, v=x, w=y,and if the new joint pdf is q, we have:

$$q(u,v,w)=2 \, C. \, f(v) \, f(w) \, f(2u-v+2w) \, (F(w)-F(v))^{n-1}$$

$$(F(2u-v+2w)-F(w))^{n-1} \qquad \text{eq.8}$$

$$F_u(x) = \int_{-\infty}^{x} \int_{Dv} \int_{Dw} q(u, v, w) \, du dv dw \qquad \text{eq.9}$$

These integrals can be numerically computed in order to analyse the effect of the noise distribution, and of the filter size.

Influence of the noise distribution : By comparing the computed cdf of random variable U, we noticed that with gaussian or uniformly distributed input noises, results are similar (Fig.2-3). If the kurtosis of the input distribution is increased, as in the case of double-exponential pdf (so-called exponential noise), the probability of false alarm is increased, for fixed t (Fig.4).

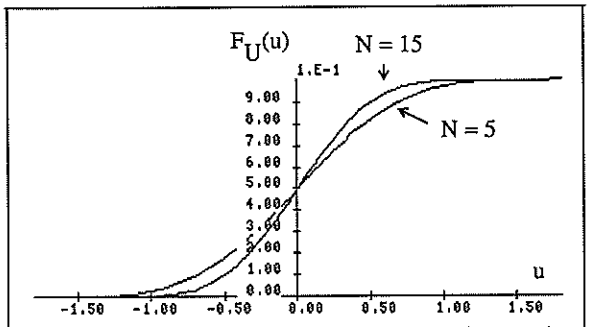


Fig.2 cdf of variable U. unit variance uniform input noise

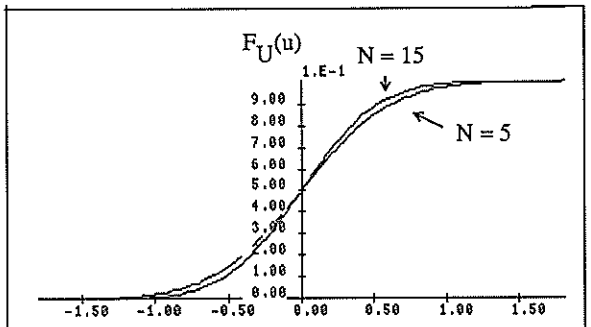


Fig.3 cdf of variable U. unit variance gaussian input noise

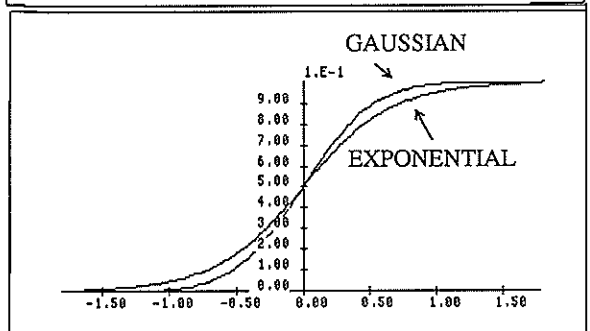


Fig.4 cdf of variable U. Filter size N=15

Influence of the filter size : Considering the curves presented on Fig.2-3, we remark that probability Pf generally decreases as the filter size increases. It should be noticed that this is not true for noises having heavy-tailed distributiuons. The probability that 'outliers' occur inside the filter window increases with the filter size. This phenomenon tends to disturb the selecting process.

Detection of a transition region : In this section we are comparing the performances of the selected filter Φ to the one proposed by Lee and Fam. We consider a noisy ramp edge, starting at pixel no. k0 (contrast:h, width:w). σ denotes the standard deviation of the noise, and RSB=h/σ a signal-to-noise ratio. We want to determine the probability of detecting a transition region at the starting pixel (Pd:probability of detection) and in a stationary area (Pf:probability of false alarm). A sequence of noisy edges (width w=15) is simulated. The sequence length is Ns=30 000. Hence,the experimental probabilities can be regarded as unbiased estimates of the real ones, with a standard deviation σ_p that is less than 0.003.

As showed by the characteristic curves on Fig.5, the performances of AOF are generally better than the ones of CS filter, except with an exponential noise (Fig.6)

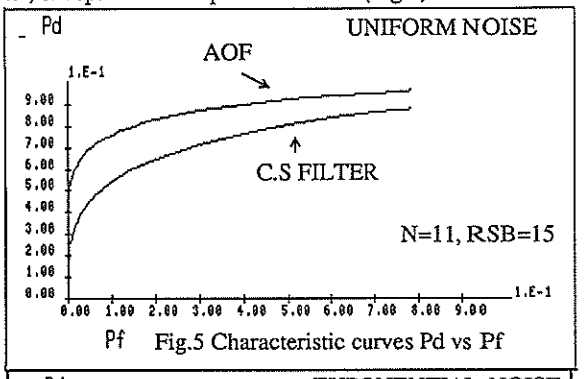


Fig.5 Characteristic curves Pd vs Pf

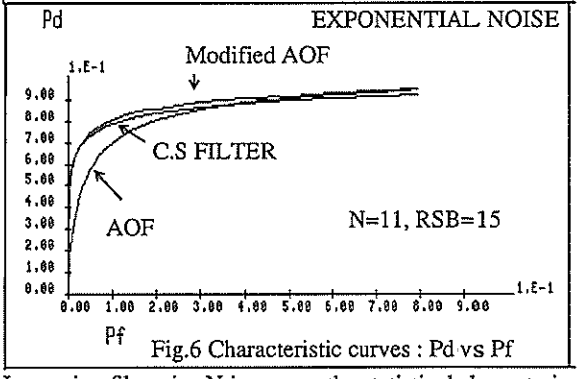
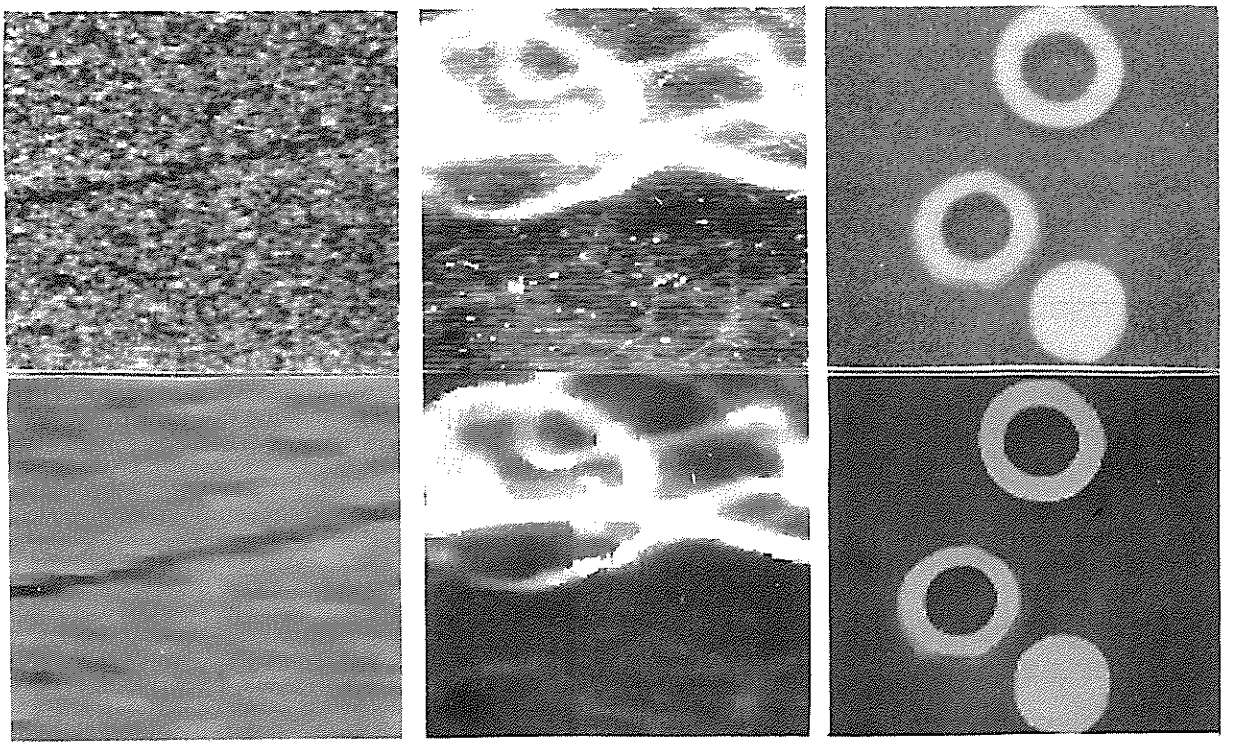


Fig.6 Characteristic curves : Pd vs Pf

Increasing filter size N improves the statistical characteristics of the selecting stage. In the case of heavy-tailed distributions, the coefficients of filter Φ should be modified by choosing : $f_2 = f_{N-1} = 1/2$, f_{n+1}=-1 (Modified AOF).

IV - EXPERIMENTAL RESULTS

Some experimental results are presented on figure 7. Images presented Fig.7a and 7c come from the CNRS database. The first one is a X-radiography of a crack in a metallic workpiece. The second one is a numerical angiography. Fig. 7e represents an industrial scene viewed by means of an unfocused camera. A white gaussian noise has been added. Signal-to-noise parameter RSB is low (7a,7c) or moderate (7e). These images are filtered by means of separable AOF.



a	c	e
b	d	f

Fig.7 Experimental Results

a/ CNRS database/fissures.ima (crack) RSB = 2
b/ image filtered by separable AOF 15x9
c/ CNRS database/angiogra.ima (angiography) RSB = 2 to 10
d/ image b filtered by separable AOF 9x7
e/ industrial scene (gaussian noise) RSB = 10
f/ image e filtered by separable AOF 7x9

V - CONCLUSION

Adaptive Order Filters (AOF) are local operators that are useful for image enhancement, in the case of noisy images having blurred edges. As a first processing step, before gradient estimation, they produce sharp edges which are not moved with respect to their original location.
Because of their symmetrical structure, AOF presented in this paper have better performances than CS-filters proposed by Lee and Fam, in the case of noisy images. Moreover, good results can be obtained by means of separable filters, that have a lower complexity than 2D ones.
However, further work aiming at designing optimal filters F1 and F3, and at describing the statistical characteristics of the filter output, has to be completed.

VI - REFERENCES

[1] N.C.Gallagher and G.L.Wise, 'A Theoretical Analysis of the Properties of Median Filters' IEEE Trans.ASSP vol.29, pp.1136-1141, 1981
[2] A.Restrepo and A.C.Bovik, 'Adaptive Trimmed Mean for Image Restoration', IEEE Trans. ASSP vol.36 no.8, pp.1326-1337, Aug. 1988.
[3] P.Zamperoni, 'Some Adaptive Rank Order Filters for Image Enhancement', Pattern Recognition Letters vol.11, pp.81-86, 1990.
[4] H-M Lin and A.N.Wilson, 'Median Filters with Adaptive Length', IEEE Trans. CAS, vol.35, no.6, pp.675-690, 1988.
[5] P.Zamperoni, 'Feature Extraction by Rank Order Filtering for Image Segmentation', IJPRAI vol.2 no.2 pp. 301-319, 1988.
[6] Y.H.Lee and A.T.Fam, 'An Edge Gradient Enhancing Adaptive Order Statistic Filter', IEEE Trans. ASSP, vol.35, no.5, pp.680-695, may 1987.
[7] J.L.Fruttaz, 'Etude des propriétés statistiques d'un filtre rehausseur de contraste', Internal LAMII Report, sept.1989
[8] A.C.Bovik, T.S.Huang, D.C.Munson : " A generalization of median filtering using linear combination of order statistics", IEEE Trans. ASSP, vol. 31 no12, pp.1342-1350, dec. 1983
[9] H.A.David, 'Order Statistics', Wiley Interscience, 1981

VII - ACKNOWLEDGEMENTS

Numerical integrals were computed by means of routines from Harwell library.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

CONSIDERATIONS IN THE IDENTIFICATION AND RESTORATION OF BLURRED PHOTOGRAPHIC IMAGES

A. Murat Tekalp, Schlomo Koch†, Reginald Lagendijk#, Gordana Pavlović and Howard Kaufman†

Electrical Engineering Department, University of Rochester, Rochester, New York 14627, USA
†ECSE Department, Rensselaer Polytechnic Institute, Troy, New York, 12180-3590, USA
#Electrical Engineering Department, Delft University of Technology, 2600 GA Delft, The Netherlands

In this paper, we discuss some issues involved in the practical restoration of blurred images, and in particular blurred photographic images. We review the assumptions under which most blur identification and image restoration algorithms were developed, and elaborate on the validity of these assumptions in general, and in the case of photographic images. In particular, we discuss incorporating the nonlinearities introduced by the photographic film and the scanner in the case of photographic images. We also discuss some general issues such as the problems encountered in blur identification, how to account for the spatial variations in the image and blur models, and the relation between signal-to-noise ratio and resolution. Some experimental results are presented.

1. INTRODUCTION

Image restoration may be an important tool in recovering important information from one of a kind blurred photographic images and documents. Image restoration has applications in law enforcement and forensic science, medical imaging, space explorations, image communications, and consumer photography. This paper aims to discuss the issues involved in the practical use of blur identification and image restoration algorithms.

The basic premises in the image restoration literature are:

(i) The blur formation is considered as linear and space-invariant. Then, the resulting image plane intensity distribution can be represented by the 2-D convolution of the ideal image with the point spread function (PSF) of the image formation system, as follows:

$$b(m,n) = \sum_{(o,p)\in R} h(m-o, n-p)s(o,p), \qquad (1)$$

where R denotes the blur model support and $\{h_{i,j}\}$ are the PSF coefficients.

(ii) Image recording is modeled as a linear process, and the effect of all noise processes appearing during image formation and recording can be modeled as an additive process that is uncorrelated with the image signal. As a result, the recorded image is given by

$$r(m,n) = b(m,n) + v(m,n), \qquad (2)$$

where $v(m,n)$ is a white, Gaussian process, with zero-mean and variance σ_v^2. We assume high signal-to-noise ratios (SNR) are available for the image restoration problem.

(iii) The ideal image can be more-or-less represented by a homogeneous random field, and its power spectrum can be estimated using classical nonparametric methods. Alternatively, under the Gauss-Markov assumption, the ideal image can be represented by a non-symmetric half plane (NSHP) causal autoregressive (AR) difference equation model [1]:

$$s(m,n) = \sum_{(k,\ell)\in R_{\oplus+}} c_{k,\ell}\, s(m-k, n-\ell) + w(m,n) \qquad (3)$$

where $\{c_{k,\ell}\}$ are the NSHP model coefficients, and $w(m,n)$ is a zero-mean, white Gaussian field with variance σ_w^2, to account for the uncertainty in the model. The region $R_{\oplus+}$ denotes the NSHP model support.

Based on these assumptions, some cepstral domain [2,3], and maximum likelihood (ML) blur identification algorithms [4] were developed, as well as a wide range of frequency domain, recursive and iterative restoration methods [5]. However, these algorithms were mostly tested under fairly restrictive and controlled conditions, namely where the blur and noise were synthetically introduced so as to conform with the assumptions. Recent experiments with the restoration of blurred photographic images indicate that in many cases these algorithms do not work with the real data. Some of the reasons for this are: (i) image sensors and scanners have nonlinear characteristics, (ii) image and blur model parameters usually vary as a function of spatial coordinates, (iii) there are numerical difficulties associated with the ML identification of blurs with large supports, (iv) available SNR is usually not high enough to restore small details in the images. This paper addresses the above problems in detail, and discusses recent solution strategies.

2. INCORPORATING SENSOR NONLINEARITY

In scanning images recorded on photographic film, the resulting image signal is proportional to the optical density where there exists a logarithmic relation between the optical density and the exposure (integral of the light

intensity during the time of exposure). This relation is usually provided by the manufacturer as a d—log e curve. Assuming that all density values of the recorded image fall into the linear region of a typical $d - \log e$ curve we can write a parametric model for the $d - \log e$ curve as

$$D(E) = \alpha \log(E) + \beta, \tag{4}$$

where α stands for the slope of the linear region, β is the density-offset of the extrapolated linear region with respect to the origin

Combining (2) and (4), the observed image can be expressed as

$$r_d(m,n) = \alpha \log\{h(m,n) * * s_e(m,n)\} + \beta + v_d(m,n), \tag{5}$$

where the subscripts d and e are used to explicitly indicate the density and the exposure domains, respectively.

As a first step in processing, we transform the observed blurred and noisy image to a so-called "exposure domain" using the inverse of the transformation (4), in order to establish a linear convolutional relationship between the observed and original images, to obtain

$$r_e(m,n) = v_e(m,n)[h(m,n) * * s_e(m,n)] \tag{6}$$

where $v_e(m,n) = 10^{\frac{v_d(m,n)}{\alpha}}$, and $r_e(m,n)$ denotes the transformed observations. We observe that the additive observation noise in the density domain manifests itself as multiplicative noise in the exposure domain.

2.1 Wiener Filter with Multiplicative Noise

We derive the LMMSE filter (the Wiener filter) for deconvolution, in the exposure domain, in the presence of multiplicative observation noise [6]. It is well-known that the 2-D Wiener filter has the frequency domain transfer function

$$\mathcal{W}(\omega_1,\omega_2) = \frac{S_{s_e r_e}(\omega_1,\omega_2)}{S_{r_e r_e}(\omega_1,\omega_2)}, \tag{7}$$

where $S_{s_e r_e}(\omega_1,\omega_2)$ is the cross-spectral density of the observed image $r_e(m,n)$ and the original image $s_e(m,n)$, $S_{r_e r_e}(\omega_1,\omega_2)$ is the spectral density of the observed image $r_e(m,n)$.

We can express the frequency response of the Wiener filter in the presence of multiplicative noise, independent of the image signal $s_e(m,n)$, with arbitrary distribution as

$$\mathcal{W}_M(\omega_1,\omega_2) = \frac{\mathcal{E}\{v_e(i,j)\}H^*(\omega_1,\omega_2)S_{ss}(\omega_1,\omega_2)}{\{|H(\omega_1,\omega_2)|^2 S_{ss}(\omega_1,\omega_2)\} * * S_{v_e v_e}(\omega_1,\omega_2)}. \tag{8}$$

Given that observation noise in the density domain, $v_d(m,n)$, is white, Gaussian with zero mean with variance $\sigma^2_{v_d}$, the frequency response of the Wiener filter can be expressed as

$$\mathcal{W}_M(\omega_1,\omega_2) = \tag{9}$$
$$(\frac{1}{\sqrt{\gamma}})\frac{H^*(\omega_1,\omega_2)S_{ss}(\omega_1,\omega_2)}{|H(\omega_1,\omega_2)|^2 S_{ss}(\omega_1,\omega_2) + (\gamma-1)\{\overline{|H(\omega_1,\omega_2)|^2 S_{ss}(\omega_1,\omega_2)}\}}$$

where $\gamma = e^{\frac{\sigma^2_{v_d}\ln^2 10}{\alpha^2}}$ and $\overline{X(.,.)}$ denotes averaging over all values of the independent variables.

We observe that the optimal linear filter $\mathcal{W}_M$, in the case of the particular type of multiplicative noise $v_e(m,n) = 10^{\frac{v_d(m,n)}{\alpha}}$ where $v_d(m,n)$ is zero-mean Gaussian and uncorrelated with the original image, differs from the classical Wiener filter in the case of additive white Gaussian noise, by a DC gain $1/\gamma$ and in the noise power factor. However, in general, if $v_d(m,n)$ is not Gaussian, it is possible to obtain from (8) filter structures that are fundamentally different from the classical Wiener filter.

2.2 Results

We processed scanned photographic "License plate" image, shown in Fig. 1 (a), which is blurred due to an out-of-focus camera. This image was taken using a 35 mm camera and Kodak VR-G 200 film. The image was scanned by using a microdensitometer (courtesy of Eastman Kodak Company). The scanner characteristics and the $d - \log e$ curve of the film are provided by Eastman Kodak Company.

The PSF of the out-of-focus blur is modeled by a uniform disk and identified in the exposure domain, using the method proposed by Gennery [2].

We estimated the diameter of the PSF to be 14 pixels and the signal-to-noise ratio (SNR) as 17 dB. The image that is restored by our new filter is shown in the exposure domain in Fig. 1 (c). The restoration obtained by using the classical Wiener filter in the density domain is shown in Fig. 1 (b). There is no visible improvement in the case of the conventional Wiener filter

3. HIERARCHICAL BLUR IDENTIFICATION

Various implementations of the ML estimator have been discussed in [4]. Each of these implementations suffer from the following drawbacks: (i) A reasonably good *a priori* estimate of the support of PSF is required. (ii) The likelihood function contains several local maxima, and the number of local maxima increases with the number of free parameters to be estimated. Hence, it is desirable to keep the number of free parameters low, and to have reasonably good initial estimates to prevent convergence to one of the local maxima. A hierarchical blur identification and image restoration strategy was proposed recently [7] which greatly circumvents these drawbacks. Although the proposed hierarchical strategy is essentially independent of the chosen implementation of the ML estimator, we focus on the EM implementation in this paper.

Blur is a resolution dependent phenomenon, because the severity of the blur in a discrete image depends on how the continuous blurred image is sampled. Blurred images sampled at a coarser resolution look (from the same distance) less blurred than images sampled at a

higher resolution at the expense of containing less information. Hence, if we need to identify a discrete image blurred by a PSF with a large support, the dimensionality of the problem can (initially) be reduced by considering the identification of a subsampled version of the image. In this way, the effective PSF support is reduced, and fewer PSF coefficients have to be identified. In addition to this, the likelihood function contains fewer local optima, and the identification process is computationally less expensive because it operates on less data. Once the PSF and the image model of the subsampled image have been identified, an interpolated version of these parameters and/or the restoration result can be used to initialize the identification/restoration algorithm on the full resolution blurred image. Repeating this procedure p times using a subsampling factor of 2^p for the p-th subsampled image, we obtain a hierarchical identification/restoration method that "builds up" a possibly very large PSF by gradually increasing its resolution.

3.1 Downsamling the Blurred Image

The filtered and downsampled blurred images $r_p(m,n)$ are related to the full resolution image $r(m,n)$ as

$$r_p(m,n) = \{\ell_p(i,j) * *r(i,j)\}_{(i,j)=(2^p m, 2^p n)}, \qquad (10)$$

where $\ell_p(i,j)$ is the low pass filter which restricts the bandwidth of $r(m,n)$ before downsampling in order to prevent aliasing effects. If we neglect the noise contribution, then

$$r_p(m,n) = \{\ell_p(i,j) * *h(i,j) * *s(i,j)\}_{(i,j)=(2^p m, 2^p n)}. \quad (11)$$

Associated with $\{r, r_1, \ldots, r_p\}$ are the downsampled ideal images $\{s, s_1, \ldots, s_p\}$ and the downsampled PSF's $\{h, h_1, \ldots, h_p\}$. We use a lowpass filter which satisfy the following conditions: (i) if the blurred image is downsampled by a factor of 2^p then the support of the PSF has to be reduced by the same factor, (ii) the filter impulse responses $\ell_p(i,j) > 0$ to ensure that $h_p(i,j) = \ell_p(i,j) * *h(i,j) > 0$, (iii) the filter must have a cut-off frequency approximately at $w_c = \pi 2^{-p}$. A separable lowpass filter impulse response which satisfies these conditions can be obtained from a truncated sinc function windowed by a Hamming function

$$\ell_p(k) = \frac{sin(\pi k 2^p)}{\pi k 2^p}(0.54 + 0.46 cos(\pi k 2^p)), \quad -2^p+1 \le k \le 2^p-1. \qquad (12)$$

The parameters identified at the resolution level $(p+1)$ can be used to initialize either the E-step or the M-step of the EM-algorithm at the resolution level p. To this end, either the identified PSF or the restored image at level $(p+1)$ needs to be interpolated. We initialize the EM algorithm at the next resolution level by the interpolated restoration result from the previous level obtained by pixel replication.

3.2 Results

We consider the identification of the "Cameraman image", blurred by the symmetric horizontal motion over 9 pixels and with SNR of 40 dB. Table 1 shows the identified PSF coefficients at each level (since PSF is symmetric only half coefficients are listed). The (initial) parameters computed by the M-step in the first iteration from the interpolated restoration result are shown in this table. At the resolution level 2, the value of the likelihood function indicates that the 1x3 PSF is to be preferred to the 1x5 PSF. At the resolution level 1, the likelihood function for the 1x5 PSF support is significantly smaller than for the 1x3 PSF support size. Hence at the next level we need to consider a 1x7 and 1x9 PSF support size. Finaly for the full resolution image, preference has to be given to the support of the size 1x9.

4. SPATIAL VARIATIONS IN PARAMETERS

Because most real images are by nature nonhomogeneous, it is necessary that restoration procedures account for image parameter variations. Furthermore, because motion and out of focus blurs may be spatially varying, the observation equation will also contain parameter variations.

In the extreme, a different model for each pixel would be required to represent an inhomogeneous image. The image model would then be given by:

$$s(m,n) = \sum_{(k,\ell)\epsilon R_{\Theta+}} c_{k,l}(m,n) s(m-k, n-\ell) + w(m,n) \quad (13)$$

where $\{c_{k,\ell}(m,n)\}$ are the NSHP model coefficients at pixel (m,n).

The blurred image would then be described by:

$$r(m,n) = \sum_{o=0}^{M} \sum_{p=0}^{N} d(m,n;o,p) s(o,p) + v(m,n) \qquad (14)$$

where $r(m,n)$ is the degraded image, $s(o,p)$ is the ideal image to be determined, $d(m,n;o,p)$ is the point spread function (PSF) of the degradation, and $v(m,n)$ is the observation noise. In (4.2), the PSF is a function of input image coordinates (o,p) and output image coordinates (m,n), and hence it is space-variant.

4.1 Restoration procedures

Because the image blur parameters are continuous in the spatial domain, they can be modeled as additional state variables that will be estimated simultaneously with the image intensity $s(m,n)$. For example we could model an image parameter by the equation:

$$c_{k,\ell}(m,n) = c_{k,\ell}(m-1,n) + w_c \qquad (15)$$

where w_c is a noise term that represents the anticipated variation or uncertainty in the parameter $c_{k,\ell}$. Since (4.4) represents only a one dimensional variation in $c_{k,\ell}$, it may be more desirable to impose a NSHP 2-D support region for $c_{k,\ell}$ giving a more complex equation of the form:

$$c_{k,\ell}(m,n) = \sum \alpha_{k,\ell}\, c_{k,\ell}(m-k, n-\ell) + w_c(m,n) \quad (16)$$

where the coefficients $\alpha_{k,\ell}$ would be apriori specified.

In either case, the resulting estimation problem is now nonlinear because it involves products of unknowns (i.e., coefficients and image intensities).

Thus whereas in the presence of known coefficients, restoration was possible with a linear Kalman filtering procedure [8], it is now necessary to use a linearized or extended Kalman based restoration procedure [9].

4.2 Results

To date a series of synthetic images has been generated and extended Kalman filtering has been used only for single blur parameter identification. Results indicated that extensive tuning procedures are required because of the effects of linearization.

The extended Kalman filter is sensitive to the value of the boundary conditions. Knowing the correct values results in a fast convergence and accurate estimate of the blur parameter. If boundary conditions are far from their true values, because of the linearization the convergence is slow, and not always to the correct values.

Another way to improve parameter identification is to increase the value of the plant noise covariance matrix. Tuning this improves the results of parameter identification matrix.

	s_d	Point-spread function $d(0,0)$	$d(0,1)$	$d(0,2)$	$d(0,3)$	$d(0,4)$	σ_w^2
Resolution level 4							
initial	1x3	1.000	0.000				10.0
final		0.575	0.213				1.10
Resolution level 3							
initial	1x3	0.546	0.272				282.6
final		0.582	0.210				1.53
Resolution level 2							
initial	1x3	0.307	0.301				180.2
final		0.506	0.247				0.10
initial	1x5	0.400	0.288	0.012			189.2
final		0.483	0.247	0.011			1.10
Resolution level 1							
initial	1x3	0.282	0.359				108.3
final		0.554	0.223				0.36
initial	1x5	0.302	0.232	0.117			96.7
final		0.242	0.257	0.122			0.11
Resolution level 0							
initial	1x7	0.098	0.143	0.129	0.179		62.6
final		0.140	0.101	0.141	0.188		0.37
initial	1x9	0.120	0.124	0.138	0.108	0.069	59.2
final		0.123	0.124	0.124	0.129	0.061	0.29
ideal		0.125	0.125	0.125	0.125	0.063	0.35

Table 1 Initial and identified parameters at 5 resolution levels for horizontal motion blur an SNR=40 dB.

REFERENCES

1. J.W. Woods, 'Markov Image Modeling,' *IEEE Trans. Auto. Control*, vol. AC-23, pp. 846-850, Oct. 1978.
2. D.B. Gennery, "Determination of Optical Transfer Function by Inspection of Frequency-Domain Plot," *J. Opt. Soc. Amer.*, vol. 63, pp. 1571-1577, Dec. 1973.
3. M. Cannon, 'Blind Deconvolution of Spatially Invariant Image Blurs with Phase,' *IEEE Trans. Acoust., Speech, Sign. Proc.*, vol. ASSP-24, pp. 58-63, Feb. 1976.
4. R. L. Lagendijk, A. M. Tekalp and J. Biemond, 'Maximum Likelihood Image and Blur Identification, A Unifying Approach', *J. Opt. Eng.* May 1990.
5. M. I. Sezan and A. M. Tekalp, 'A Survey of Recent Developments in Digital Image Restoration', *J. Opt. Eng.* May 1990.
6. G. Pavlović, A. M. Tekalp , 'Restoration in the Presence of Multiplicative Noise with Application to Scanned Photographic Images', *Proc. IEEE Int. Conf. ASSP*, 1990.
7. R. L. Lagendijk, J. Biemond, D. E. Boekee, 'Hierarchical Blur Identification', *Proc. IEEE Int. Conf. ASSP*, 1990.
8. D. Angwin and H. Kaufman, "Image Restoration Using Reduced Order Models," *Signal Processing*, vol. 16, pp. 21-28, Dec. 1988.
9. L. Ljung, "Asymptotic Behavior of the Extended Kalman Filter as a Parameter Estimator for Linear Systems", *IEEE Trans. Auto. Control*, Feb. 1979.

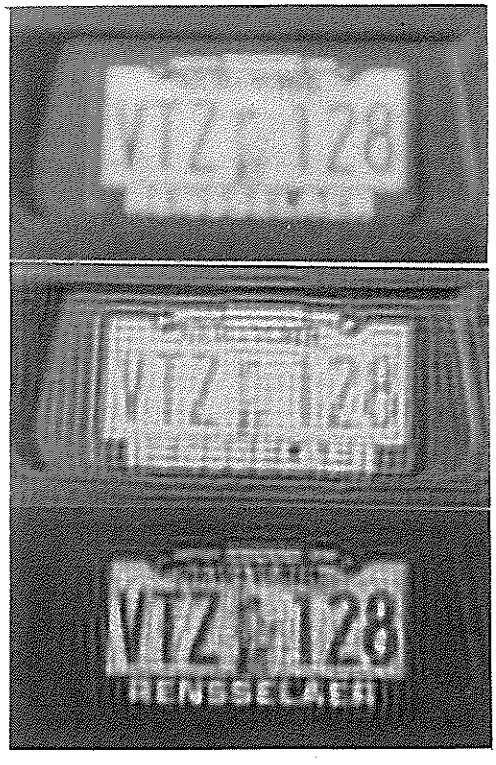


Figure 1 (a) Blurred photographic image displayed in density domain, (b) Restored image using the conventional Wiener filter displayed in density domain, (c) Restored image usig the proposed filter displayed in exposure domain.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

MULTI-SCALE IMAGE RESTORATION

J.M. Bruneau, M. Barlaud, P. Mathieu

LASSY Equipe I3S CNRS Université de NICE-SOPHIA ANTIPOLIS
Bat 4 S.P.I. - Rue Albert Einstein 06560 Valbonne (FRANCE)

ABSTRACT **: In this paper we propose a new method for image restoration using a Biorthogonal Wavelet Transform. First we use a pyramidal algorithm architecture based on Biorthogonal Wavelet Transform in order to obtain a set of images at different scales. The blurred image is restored at a coarse resolution and then this process is iterated, gradually increasing the resolution.This coarse to fine restoration process results in a progressive restoration of the blurred image.**

I. INTRODUCTION TO THE WAVELET TRANSFORM

Wavelets are functions generated from one single function ψ by dilations and translations :

(1) $\psi_n^j(x) = 2^{-j/2}\,\psi(2^{-j}x - n).$

The basic idea of the Wavelet Transform is to represent any arbitrary function f as a superposition of wavelets.

(2) $f = \sum c_n^j(f)\,\psi_n^j$

The ψ_n^j constitute an orthonormal basis so that :

(3) $c_n^j(f) = \langle \psi_n^j, f \rangle = \int dx\, \psi_n^j(x)\, f(x)$

In a multiresolution analysis, one really has two functions : the mother wavelet ψ and a scaling function ϕ.

One also introduces dilated and translated versions of the scaling function :

(4) $\phi_n^j(x) = 2^{-m/2}\,\phi(2^{-m}x - n).$

If we want :

i) to use a pair of exact reconstruction filters corresponding to an orthonormal basis ;

ii) to have fast computation (the filters should be short) ;

iii) to use symmetric filters as well ;

the only symmetric exact reconstruction filters are those corresponding to the Haar basis. Thus we propose to use the Biorthogonal Wavelet Transform.

II. THE BIORTHOGONAL WAVELET TRANSFORM

II.1. Definition

The goal is to preserve symmetry by relaxing the orthonormality requirement.

The ψ_n^j no longer constitute an orthonormal basis so the computation of the coefficients $c_n^j(f)$ is carried out via the dual basis :

(5) $c_n^j(f) = \langle \tilde{\psi}_n^j, f \rangle$

where $\tilde{\psi}$ is a different function.

When f is given in sampled form, one can take these samples for the coefficients a_n^j and :

(6) $a_n^{j+1}(f) = \sum_k h_{2n-k}\, a_n^j\,(f)$

(7) $c_n^{j+1}(f) = \sum_k g_{2n-k}\, a_k^j\,(f)$

Which describes a subband algorithm with :

(8) $h_n = 2^{1/2} \int dx\, \phi(x-n)\, \phi(2x)$ ==> Low pass filter

(9) $g_l = (-1)^l\, h_{-l+1}$ ==> High pass filter

Consequently, the exact reconstruction is given by :

(10) $a_l^j(f) = \sum_n [\, \tilde{h}_{2n-l}\, a_n^{j+1}(f) + \tilde{h}_{2n-l}\, c_n^{j+1}(f)\,]$

The relation between the different filters is given by :

(11) $\tilde{g}_n = (-1)^n h_{-n+1}$ and $g_n = (-1)^n \tilde{h}_{-n+1}$

(12) $\sum_n h_n\, \tilde{h}_{n+2k} = \delta_{k,0}$

II.2. Computation of Filter coefficients

Let $H(\xi)$ and $\tilde{H}(\xi)$ be respectively the Fourier Transform of h and $\tilde{h}$ then :

i) regularity for ψ and $\tilde{\psi}$ implies that $H(\xi)$ and $\tilde{H}(\xi)$ are divisible by $(1 + e^{j\xi})$ to some power ;

ii) we want the filter lengths of h and $\tilde{h}$ to be approximately equal.

The problem statement is, for symmetric filters, to find two trigonometric polynomials $H(\xi)$ and $\tilde{H}(\xi)$ **[CDF]**, **[Baub1,2]**, **[ABMD1,2]** such that :

(13) $H(\xi)\,\tilde{H}(\xi) + H(\xi+\pi)\,\tilde{H}(\xi+\pi) = 1$

(14) $H(\xi)\tilde{H}(\xi) = \cos(\xi/2)^{2l}\,.$

$$[\sum_{j=0}^{l-1} \binom{l-1+j}{j} \sin(\xi/2)^{2j} + \sin(\xi/2)^{2l}\, R(\xi)]$$

where $R(\xi)$ is an odd polynomial in $\cos(\xi)$.

II.3. Example

The impulse responses of H and $\tilde{H}$ are :

$h_0 = .520897409718$ $\qquad$ $\tilde{h}_0 = .636046869922$

$h_1 = h_{-1} = .244379838485$ $\qquad$ $\tilde{h}_1 = \tilde{h}_{-1} = .337150822538$

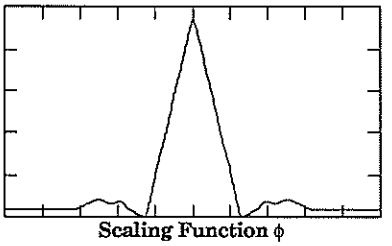
$h_2 = h_{-2} = -.0385117141551$ $\quad \tilde{h}_2 = \tilde{h}_{-2} = -.0661178056048$

$h_3 = h_{-3} = .0056201615151$ $\quad \tilde{h}_3 = \tilde{h}_{-3} = -.0966661530487$

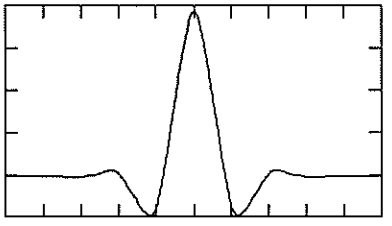
$h_4 = h_{-4} = .0280630092963$ $\quad \tilde{h}_4 = \tilde{h}_{-4} = -.00190562935635$

$\tilde{h}_5 = \tilde{h}_{-5} = .00951533051121$

Because of the length of h and $\tilde{h}$ we call these two filters "9-11". The scaling functions ϕ and $\tilde{\phi}$ corresponding respectively to the analysis filter H and synthesis filter $\tilde{H}$ are :



Scaling Function ϕ



Scaling Function $\tilde{\phi}$

II.4. Matrix Formulation of the Wavelet Transform

We now turn to a matrix formulation of the Biorthogonal Wavelet Transform, as defined hereafter.

Let $\underline{a}^T = [a_1, a_2, .., a_p]$ be a vector of dimension $p=2^k$. We define a vector $\underline{a}^{jT}$ of dimension 2p at resolution j as follows:

(15) $\underline{a}^{jT} = [a_1, a_2, .., a_p, x, a_p, .., a_1]$

where x is a point which has been interpolated using a_p, a_{p-1},... We then set $\frac{N}{2^j} = 2p$.

Using the notation in section **II.1.**, we define the filtering operators at resolution j - $H^j, \tilde{G}^j$, $\tilde{H}^j$ and G^j - as follows :

Let $(m,n) \in [1, \frac{N}{2^j}] \times [1, \frac{N}{2^j}]$. We set :

(16) $H^j(m,n) = h_{|m-n|}$ if $|m-n| \in [1, \frac{N}{2^{j+1}}]$

$= h_{|\,(|m-n|-\frac{N}{2^j})|}$ if $|m-n| \in [\frac{N}{2^{j+1}}, \frac{N}{2^j}]$

(17) $\tilde{G}^j(m,n) = 2\,(-1)^{|m-n|} H^j(m,n)$

(18) $\tilde{H}^j(m,n) = 2\tilde{h}_{|m-n|}$ if $|m-n| \in [1, \frac{N}{2^{j+1}}]$

$= 2\tilde{h}_{|\,(|m-n|-\frac{N}{2^j})|}$ if $|m-n| \in [\frac{N}{2^{j+1}}, \frac{N}{2^j}]$

(19) $G^j(m,n) = \frac{1}{2}(-1)^{|m-n|} \tilde{H}^j(m,n)$

These four matrices are circulant and symmetric.

We also define the decimation operators on the odd and even rows at resolution j : D_o^j et D_e^j.

Let $(m,n) \in [1, \frac{N}{2^{j+1}}] \times [1, \frac{N}{2^j}]$. We have:

(20) $D_o^j(m,n) = 1$ if n=2m-1

$= 0$ elsewhere

(21) $D_e^j(m,n) = 1$ if n=2m

$= 0$ elsewhere

II.4. Analysis and Synthesis of a vector

Let $\underline{a}^{j+1}$ be vector $\underline{a}$ at resolution j+1 and $\underline{c}_a^{j+1}$ the associated wavelet coefficients; we have, in augmented vector form :

(22) $$\begin{pmatrix} \underline{a}^{j+1} \\ \underline{c}_a^{j+1} \end{pmatrix} = \begin{pmatrix} (D_o^j H^j) & 0 \\ 0 & (D_e^j G^j) \end{pmatrix} \begin{pmatrix} \underline{a}^j \\ \underline{a}^j \end{pmatrix}$$

with Dim $\underline{a}^{j+1}$ = Dim $\underline{c}_a^{j+1} = \frac{N}{2^{j+1}}$

let $\underline{a}^{j+1}$ et $\underline{c}_a^{j+1}$ be respectively vector $\underline{a}$ at resolution j+1 and the associated wavelet coefficients ; vector $\underline{a}^j$ at resolution j is given by :

(23) $$\underline{a}^j = \begin{pmatrix} (D_o^j \tilde{H}^j)^T & (D_e^j \tilde{G}^j)^T \end{pmatrix} \begin{pmatrix} \underline{a}^{j+1} \\ \underline{c}_a^{j+1} \end{pmatrix}$$

II.5. Decomposition of a convolution operator

Let F^j be a filtering operator at resolution j. F^j is defined as follows :

Let $(m,n) \in [1, \frac{N}{2^j}] \times [1, \frac{N}{2^j}]$. We set :

(24) $F^j(m,n) = f_{|m-n|}$ if $|m-n| \in [1, \frac{N}{2^{j+1}}]$

$= f_{|\,(|m-n|-\frac{N}{2^j})|}$ if $|m-n| \in [\frac{N}{2^{j+1}}, \frac{N}{2^j}]$

The definition of F^j requires that it be both symmetric and circulant. In addition, it is assumed that :

(25) $\sum_k f_k = 1$ and $f_{-k} = f_k$

where f_k is the Point Spread Function.

Let $\underline{b}^j$ be vector $\underline{b}$ at resolution j. It is assumed that $\underline{a}^j$ results from filtering of F^j by $\underline{b}^j$, plus a term $\underline{w}^j$ of Gaussian zero mean white noise :

(26) $\underline{a}^j = F^j \underline{b}^j + \underline{w}^j$

Let $\underline{b}^{j+1}$ and $\underline{a}^{j+1}$ be respectively vectors $\underline{b}$ and $\underline{a}$ at resolution j+1, and $\underline{c}_b^{j+1}$ and $\underline{c}_a^{j+1}$ the associated wavelet coefficients.

Our aim is to find how these four terms are related at resolution j+1, from our knowledge of relation (26) at resolution j.

From equations (22) and (26), we deduce :

(27) $$\begin{pmatrix} \underline{a}^{j+1} \\ \underline{c}_a^{j+1} \end{pmatrix} = \begin{pmatrix} (D_o^j H^j) & 0 \\ 0 & (D_e^j G^j) \end{pmatrix} \begin{pmatrix} F^j \underline{b}^j + \underline{w}^j \\ F^j \underline{b}^j + \underline{w}^j \end{pmatrix}$$

and relation (23) giving exact reconstruction for $\underline{b}^j$ is written as :

(28) $$\underline{b}^j = \begin{pmatrix} (D_o^j \tilde{H}^j)^T & (D_e^j \tilde{G}^j)^T \end{pmatrix} \begin{pmatrix} \underline{b}^{j+1} \\ \underline{c}_b^{j+1} \end{pmatrix}$$

Substituting (28) into (27), we obtain :

$$(29)\qquad \begin{pmatrix} \underline{a}^{j+1} \\ \underline{c}_a^{j+1} \end{pmatrix} = \begin{pmatrix} F_{H\tilde{H}}^{j+1} & F_{H\tilde{G}}^{j+1} \\ F_{G\tilde{H}}^{j+1} & F_{G\tilde{G}}^{j+1} \end{pmatrix} \begin{pmatrix} \underline{b}^{j+1} \\ \underline{c}_b^{j+1} \end{pmatrix} + \begin{pmatrix} \underline{w}^{j+1} \\ \underline{c}_w^{j+1} \end{pmatrix}$$

An observation of (29) reveals that the relations between $\underline{a}^{j+1}$ and $\underline{b}^{j+1}$, and between $\underline{c}_a^{j+1}$ and $\underline{c}_b^{j+1}$ are not trivial, due to the presence of the terms $F_{G\tilde{H}}^{j+1}$ et $F_{G\tilde{H}}^{j+1}$.

It can be shown that the operators $F_{H\tilde{H}}^{j+1}$, $F_{H\tilde{G}}^{j+1}$, $F_{G\tilde{H}}^{j+1}$ and $F_{G\tilde{G}}^{j+1}$ exhibit the same properties as F^j,i.e. , they are circulant and symmetric matrices.

Let $f_{k,H\tilde{H}}^{j+1}$ denote the k^{th} column of matrix $F_{H\tilde{H}}^{j+1}$. By the same token, let f_k^j denote the k^{th} column of matrix F^j. The following can be stated :

$$(30)\qquad \sum_n f_{k,H\tilde{H}}^{j+1}(n) = \sum_n f_k^j(n) = 1$$

In practice, due to the properties of these matrix, only one line needs to be computed, and the others are found by circular permutation. The following is then true :

$$(31)\qquad f_{k,H\tilde{H}}^{j+1} = (D_o H^j \tilde{H}^j)\, \underline{f}_{2k-1}^j$$

III. DECONVOLUTION AND BIORTHOGONAL WAVELET TRANSFORM

III.1. Approximation of equation (29)

From a practical viewpoint, if F^j is a low pass filter, it can be shown that, for the first line of the partitioned matrix, the following property exists :

$$(32)\qquad \| F_{H\tilde{G}}^{j+1} \| << \| F_{H\tilde{H}}^{j+1} \|$$

Furthermore, the following is also generally true :

$$(33)\qquad \| \underline{c}_b^{j+1} \| < \| \underline{b}^{j+1} \|$$

We can then write :

$$(34)\qquad \underline{a}^{j+1} = F_{H\tilde{H}}^{j+1} \underline{b}^{j+1} + [\underline{e}^{j+1} + \underline{w}^{j+1}]$$

where the error $\underline{e}^{j+1}$ is given by :

$$(35)\qquad \underline{e}^{j+1} = F_{H\tilde{G}}^{j+1} \underline{c}_b^{j+1}$$

Figure (III-1) represents the SNR between $\underline{a}^{j+p}$ and $F_{H\tilde{H}}^{j+p}$ $\underline{b}^{j+p}$ with p ∈ [1,4] and j=0, and the noise term $\underline{w}^j$ null. Vector $\underline{b}$ is a row of the image. The PSF is a Gaussian with a width of 7 and variance $\sigma^2 = 5$.

It was observed that, for our filter, $\underline{e}^{j+p}$ actually is negligible. In fact, the error obtained is of the same order of - or even lower than - that introduced by 8-bit linear quantization. Henceforth, we will consider only the noise term $\underline{w}^{j+p}$ as suggested by inequations (32) and (33).

It is then deduced that, for resolutions j+p, the following applies recursively :

$$(36)\qquad \forall p \in [1,J] \qquad \underline{a}^{j+p} = F_{H\tilde{H}}^{j+p} \underline{b}^{j+p} + \underline{w}^{j+p}$$

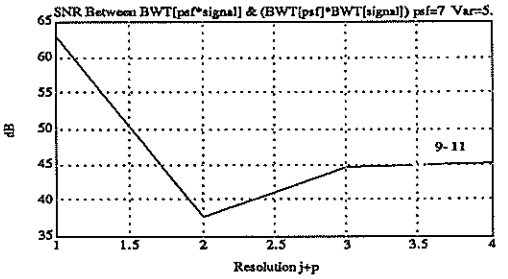


Figure (III-1)

III.2. A multiresolution deconvolution method

Proposition :

Let $\underline{a}^j$ and $\underline{b}^j$ be related as follows, at resolution j :

$$(37)\qquad \underline{a}^j = F^j \underline{b}^j + \underline{w}^j$$

An integer Jmax exists such that :

$$(38)\qquad F_{H\tilde{H}}^{j+Jmax} \approx I^{j+Jmax}$$

In a practical illustration of the above proposition, we show, in **figure (III-2)** the variation of the PSF versus resolution, for j+p = 1 to 4. The PSF is the same as in **figure (III-1)**. The filter used for the Biorthogonal Wavelet Transform is the "9-11" too.

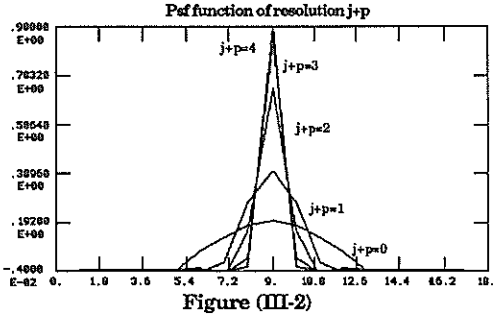


Figure (III-2)

It can be observed that for j+p ≥3, the PSF is very similar to a Dirac function. Consequently, when the resolution encreases the Fourier Transform of the PSF lead to a constant : the Zero of the PSF Power Spectral Density tend to desappear.

In **Figure (III-3)** depicts the Peak SNR between a blurred image and a noised blurred image at different resolutions (the 2D PSF is a separable form obtained with the 1D PSF above). It is observed another interesting property of the multi-resolution analysis : more the resolution encreases less the noise effects will be felt.

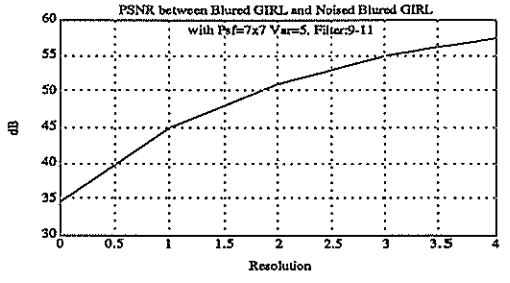


Figure (III-3)

According to the above considerations, the greater the resolution, the easier the restoration is. Therefore we are going to restore the signal at a great resolution using equations (31,36) and then try to obtain the restored signal at the initial resolution.

IV. IMAGE DECONVOLUTION AND BIORTHOGONAL WAVELET TRANSFORM

Following [Mal] we use a 2D Wavelet Transform in which horizontal and vertical orientations are considered preferential.

We suppose that the 2D PSF is a separable form, therefore our deconvolution method is extented easily to 2D signals.

In our restauration example, we choose the following algorithm :

i) The image at resolution j=0 is blurred with the 2D PSF defined in **figure (III-3)**.

ii) The image i) is decomposed at resolution j+p=1 using the 2D Wavelet Transform algorithm.

iii) The image ii) is restored using a Kalman filter [BFBM].

iv) Using only the image iii), we synthetise an image at resolution j=0.

The image v) is the original one at the resolution j=0.

i) Blurred Image for j=0 PSNR = 20.8 dB

ii) Blurred Image for j+p=1 PSNR = 28.6 dB

iii) Restored Image for j+p=1 PSNR = 36.9 dB

Original Image for j+p=1

iv) Restored Image for j=0 PSNR = 29.3 dB

v) Original Image for j=0

V. CONCLUSIONS

It can be observed that at resolutions j+p=1 and j=0 the PSNR gain is about 8 dB. The restoration at scale 1 is easier rather than scale 0 (the SNR is higher). Furthermore the processing time ratio between the restoration at resolutions j=0 and j+p=1 is about four.

Future works will include restoration in wavelet subspaces at several resolutions.

VI. REFERENCES

[ABMD1] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies "Image Coding Using Vector Quantization in the Wavelet Transform Domain" IEEE ICASSP 90.

[ABMD2] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies "Image Coding Using Wavelet Transform" submitted for publication.

[CDF] A. Cohen, I. Daubechies and J.C. Fauveau, "Biorthogonal bases of compactly supported wavelets", AT&T Bell Laboratories preprint.

[Daub1] I. Daubechies, "Orthonormal bases of compactly supported wavelets", Comm. Pure Appl. Math. 41 (1988) 909-996.

[Daub2] I. Daubechies, "Orthonormal bases of compactly supported wavelets. II. Variations on a theme, AT&T Bell Laboratories preprint.

[Mal] S. Mallat "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", IEEE Trans on Pattern Anal. and Mach. intel. Vol. 11 No.7, Jul 89.

[LBB] R. L. Lagendijk, J. Biemond and D. E. Boekee " Hiearchical Blur Identification" IEEE ICASSP 90.

[BFBM] L.Blanc-Feraud, M. Barlaud P. Mathieu " High efficiency images restoration using mirrors images and Kalman filtering " Signal Processing IV : Theories and applications, EURASIP 1988.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

REALIZATION AND PERFORMANCE EVALUATION OF A CLASS OF DISCRETE STATE-SPACE MODELS FOR LINEAR RECURSIVE FILTERING OF NOISY IMAGES

Maria Angela Bedini, Leopoldo Jetto
Dep. of Eletctronics and Automatica, University of Ancona
via Brecce Bianche, 60131 Ancona, Italy

ABSTRACT. *Different causal and semi-causal image correlation models are proposed and a discrete, linear, stochastic state-space representation is derived for each of them. Their performance in restoring noisy images is evaluated and compared in terms of maximum precision theoretically obtainable, signal-to-noise ratio improvement and computational cost.*

1. INTRODUCTION

Image models usually proposed in the literature require either the "a priori" knowledge of the image autocorrelation function or the use of identification-estimation algorithms [1-4].

A new approach recently proposed [5] is based on the following hypotheses:

SMOOTHNESS ASSUMPTION:*the image is modelled by the union of open disjoint subregions whose interior is enough regular to be well described by a 2-D differentiable surface.*

In this case the signal which represents the gray level and its spatial derivatives up to a fixed order $\bar{n}$ can be assumed to be a state vector.

STOCHASTIC ASSUMPTION:*all the derivatives of order $\bar{n}+1$ of the 2-D signal are modelled by means of zero-mean independent gaussian random fields.*

The boundary of each homogeneous subregion is constituted by the image edges. Clearly the image is not differentiable on these discontinuities. The filter can be adapted to their presence by introducing a linear edge detector operator according to the technique described in [5].

In this paper it is shown that, from the previous assumptions, several different image models can be constructed without the knowledge "a priori" of the image autocorrelation function or the use of identification-estimation algorithms. The only information required "a priori" is the geometry of the dependence scheme assumed between adjacent pixels. A class of these image models is here considered, a state-space realization is derived and their performance in restoration problems is evaluated and compared.

2. STATE-SPACE REALIZATION

Let x(r,s) be the value of the original homogeneous monochromatic image at spatial coordinate (r,s). Because of the smoothness assumption the following vectors can be defined

$$X(r,s) := [\partial^n / \partial r^{n-\alpha} \partial s^{\alpha}(x(r,s)); \quad n = 0,1,...,\bar{n} \ ; \ \alpha = 0,1,...,n]^T,$$
$$(\bar{n}+1)(\bar{n}+2)/2 \times 1$$

$$W_r(r,s) := [\partial^{\bar{n}+1} / \partial r^{\bar{n}-\alpha+1} \partial s^{\alpha}(x(r,s)), \alpha = 0,1,...,\bar{n}]^T,$$
$$(\bar{n}+1) \times 1$$

$$W_s(r,s) := [\partial^{\bar{n}+1} / \partial r^{\bar{n}-\alpha} \partial s^{\alpha+1}(x(r,s)) \ , \alpha = 0,1,...,\bar{n}]^T.$$
$$(\bar{n}+1) \times 1$$

X(r,s) is assumed to be the state vector, it is composed of $N=(\bar{n}+1)(\bar{n}+2)/2$ elements. Now if it is put: $r = r(u) = r_0 + g\,u$, $s = s(u) = s_0 + bu$, it can be shown that the following *Homogeneous Image Equation* (*HIE*) holds [6]

$$\dot{X}(r(u),s(u)) = \dot{r}(u)\, A\, X(r(u), s(u)) + \dot{s}(u)\, A' X(r(u)\, s(u)) + \dot{r}(u)\, B\, W_r(r(u), s(u)) + \dot{s}(u)\, B\, W_s(r(u), s(u)). \tag{2.1}$$

where A, A' and B are suitably defined (NxN) matrices with elements 0 and 1. Moreover A and A' commute.

It is considered the situation where the value of the sampled image $x_{i,j} := x(i\Delta_r, j\Delta_s)$ is observed under additive white gaussian noise $v_{i,j} \sim \mathcal{N}(0, \sigma_v^2)$:

$$y_{i,j} = x_{i,j} + v_{i,j} \ . \tag{2.2}$$

To obtain a state-space representation of the image, the following semi-causal dependence model is assumed:

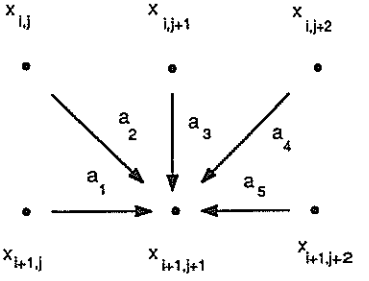


Fig. 1

The coefficients a_i, (i=1,...,5) may be one or zero according to whether the corresponding pixel connection is considered or not. Hence several correlation models can be derived from the general scheme of Fig. 1. In this paper the following dependence models are considered:

(1) *One-Point One-Dimensional Causal Model*: $a_1=1, a_2=a_3=a_4=a_5=0$.
(2) *Two-Points Two-Dimensional Causal Model*: $a_1=a_3=1, a_2=a_4=a_5=0$.
(3) *Three-Points Two-Dimensional Causal Model*: $a_1=a_2=a_3=1, a_4=a_5=0$.
(4) *Three-Points Two-Dimensional Semicausal Model*: $a_1=a_3=a_5=1, a_2=a_4=0$.
(5) *Five-Points Two-Dimensional Semicausal Model* : $a_1=a_2=a_3=a_4=a_5=1$.

The relations between the state $X_{i+1,j+1}$ evaluated at pixel (i+1,j+1) and the state evaluated at the neighbouring pixels can be obtained by integrating the *HIE* along the directions indicated in Fig.1. Exploiting the commutativity of the two semigroups $e^{A'\Delta_s}$ and $e^{A'\Delta_r}$ one obtains the following *Constitutive Equations* of the discrete model:

$$X_{i+1,j+1} = H_1 X_{i+1,j} + W^{(1)}_{i+1,j+1}, \quad (2.3)$$
$$X_{i+1,j+1} = H_2 X_{i,j} + W^{(2)}_{i+1,j+1}, \quad (2.4)$$
$$X_{i+1,j+1} = H_3 X_{i,j+1} + W^{(3)}_{i+1,j+1}, \quad (2.5)$$
$$X_{i+1,j+1} = H_4 X_{i,j+2} + W^{(4)}_{i+1,j+1}, \quad (2.6)$$
$$X_{i+1,j+1} = H_5 X_{i+1,j+2} + W^{(5)}_{i+1,j+1}. \quad (2.7)$$

where: $H_1 := e^{A'\Delta_s}$, $H_3 := e^{A\Delta_r}$, $H_5 := e^{-A'\Delta_s}$,
$H_2 := e^{A'\Delta_s + A\Delta_r} = H_1 H_3$, $H_4 := e^{-A'\Delta_s + A\Delta_r} = H_3 H_5$.

and where $W^{(l)}_{i+1,j+1}$, (l=1,...,5), is given by the convolution integral representing the forced state evolution.

An ensemble of pixels composed of L adjacent columns j= j1,...,jL, and two adjacent rows i and i+1 is now considered. Eqns. (2.3)-(2.7) can be written for each pixel (i+1,j) with (j=j2,...,JL-1), while (2.3)-(2.5) and (2.5) - (2.7) can be written for pixels (i+1,j1) and (i+1,jL) respectively.The following notation is now introduced

$$a_{l,l+k} := a_l + a_{i+1} + \ldots + a_{l+k}$$

and the following vectors and band-matrices are defined:

$$\underset{NLx1}{\mathbf{X}_{i+1}} := [X_{i+1,j1},\ldots,X_{i+1,jL}]^T,$$

$$\underset{NLx1}{Z_{k,i}} := \Big[\frac{1}{a_{3,5|=3}}\sum^{5} W^{(l)}_{i+1,j1} \;\; \frac{1}{a_{1,5|=1}}\sum^{5} W^{(l)}_{i+1,j2} \cdots \frac{1}{a_{1,5|=1}}\sum^{5} W^{(l)}_{i+1,jL-1} \;\; \frac{1}{a_{1,3|=3}}\sum^{5} W^{(l)}_{i+1,jL}\Big]^T. \quad (2.8)$$

$$\underset{NLxNL}{F_k} = \begin{bmatrix} 0 & a_5 H_5/a_{1,3} & 0 \ldots & 0 \\ a_1 H_1/a_{1,5} & 0 & a_5 H_5/a_{1,5} \ldots & 0 \\ \cdot & \cdot & \cdot & \\ 0 \ldots & 0\; a_1 H_1/a_{1,5} & 0 & a_5 H_5/a_{1,5} \\ 0 \ldots & 0 & a_1 H_1/a_{1,3} & 0 \end{bmatrix} \quad (2.9)$$

$$\underset{NLxNL}{E_k} = \begin{bmatrix} a_3 H_3/a_{3,5} & a_4 H_4/a_{3,5} & 0 \ldots & 0 \\ a_2 H_2/a_{1,5} & a_3 H_3/a_{1,5} & a_4 H_4/a_{1,5}\; 0 \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 \ldots & a_2 H_2/a_{1,5} & a_3 H_3/a_{1,5} & a_4 H_4/a_{1,5} \\ 0 \ldots & 0 & a_2 H_2/a_{1,3} & a_3 H_3/a_{1,3} \end{bmatrix} \quad (2.10)$$

The subscript k ,(k=1,...,5) introduced in the above vectors and matrices identifies one particular dependence model among those here considered. For each of them the corresponding $Z_{k,i}$, F_k and E_k can be derived from (2.8)-(2.10) by suitably assigning to the a_i 's the values zero or one.

The totality of equations that can be written for the pixels (i+1,j), j=j1,...,jL, can be expressed in the compact form of the equation:

$$\mathbf{X}_{i+1} = (I-F_k)^{-1} E_k \mathbf{X}_i + (I-F_k)^{-1} Z_{k,i}. \quad (2.11)$$

From (2.2) the following measure equation can be associated to eqn. (2.11):

$$Y_{i+1} = C X_{i+1} + V_{i+1}. \quad (2.12)$$

where:
$\underset{Lx1}{Y_{i+1}} := [y_{i+1,j1},\ldots,y_{i+1,jL}]^T$, $\underset{Lx(NL)}{C} := \text{diag}\{C'\}$,

$\underset{Lx1}{V_{i+1}} := [v_{i+1,j1},\ldots,v_{i+1,jL}]^T$, $\underset{1xN}{C'} = [1,\ldots,0]$.

Exploiting the stochastic assumption it can be proved that $Z_{k,i}$ is a white noise sequence so that eqns. (2.11) and (2.12) are suitable for Kalman filtering implementation. The expression of the covariance matrix $Q_{z,k} := \mathcal{E}[Z_{k,i} Z_{k,i}^T]$ is reported in [6].

3. PERFORMANCE EVALUATION

Numerical experiments have been carried out on the two following different kinds of images:

A) homogeneous field (64 x 64) generated by

$x_{i,j} = A_0 + \Sigma_{l=1,4} A_l \cos(l\,\omega_0(i-1)\Delta + (j-1)\Delta)$

where : $A_0 = 5$, $A_1 = 10$, $A_2 = 9$, $A_3 = 8$, $A_4 = 7$,

$\omega_0 = 0.4\,\pi$, $\Delta = 5/63$.

B) simulated eight-bit image (256 x 256) consisting of concentric rings (see Fig.2)

The originals have been corrupted by zero-mean white gaussian noise with a variance such that Signal variance/Noise variance = 1. According to the image representation of eqns. (2.11) and (2.12) the Kalman algorithm has been implemented as a strip filtering processor [2] by partitioning the image into parallel strips of width L. Both images have been processed using the steady-state value of the filter gain computed off line. The filter performance has been measured through the signal-to-noise ratio improvement η defined as:

$$[\eta]_{dB} = 10 \log_{10} \Sigma_i \Sigma_j (y_{i,j} - x_{i,j})^2 / \Sigma_i \Sigma_j (\bar{x}_{i,j} - x_{i,j})^2 .$$

The homogeneous field has been processed with three models of increasing order corresponding to the choice $\bar{n} = 0,1,2$ respectively. The image B has been filtered with a model of order corresponding to the choice $\bar{n} = 0$ in that it can be considered a piecewise constant image. In this case the filter has been adapted to the presence of the image edges by introducing a linear edge-detector operator and using the technique proposed in [5]. An "a priori" evaluation of the filter performance has been carried out by computing the inferior limit of $\|P_\infty\|$ (where P_∞ is the steady-state solution of the Riccati equation). This limit represents the maximum precision theoretically obtainable for the state. It is given by [6]

$$\|P_\infty\| \geq \gamma_1/(1 + \sqrt{1 + \gamma_1\gamma_2}) ,$$

where :

$\gamma_1 = \|(I-F_k)^{-1}Q_{zk}(I-F_k)^{-T}\| / (1+\|(I-F_k)^{-1}E_k\|^2)$,

$\gamma_2 = \|C^T R^{-1} C\|$, $R := \mathcal{E}[V_i V_i^T]$.

Numerical results are summarized in the tables and figures reported in the end of the paper.

4. CONCLUSIONS

As expected the monodimensional model produces less good filter performances in comparison with the other models. Note the creeping effect on Fig. 4. A considerable improvement in the filter performance is introduced by the two-dimensional causal models both in terms of η and in terms of visual quality of the filtered image (see Table 1 and Figs. 5,6).A further improvement is introduced by the semi-causal models (see Table 1 and Figs. 7,8).

All the two-dimensional models have an approximately equal storage requirement while the computational cost of the semicausal models tends to become prevailing for increasing values of $\bar{n}$. In fact they require a higher number of iterations for filter stabilization. Note also that both for causal and semi-causal models the filter performance is lightly better without the pixels along the diagonal directions (as foreseen by the corresponding values of $\|P_\infty\|$). This is probably due to the fact that all the information contained in the diagonal direction can be decomposed into a vertical and a horizontal component as indicated by the relations existing between the semigroups H_1, H_2 and H_3. So, it can be conjectured that the diagonal directions add no significant contribution to the information carried by the horizontal and vertical directions.

In conclusion a new class of image models has been proposed whose main feature is the generality of underlying hypotheses. This makes it possible to derive a state-space representation of the image in a straightforward manner, whichever is the dependence model between adjacent pixels that has been assumed "a priori". Moreover, on the basis of experimental results high filter performances seem to be really attainable.

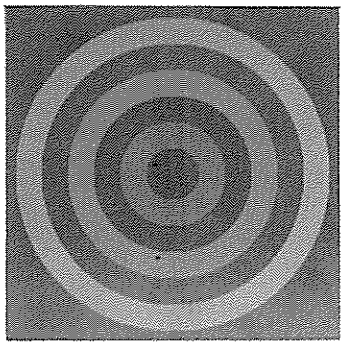


Fig. 2 - Original Image B.

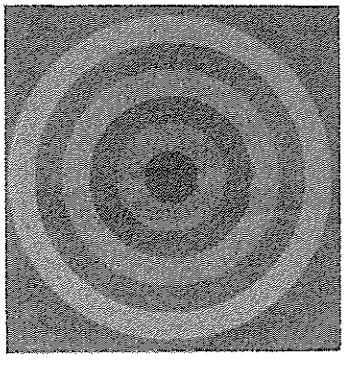


Fig. 3- Noisy Image B.

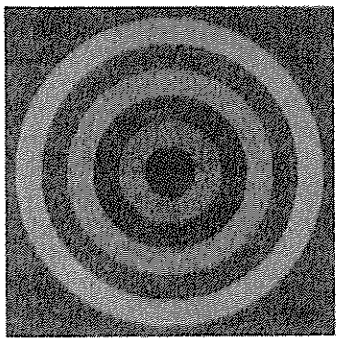


Fig. 4 - Filtered Image B (Mod. N° 1).

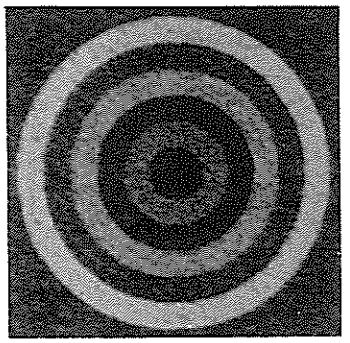


Fig. 5 - Filtered Image B (Mod. N° 2).

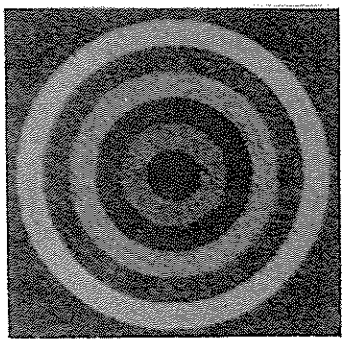


Fig 6 - Filtered Image B (Mod. N° 3).

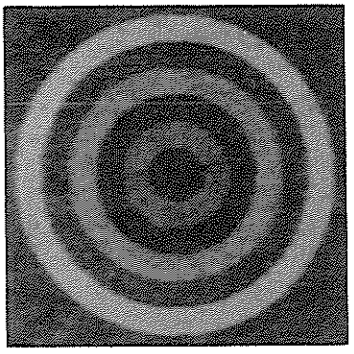


Fig. 7 - Filtered Image B (Mod. N° 4).

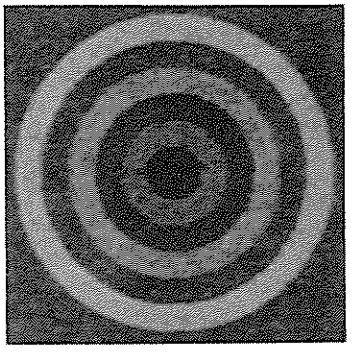


Fig. 8 - Filtered Image B (Mod. N° 5).

MODEL		[η]		$\|P_\infty\|$		q	
		A	B	A	B	A	B
	$\bar{n}=0$	3.59	5.56	4.74	11.4	18	16
1	$\bar{n}=1$	3.77		89.9		25	
	$\bar{n}=2$	2.83		791		33	
	$\bar{n}=0$	5.51	8.84	7.39	17.7	16	14
2	$\bar{n}=1$	6.70		128		26	
	$\bar{n}=2$	7.60		1010		42	
	$\bar{n}=0$	5.42	8.40	7.44	17.8	16	14
3	$\bar{n}=1$	6.37		128		26	
	$\bar{n}=2$	6.99		1010		48	
	$\bar{n}=0$	6.41	10.18	4.74	10.7	17	17
4	$\bar{n}=1$	8.22		89.8		60	
	$\bar{n}=2$	8.87		790		118	
	$\bar{n}=0$	6.48	10.17	4.81	11.2	17	17
5	$\bar{n}=1$	8.10		90.8		62	
	$\bar{n}=2$	8.58		797		147	

Table 1. Results relative to Imges A and B, q is the iteration number for filter stabilization.

REFERENCES

[1] S. R. POWELL and L. M. SILVERMAN: *Modeling of two-imensional covariance functions with applications to image enhancement*, IEEE Trans. Autom. Contr., Vol. AC-19, pp. 8-13,1974

[2] J. W. WOODS and C. H. RADEWAN: *Kalman filtering in two dimensions*, IEEE Trans. Inf.Theory, Vol. IT-23,pp. 473-482, I977.

[3] H. R. KESHAWAN and M. D. SRINATH: *Enhancement of noisy images using an interpolative model in two dimensions*, IEEE Trans. on Syst., Man, and Cybern., Vol. SMC-8, pp. 247-258, I978.

[4] H. KAUFMAN, J. W. WOODS, S. DRAVIDA and A. M. TEKALP: *Estimation and identification of two-dimensional images*, IEEE Trans. Aut. Contr., Vol. AC-28, pp. 745-756, I983.

[5] A. GERMANI and L. JETTO: *Image modeling and restoration : a new approach,* Cicuits Syst. and Signal Process, Vol. 7,pp427-457,1988

[6] M.A. BEDINI and L.JETTO: *Realization and performance evaluation of a class of discrete state-space models for linear rcursive filtering of noisy images,* Int. Rep. n1/90, Dip. di Elettronica e Automatica, Univ. di Ancona,1990.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Comparison of some morphological segmentation algorithms based on contrast enhancement - Application to automatic defect detection -

Philippe Salembier
Signal Processing Laboratory
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland.
Tel : (21 41) 693 27 08

Abstract : In this paper, several morphological algorithms for contrast enhancement are defined and compared. They are used to highlight peaks and ridges in an image in order to perform a segmentation based on local contrast. Starting with the classical *top hat* transform, improvements can be obtained by two different approaches. The first one consists of taking into account the possible anisotropy contained in an image. The second one tries not only to compensate for the slow variation of the local grey level mean but also to reduce the background noise. The performances of the various algorithms are compared with a Student *t* test which evaluates the separation in mean of two populations. The algorithms which give the best results, are then applied to an automatic defect detection problem. It is shown that, once the contrast enhancement has been performed, the actual detection part is a very simple and easy operation.

1 Introduction

Mathematical morphology provides a shape based approach to image analysis. It was derived from set theory, and can be used for both binary and grey scale images [1],[2]. The basic principle of any processing operation is to transform the original image into a simpler one by removing irrelevant information. To this goal, the image is locally compared to a predefined object called the structuring element.

This technique is becoming increasingly useful for applications in signal processing and machine vision [3]. One important industrial application is automated visual inspection and more particularly surface defect detection. This study is related to the integration of intelligent process control and inspection in robot finishing. It addresses polishing of castings. Image analysis is used to perform defect detection on 3D metallic surface. Very often, surface defects on products such as metallic pieces, textiles, paper (etc...) can be identified by the analysis of three parameters: the brightness, the shape and the texture. The first two parameters at least, naturally call for mathematical morphology techniques.

In this paper, we are concerned with segmentation algorithms based on contrast enhancement. Mathematical morphology offers a simple algorithm known as *top hat* transform [4]. It is based on the difference between the original image and its morphological opening. The opening process generates a kind of average picture while removing peaks and ridges. Then, the difference with the original highlights those peaks and ridges.

In the next section, it will be shown how this classical algorithm can be improved with two different approaches. The first one consists of taking into account the possible anisotropy contained in the picture, and the second one tries not only to compensate for the slow variations of the DC component but also to reduce the detection noise. The performances of the various algorithms will be compared in the third section. To this end, a Student t test will be used. The last section will be devoted to the application of these algorithms to a particular defect detection problem.

2 Morphological contrast enhancement algorithms

Mathematical morphology applied to digital grey scale images is based on the basic operations of dilation $\oplus$, erosion $\ominus$, opening o and closing $\bullet$ [5]. If I, S and K respectively denotes the original image, the structuring element, and its support region, then :

$$\text{Erosion} : (I \ominus S)(x) = \min_{z \in K} [I(x+z) - S(z)]$$

$$\text{Dilation} : (I \oplus S)(x) = \max_{z \in K} [I(x-z) + S(z)]$$

$$\text{Opening} : I \circ S = (I \ominus S) \oplus S$$

$$\text{Closing} : I \bullet S = (I \oplus S) \ominus S$$

From a visual point of view, erosion shrinks the bright parts of the image whereas dilation expands them. Opening and closing generate nonlinear smooth versions of the original picture.

As stated in the previous section, contrast enhancement can be performed by a *top hat* transform [1],[4] defined by:

Algorithm 1:

$$T = I - (I \circ S)$$

if bright areas have to be detected.

The opening results in a "lower average" (an average which remains below the signal) of the original image, as shown in figure 1 for a 1D signal. Then, the resulting image is insensitive to slow variations of grey tones and it consists of the enhanced peaks and ridges embeded in a random signal which can be considered as a noise.

Of course, the size and shape of the structuring element are of primary importance. For example, if edges have to be detected with accuracy, then the structuring element should be flat. In the same way, if shape information about the area to detect is available, it should be used to define the structuring element. If no such information exists, an isotropic element may be used (circle and its various approximations: octogon, hexagon, square...).

In order to improve the efficiency of the enhancement, the noise power can be reduced. Several techniques can be used to perform this improvement, and are described below:

* Very often, the peaks to be detected are surrounded by non isotropic areas such as textured regions, boundaries of objects etc... If this anisotropy is not taken into account, the opened signal may be far below the original, and this will increase the noise power. This problem can be overcome by using a kind of adaptive structuring element: instead of using a single opening, a linear structuring element is used to perform various openings in several directions. The opening giving the highest value is chosen as being the final results. This improved algorithm can be defined as follows:

  Algorithm 2:

$$T = I - \max_{Direction} \left[I \circ S_{Line,Direction} \right]$$

  where $S_{Line,Direction}$ represents a linear structuring element oriented according to the parameter Direction.

  As shown in the next section, this algorithm has higher performance than the classical *top hat* transform, but the noise power reduction can also be obtained by a completely different approach.

* The noise presence is mainly due to the fact that the opening produces a "lower average" of the original. We propose now to use an "upper average" i.e. a reference which is above the noise signal, but which is not affected by peaks and ridges. This reference can be obtained by a closing followed by an opening on the original as illustrated in figure 1. It is then very easy, by comparison (minimum), to get an image equal to the original one everywhere except for peaks and ridges, and by subtraction to dramatically enhance the contrast. This new algorithm can be summarized by:

  Algorithm 3:

$$T = I - \mathrm{Min}\left[(I \bullet S) \circ S, I \right]$$

* The last algorithm is based on the previous one, but it also takes advantage of the possible anisotropy of the surrounding area. It is defined by:

  Algorithm 4:

$$T = I - \mathrm{Min}\left[\max_{Direction} \left\{ (I \bullet S) \circ S_{Line,Direction} \right\}, I \right]$$

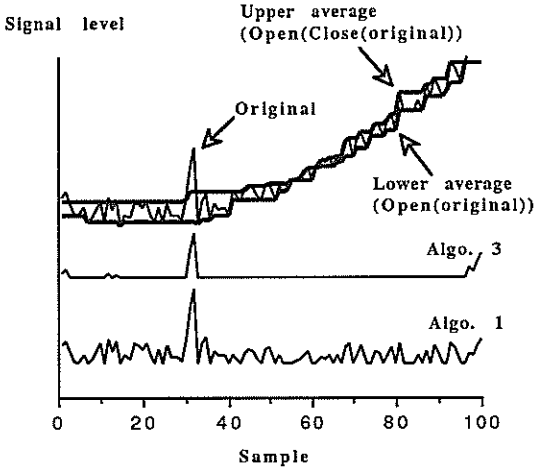


Figure 1: Comparison of lower and upper averages

All these algorithms where defined to detect bright areas. If dark areas have to be extracted, corresponding algorithms can be defined by duality:

Algorithm 1':

$$T = (I \bullet S) - I$$

Algorithm 2':

$$T = \min_{Direction} \left[I \bullet S_{Line,Direction} \right] - I$$

Algorithm 3':

$$T = \mathrm{Max}\left[(I \circ S) \bullet S, I \right] - I$$

Algorithm 4':

$$T = \mathrm{Max}\left[\min_{Direction} \left\{ (I \circ S) \bullet S_{Line,Direction} \right\}, I \right] - I$$

3 Performances of contrast enhancement algorithms

Ultimately, the aim of a contrast extraction algorithm is to distinguish between two populations of pixels. Thus, the algorithm efficiency can be defined in term of a statistical test. Moreover, the two populations are finally

split on the basis of their grey level value using a threshold. This means that, a statistical test sensitive to the average grey level such as the Student t test, can be used. In order to assess the relative performance of the various algorithms, the optimal segmentation result has to be defined. It may be known a priori in the case of synthetic images, or characterized by a human observer in the case of natural images. The t test is then performed on the two populations of pixels derived from the optimal segmentation reference. If m, σ^2 and n denote respectively the mean, the variance and the size of a population, and if the subscript 1 and 2 refers to the two populations, the t test is defined by [6]:

$$t \approx \frac{|m_1 - m_2|\sqrt{n_1 n_2}}{\sqrt{n_1\sigma_1^2 + n_2\sigma_2^2}} \quad , \quad \text{for } n_1, n_2 >> 1$$

This t value can be used to compare the performance of the various algorithms and to optimize their parameters. In order to have reliable conclusions in the case of natural images, the influence of a small error in the definition of the optimal segmentation reference should be examined. The results of such an investigation on Image 1 are reported in figure 2.

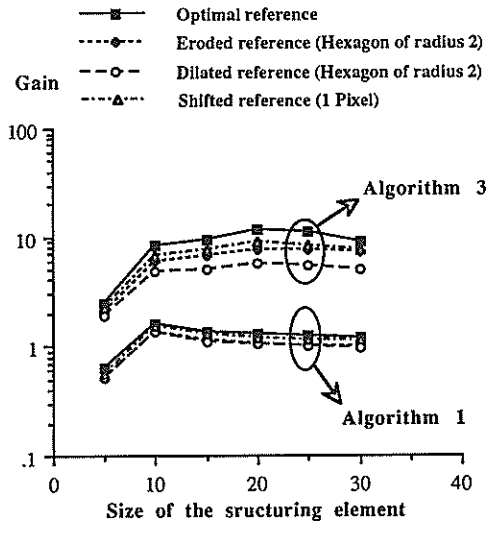


Figure 2 : Influence of the segmentation reference

The separation gain, which is the ratio between the t value after contrast enhancement and the t value computed on the original image, are presented for two different algorithms (Algorithms 1 and 3). In each cases, results are given for the optimal, an eroded, a dilated and a shifted version of the reference. Of course, the absolute value of the gain depends upon this segmentation reference, but it can be concluded that if the error on the reference remains moderate, this t test can be used to perform comparisons and parameters optimization.

The performances of the four contrast enhancement algorithms are presented in figure 3 and 4 corresponding to Image 1 and 2 respectively. These pictures represent views of 3D metallic surface being polished, and were captured with a CCD camera.

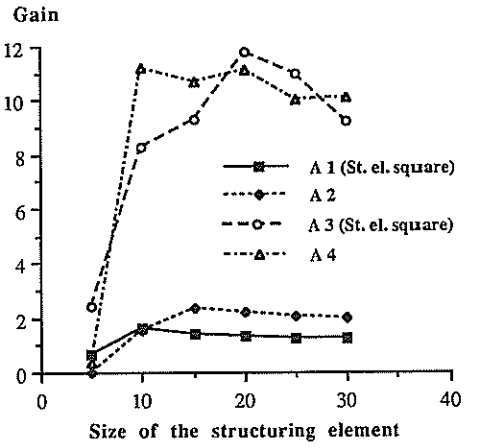


Figure 3 : Gain in contrast for Image 1

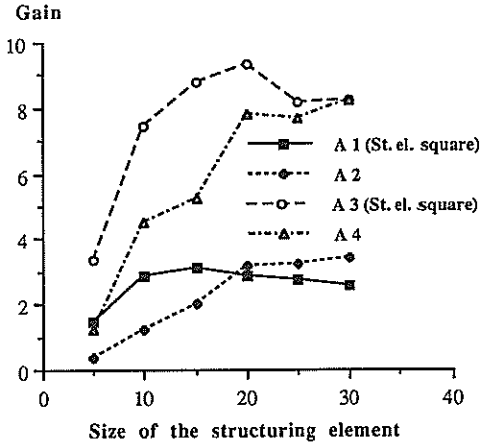


Figure 4 : Gain in contrast for Image 2

It can be seen that, the t test allows the definition of the optimal size of the structuring element, and that in practice, this size should be chosen to be greater than the greatest object to be detected. Moreover, Algorithm 2 which takes advantage of the possible anisotropy performs better than the classical *top hat* transform. This is mainly due to the boundaries between object and background, or between areas with specular reflection or with texture, and finally to the anisotropy of the texture

itself. But a much higher gain can be obtained with the algorithms 3 and 4. One can see the advantage of using an "upper average" instead of a "lower average" as reference. In this case, the background noise has almost completely disappeared.

4 Application to defect detection

As stated in the introduction, this study is related to automatic defect detection on 3D metallic surface. Based on the performance results shown in the previous section, algorithms 3 and 4 were selected to perform the contrast enhancement. Indeed with algorithms 1 or 2, the defects are actually enhanced but the presence of an important background detection noise makes the detection unreliable: it is difficult to chose a segmentation threshold and the shape of the defects depends on this threshold.

With algorithm 3 or 4, the background noise has almost completely disappeared. It consists of very small isolated spikes. Then, it can be considered that each connected set of non zero pixels after the contrast enhancement is a potential defect. In this case, the final segmentation simply consists in removing the detection errors. To this goal, several characteristics such as the surface, the maximum value, the volume (etc...) can be associated to each potential defect. Depending on the particular application, these features can be used to eliminate the detection errors.

In our case, the minimum defect size is not fixed, this eliminates features such as the surface or the volume. But the maximum value of a potential defect is a very reliable feature because of the noise characteristics (the isolated spikes have a very small height). Note that with this procedure, the defect shape is independent of any threshold. Finally, with the set of images we have, it was not possible to see a clear difference between algorithms 3 and 4. Both of them works quite well. As an example, the segmentation results corresponding to images 1 and 2 are presented in images 3 and 4.

5 Conclusions

In this paper, various morphological algorithms for local contrast enhancement were defined and compared. With the intention of segmenting an image, the classical *top hat* transform can be improved in two different ways. The first one takes into account the possible anisotropy contained in the image. The second one not only compensates for the slow variation of the local grey level mean but it also reduces the detection noise. A Student *t* test can be used to compare the relative performances of the various algorithms. It reveals that the second approach is more efficient in term of separability between two populations. Algorithms giving the best results, are then applied to a problem of automatic defect detection on 3D metallic surface. It is shown that they allow an accurate and reliable detection.

References

[1] J.Serra, "Image Analysis and mathematical morphology", Academic Press, London, 1982.

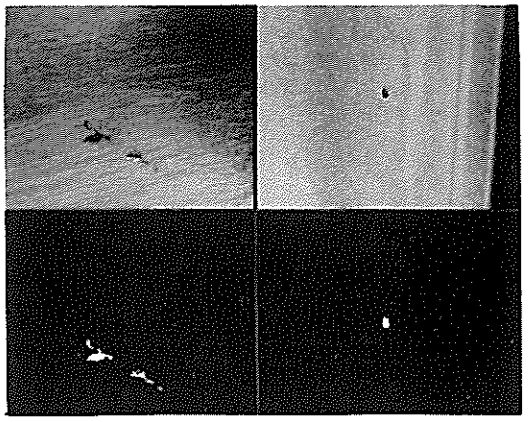
[2] S.R.Sternberg, "Grayscale Morphology", Computer Vision, Graphics and Image Processing 35, pp. 333-355, 1986.

[3] H.Joo and R.M.Haralick, "Understanding The Application of Mathematical Morphology to Machine Vision", Proc. International Conference on Circuit and System, pp. 977-981, 1989.

[4] F.Meyer, "Automatic Screening of Cytological Specimens", Computer Vision, Graphics and Image Processing 35, pp. 356-369, 1986.

[5] R.M.Haralick, S.R.Sternberg, X.Zhuang, "Image Analysis using Mathematical Morphology", IEEE trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI 9, N°4, July 1987.

[6] Y.L.Chou,"Statistical Analysis", 2nd Ed., Holt, Rinehart, Winston, New-York 1975.



Images : 1 2
3 4

Image 1 and 2 : Original pictures

Image 3 and 4 : Segmented pictures

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

MEAN FIELD ANNEALING FOR EDGE DETECTION AND IMAGE RESTORATION[0]

Josiane Zerubia[1,2] and Rama Chellappa[2]

1. INRIA-Sophia - 2004 route des lucioles - 06560 - Valbonne, France
2. USC - Signal and Image Processing Institute - Los Angeles, CA 90089-0272, USA

In this paper, we consider the problem of edge detection and image restoration with images corrupted by an additive Gaussian noise.

We propose a deterministic relaxation method based on Mean Field Annealing with a Compound Gauss-Markov Random Field model. We present a set of iterative equations for the mean values of the intensity and both horizontal and vertical line-processes taking into account some interaction between them. We show the relationship between this technique and the one recently described by Geiger and Girosi. We emphasize on the need of an optimal step-descent method to get a robust algorithm.

Lastly, we present edge detection and image restoration results on a noisy aerial image (SNR = 5 dB) with line-process interaction and compare them with those obtained without such an interaction.

1 INTRODUCTION

We consider herein the problem of edge detection and image restoration in an image corrupted by additive Gaussian white noise. For many years, researchers have worked on this problem as edge detection and image restoration in noisy data are important preliminary steps for various high level vision processes.

Many solutions have been proposed. Among them, the method using line-processes, first introduced by Geman and Geman [6], seems the most promising. They explicitly take into account the cost of creating edges in the form of a line-process. The problem is then to minimize the energy function which is non-convex with respect to the image intensity and the line-processes. One way to obtain the minimum is to use the well-known stochastic technique called "simulated annealing" [6], [8]. This method is optimal (you are able to get the global minimum asymptotically) but it is very time-consuming. Another way to deal with the problem is to search for sub-optimal deterministic algorithms. From a theoretical point of view, the main drawback is that these techniques are not guaranted to give the global minimum. But from a practical point of view, the shape of the energy function obtained for images does not have too many minima and the results are usually good enough even if not optimal. Deterministic algorithms yield solutions at a much lower computational cost.

An interesting deterministic technique is the Graduated Non-Convexity (GNC) algorithm introduced by Blake and Zisserman [2]. An extension of this technique using a Compound Gauss-Markov Random Field (CGMRF) [7] has been derived in [12], [13] for image restoration (the edges are given as a by-product).

Another class of methods coming from equilibrium statistical mechanics [10], [11] relies on the mean field approximation. The basic idea is to substitute the fields at the neighbor-sites by their respective statistical mean value. This approximation is valid only if the fluctuations around the mean are small. It has been used for both image processing and vision problems such as pattern recognition [3] or surface reconstruction [4], [5], [9], [14]. There is no proof of convergence for this technique but it reaches an equilibrium at a given temperature much faster than simulated annealing [1].

[0]The first author would like to thank D. Geiger, Z. Lichtenstein, B. Manjunath, A. Rangarajan and T. Simchony for many useful discussions. The research of the first author during her stay at USC was supported by an INRIA post-doctoral grant.

Recently, we have proposed a method using mean field annealing for a CGMRF model [15]. Herein, we derive a set of iterative equations for the mean values of the intensity and both horizontal and vertical line-processes based on the same CGMRF model but taking into account some interaction between the line-processes. We show the relationship between this method and the one proposed by Geiger and Girosi in [4].

2 MEAN FIELD ANNEALING

2.1 Problem statement

Using a CGMRF model for the image enable us to describe homogeneous regions (with the GMRF) as well as sharp transitions from one region to another (with the line-processes). Calling y the original intensity field and x the noisy one, it can be shown using Bayesian arguments that the corresponding Gibbs energy is given by :

$$E = \frac{1}{2\sigma^2}\sum_{i,j}\{(y_{i,j}-x_{i,j})^2 + \lambda^2(1-2(\theta_x+\theta_y))y_{i,j}^2 + \theta_x(\lambda^2(y_{i,j}-y_{i-1,j})^2(1-l_{i,j}) + \alpha l_{i,j} - \epsilon\alpha\frac{l_{i,j-1}+l_{i,j+1}}{2}) + \theta_y(\lambda^2(y_{i,j}-y_{i,j+1})^2(1-m_{i,j}) + \alpha m_{i,j} - \epsilon\alpha\frac{m_{i-1,j}+m_{i+1,j}}{2})\}$$

where σ^2 is the variance of the Gaussian noise, θ_x and θ_y are the GMRF parameters, λ^2 corresponds to the regularization term which reflects the confidence we have in the data, α is the penalty to be paid to create an edge and ϵ allows to control the amount of propagation of the line due to the interaction between the line-processes ($\epsilon \in [0,1]$). Furthermore, It can be shown [12] that $\lambda^2 = \frac{\sigma^2}{\nu}$ where ν is the variance of the GMRF.

It is interesting to notice that a smoothness constraint on the discontinuity field (l or m) has been introduced (cf our previous model in [15]) by substracting an ϵ term to the energy function. So, the price to be paid to create a discontinuity will be decreased when a discontinuity at a neighboring site is present.

2.2 Proposed algorithm

Using the mean field approximation [4], [5] and [15], we can derive the following iterative equations after some algebraic manipulations:

$$\bar{y}_{i,j} = \frac{1}{1+\lambda^2(1-2(\theta_x+\theta_y))}\{x_{i,j} - \lambda^2\theta_y(\bar{y}_{i,j}-\bar{y}_{i,j+1})(1-\bar{m}_{i,j}) + \lambda^2\theta_y(\bar{y}_{i,j-1}-\bar{y}_{i,j})(1-\bar{m}_{i,j-1}) - \lambda^2\theta_x(\bar{y}_{i,j}-\bar{y}_{i-1,j})(1-\bar{l}_{i,j}) + \lambda^2\theta_x(\bar{y}_{i+1,j}-\bar{y}_{i,j})(1-\bar{l}_{i+1,j})\}$$

with

$$\bar{l}_{i,j} = \sigma_\beta\{\theta_x(\lambda^2(\bar{y}_{i,j}-\bar{y}_{i-1,j})^2 - \alpha + \epsilon\alpha\frac{\bar{l}_{i,j-1}+\bar{l}_{i,j+1}}{2})\}$$

and

$$\bar{m}_{i,j} = \sigma_\beta\{\theta_y(\lambda^2(\bar{y}_{i,j}-\bar{y}_{i,j+1})^2 - \alpha + \epsilon\alpha\frac{\bar{m}_{i-1,j}+\bar{m}_{i+1,j}}{2})\}$$

where σ_β is the sigmoid function, β is proportional to the inverse of the temperature and $\bar{y}$, $\bar{l}$, $\bar{m}$ are the mean fields corresponding to y, l, m.

The results presented above are derived after a few approximations. An exact calculus could be done but would not be of any use from a practical point of view(cf [4] for details about the Transfer Matrix method and the need to obtain a local solution). In order to reduce the computational complexity of this algorithm, we have chosen to get $\bar{y}$ without taking into account the term due to the line-process interaction. This choice has been done after comparing the simulation results obtained with and without this approximation.

The algorithm proposed in [4] by Geiger and Girosi can be obtained from the above model with $\theta_x = \theta_y = \frac{1}{4}$ (see [15] for more details). The threshold and suprathreshold for creating a contour are the same in both cases (i.e. $h_0 = \sqrt{\frac{\alpha}{\lambda^2}(1-\epsilon)}$ and $h_1 = \sqrt{\frac{\alpha}{\lambda^2}}$). It means that, at zero temperature, if the intensity gradient is greater than h_1 a contour will be detected, if it is lower than h_0 a smoothing will be done and in-between the creation of an edge will depend on the presence of an edge at the neighbor-sites.

3 SIMULATION RESULTS

It is important to note that we have used an optimal step descent technique to get the mean field $\bar{y}$ (cf arguments given in section 3.1 of [15]) and that the GMRF parameter estimation has been done using an EM algorithm (see [12] and [15]).

The temperature schedule ($\beta = 0.0002$, then $\beta = \beta * 4$ until $\beta > 1$), the initial conditions (noisy data and line-processes equal to 0.5), the choice of the boundaries and the test of convergence are the same as those used in [15] for the algorithm without line-process interaction.

We present the results obtained on an aerial image (128,128) (see Fig. 1) with a SNR equal to 5 dB (see Fig. 2). We have chosen $\epsilon = 0.3$ (giving a threshold $h_0 = 4.70$ and a suprathreshold $h_1 = 5.65$) and have got the re-

stored image (see Fig. 3) and the edge map (see Fig. 4) after 193 iterations. These results can be compared with those obtained without any line-process interaction (see Fig. 5 and 6) after 182 iterations ($h = 5.50$):

- The quality of the restored images is the same (compare Fig. 5 with Fig. 3).

- The edge map is better with line-process interaction (compare Fig. 6 with Fig. 4).

It is worth pointing out that the introduction of a simple form of interaction between the line-processes gives an algorithm with a reduced sensibility to the choice of the free parameter α.

4 CONCLUSION :

In this paper, we have proposed an extension of our previous algorithm [15] based on mean field annealing using a CGMRF model. As the proposed algorithm is global because of the use of an optimal step descent technique, it would be interesting to look at its implementation on a machine such as the Orthogonal Multi-Processor (OMP) machine presently developed at USC in order to improve its computational cost. Another interesting point would be to look at other kinds of line-process interactions in order to impose more constraints.

5 REFERENCES :

1. G. Bilbro, R. Mann, T. Miller, W. Snyder, D. Van den Bout and M. White, "Optimization by mean field annealing", *Advances in Neural Information Processing Systems*, Touretzky Ed., Vol.1, pp 91-98, 1988.

2. A. Blake and A. Zisserman, "Visual reconstruction", *MIT Press, Cambridge - MA*, 1987.

3. C. Campbell, D. Sherington and K. Y. M. Wong, "Statistical mechanics and neural networks", *"Neural Computing Architecture", MIT Press, Cambridge - MA*, I. Aleksander Ed., pp 239-257, 1989.

4. D. Geiger and F. Girosi, "Parallel and deterministic algorithms for MRFs : surface reconstruction and integration", *Proc. ECCV90*, Antibes, Apr. 1990.

5. D. Geiger and F. Girosi, "Mean field theory for surface reconstruction" *Proc. DARPA Image Understanding Workshop*, pp 617-630, Palo-Alto, May 1989

6. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", *IEEE Trans. Pattern Analysis, Machine Intel.*, Vol. PAMI-6, pp 721-741, Nov. 1984.

7. F. C. Jeng and J. W. Woods, "Image estimation by stochastic relaxation in the compound gaussian case", *Proc. ICASSP88*, pp 1016-1019, New-York, Apr. 1988.

8. S. Kirkpatrick; C. Gelatt and M. Vecchi, "Optimization by simulated annealing", *Science 220*, pp 671-680, 1983.

9. J. Marroquin, "Deterministic Bayesian estimation of Markovian random fields with applications to computational vision", *Proc. ICCV*, pp 597-601, London, Jun. 1987.

10. G. Parisi, "Statistical Field Theory", *Adisson Wesley*, 1988.

11. M. Plischke and B. Bergersen, "Equilibrium Statistical Physics", *Prentice Hall, Englewood Cliffs - NJ*, 1989.

12. T. Simchony, R. Chellappa and Z. Lichtenstein, "Pyramid implementation of optimal step conjugate search algorithms for some low level vision problems", *Proc. Conf. on Computer Vision*, Miami Beach, Dec. 1988.

13. T. Simchony, R. Chellappa and Z. Lichtenstein, "The Graduated Non Convexity algorithm for image estimation using Compound Gauss-Markov Field models", *Proc. ICASSP89*, pp 1417-1420, Glasgow, May 1989.

14. A. L. Yuille, "Energy functions for early vision and analog networks", *Biol. Cybern., 61*, pp 115-123, 1989.

15. J. Zerubia and R. Chellappa, "Mean field approximation using Compound Gauss-Markov Random Field for edge detection and image restoration", *Proc. ICASSP90*, Albuquerque, Apr. 1990.

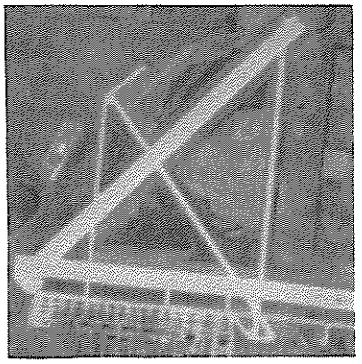


Figure 1

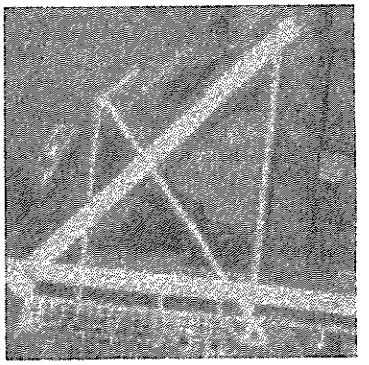


Figure 2

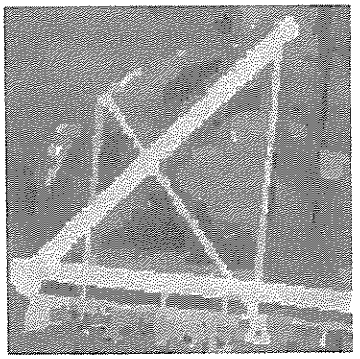


Figure 3

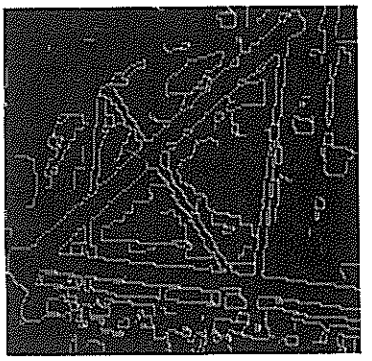


Figure 4

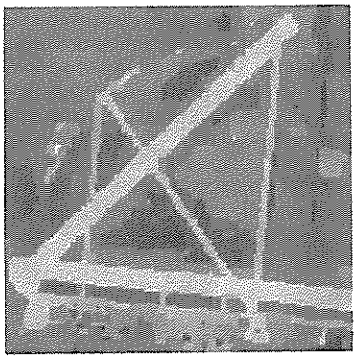


Figure 5

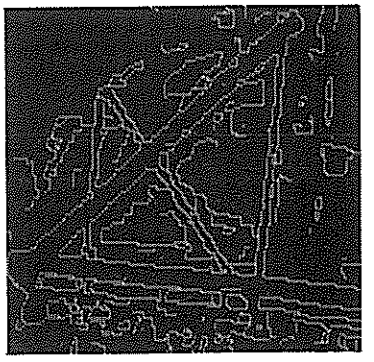


Figure 6

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

SUBBAND CODING OF MONOCHROME IMAGES USING NONSEPARABLE RECURSIVE FILTERS

Mikołaj Bleja and Marek Domański*

Politechnika Poznańska, Instytut Elektroniki i Telekomunikacji
ul. Piotrowo 3a, 60-965 Poznań, Poland

The paper extends the concept of reversive subband coding systems for image processing. Application of IIR filters reduces complexity of the filter banks in the coder and in the decoder. Application of wave digital filters in turn reduces the effects of the finite precision arithmetic. It is shown that special cases of nonseparable 2-D filter banks can be efficiently used in image coding systems. The subband of the lowest spatial frequencies is coded using the DPCM technique while all the other subbands are coded using the highly effective technique proposed by the authors. The effects of quantization of subband signals are briefly considered. The results are illustrated by some experiments with monochrome images.

1. INTRODUCTION

In recent years, *subband coding* (SBC) *of images* is of great interest as a powerful technique of image data compression. The basic idea is to split up the two-dimensional frequency band of an image into some subbands, which are then subsampled. Each of those subbands is coded separately using techniques accurately matched to the statistics of that subband and possibly to the properties of the human visual system in that subband. In the decoder, after upsampling, the subbands are filtered and then summed up in order to reconstruct the original signal.

The most complex parts of an SBC system are the filter banks which are usually implemented as FIR systems. In very recent years, some authors [1]-[7] have proposed to use IIR filters, thereby reducing the computational complexity of the filter banks. In order to avoid phase problems, it has been proposed to use the *reversive systems*, i.e., the systems that process signals in the coder in the direction which is opposite to that applied in the decoder. With the aim of reducing the effects of finite precision arithmetic, it has been proposed [4-7] to apply *wave digital filters* [8].

Because of greater computational costs of nonseparable systems, almost all of the SBC systems considered hitherto in the references are separable. Application of IIR filters reduces the computational costs, and it in turn makes the nonseparable systems more attractive.

*On visit at the Lehrstuhl für Nachrichtentechnik, Ruhr-Universität Bochum, Universitätsstrasse 150, 4630 Bochum, West Germany (as the Alexander von Humboldt Fellow).

2. 2-D WAVE FILTERS BANKS

Consider the basic reversive two-band system with 2-D *wave digital filters* (WDF) (cf. Fig. 1). The whole SBC system consists of some such basic two-band systems, thereby splitting the frequency band up into greater number of subbands.

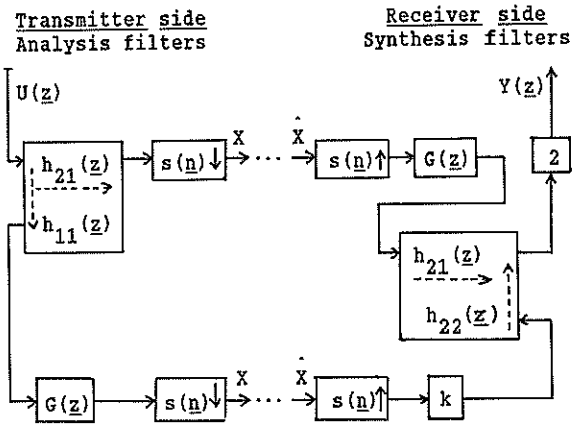


Figure 1

A WDF is described by its transfer matrix

$(\underline{z})=[h_{ij}(\underline{z})]$, where $\underline{z} = (z_1, z_2)$.

The function $s(\underline{n})$ describes sub- and upsampling of a 2-D signal. Subsampling can be viewed as modulation by $s(\underline{n})$

$$s(\underline{n}) = s(n_1,n_2)=(1+(-1)^{\nu_1 n_1+\nu_2 n_2})/2 , \qquad (1)$$

where $\nu_1,\nu_2=0$ or 1 and $\nu_1+\nu_2>0$.

Assume that the filters are either symmetric or antimmetric [9,10]. In the symmetric case,

$$h_{12}(\underline{z})=h_{21}(\underline{z})=(F_a(\underline{z})+F_b(\underline{z}))/2\ , \qquad (2a)$$

$$h_{11}(\underline{z})=h_{22}(\underline{z})=(F_a(\underline{z})-F_b(\underline{z}))/2\ , \qquad (2b)$$

and in the antimmetric case there is (2a) and

$$h_{11}(\underline{z})=-h_{22}(\underline{z})=-j(F_a(\underline{z})-F_b(\underline{z}))/2\ , \qquad (2c)$$

where $F_a(\underline{z})$, $F_b(\underline{z})$ are the all-pass functions. In the symmetric case, the functions $F_a(\underline{z})$ and $F_b(\underline{z})$ are rational with real coefficients. The real parameter k and the the function G have been defined in [5]

k=1 or -1, G=1 or z_1 or z_2 or z_1z_2.

Using the modulation property of the 2-D transforms we get

$$Y(e^{j\underline{\omega}})=U(e^{j\underline{\omega}})H_1(e^{j\underline{\omega}})+U(e^{j(\underline{\omega}+p)})H_2(e^{j\underline{\omega}}), \qquad (3)$$

where $U(\cdot)$ and $Y(\cdot)$ are the input and the output, respectively,

$$H_1(e^{j\underline{\omega}})=G(e^{j\underline{\omega}})\ h_{21}(e^{j\underline{\omega}})h_{21}(e^{-j\underline{\omega}})+$$
$$+kh_{11}(e^{j\underline{\omega}})h_{22}(e^{-j\underline{\omega}})\ , \qquad (4)$$

$$H_2(e^{j\underline{\omega}})=G(e^{j\underline{\omega}})h_{21}(e^{j(\underline{\omega}+p)})h_{21}(e^{-j\underline{\omega}})$$
$$+G(e^{j(\underline{\omega}+p)})h_{11}(e^{j(\underline{\omega}+p)})h_{22}(e^{-j\underline{\omega}}). \qquad (5)$$

The reconstructed signal Y is free of magnitude and phase distortions. The aliasing components are perfectly canceled if [5,7]

$$F_a(e^{j\underline{\omega}})=F_a(e^{j(\underline{\omega}+p)})$$

and $F_b(e^{j\underline{\omega}})=-F_b(e^{j(\underline{\omega}+p)})$. (6)

The overall phase shift of the reconstructed image is exactly zero. Nevertheless, the phase shifts of the subband signals are non-zero. Thus in each row and each column some additional elements should be added in order to preserve the whole visual information.

3. NONSEPARABLE FILTER BANKS

The IIR filter banks are significantly more efficient than the banks of the FIR quadrature-mirror filters usually used. Thus it is possible to use more sophisticated schemes of partitioning of the frequency band. In particular, nonseparable 2-D filters can be used in order to split up the 2-D frequency band into some more relevant subbands (cf. Fig. 2). The partitions of the frequency band shown in Fig. 2 is more computationally expensive than those obtained by use of separable filters. It is because of data interpolation which is needed in implementations of the respective 2-D filters [11].

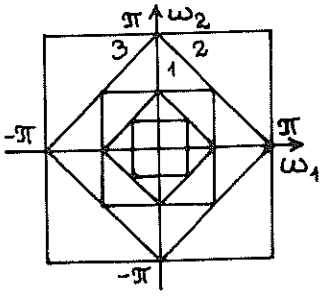


Figure 2

Nevertheless, the scheme from Fig. 2 can be obtained by combining the two-band systems described above. At each step, the three-band system (cf. Fig. 3) is used to obtain the low- and high-frequency 2-D bands. The signal is always spited up with respect to the both frequency coordinates. Between each two steps data rotation must be performed.

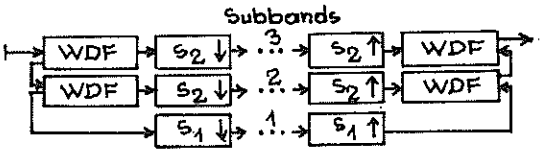


Figure 3

The 2-D WDFs are obtained from their 1-D prototypes by use of the spectral transformations of the type

$z \to z_1z_2$.

The subsampling functions are as follows

$$s_1(\underline{n})=1+(-1)^{n_1+n_2}, \qquad (7a)$$

$$s_2(\underline{n})=(1+(-1)^{n_1})(1+(-1)^{n_2})\ . \qquad (7b)$$

Nevertheless, the experiments show that, in some cases, the subjective quality of images recovered from the signals coded using the scheme from Fig. 2 is slightly better as those obtained by use of the separable scheme with the same number of subbands and the same coding technique.

4. CODING OF THE SUBBAND SIGNALS

4.1. Low-frequency component

Most of the signal energy is related to the lowest subband, therefore this subband should

be coded with the highest fidelity. Lossy DPCM coding inevitably distorts subband images and causes errors which become annoying after interpolation when distortions spread over larger numbers of points. Thus, in many cases, it is reasonable to use lossless coding techniques, in particular, when the sampling density in the subband is decreased more than 16 times with respect to the original 512x512 image [12].

4.2. High-frequency subbands

A substantial reduction of information can be achieved by coding only the detailed parts of those subband images [13]. The active areas of the high-frequency images can be defined by use of the mask $w(\underline{n})$ which is calculated from the function

$$\phi(\underline{n}) = \begin{cases} 1 & \text{if } x(\underline{n}) \geq d, \\ 0 & \text{if } x(\underline{n}) < d, \end{cases} \tag{8}$$

where $x(\underline{n})$ is a sample value of a given subband signal, d is a positive predefined threshold. Then $\phi(\underline{n})$ is processed using the nonlinear filter which acts as "hole-killer" removing single 0's or small groups of 0's from the areas surrounded with 1's [13]. Then, the obtained signal $\gamma(\underline{n})$ is passed through a FIR low-pass filter. The final mask is calculated as the filter output $w(\underline{n})$. The signal values coded and transmitted into a channel are equal to $x(\underline{n})\cdot w(\underline{n})$. The function $w(\underline{n})$ is mask (window) with smoothed edges (Fig. 4a). Edge smoothing prevents emerging of artifacts in the reconstructed images. Significant data compression can be achieved by coding only those regions where $w(\underline{n}) > 0$. The regions are approximated with squares of size from 4x4 up to 16x16. The transmitted signal for squares of size 4x4 is shown in Fig. 4b. Except of the pixel values the positions and sizes of squares must be coded and transmitted as a side information.

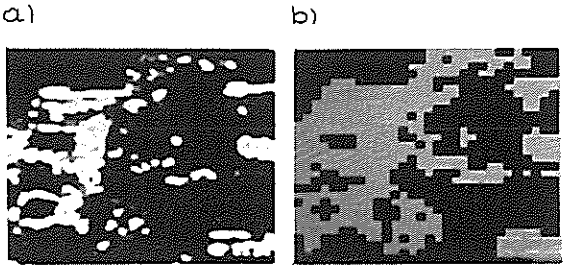


Figure 4

5. QUANTIZATION ERRORS

We update the quantization error analysis made in [14] where the "gain-plus-additive-noise" quantizer model [12] has been used. The effect of slope overload of DPCM technique can be roughly modeled by an FIR low-pass comb filter which in the 1-D case exhibits high attenuation about π. The 2-D filters obtained by use of transformations of the type

$$z \rightarrow z_1 z_2 \tag{9}$$

exhibit stopband shifted accordingly to the transformation applied. Thus some components in the middle of the spectrum of the reconstructed signal are highly attenuated. Moreover, consideration of (5) implies that this dynamic effect of the DPCM technique causes significant aliasing errors.

The analysis given in [14] is fully relevant for the technique described above. Nevertheless, it should be emphasized that the filter design error is in the case of wave digital filters equal to zero (cf. Section 2).

6. EXPERIMENTAL RESULTS

Schemes shown in Fig. 2 have been examined using 512x512 and 256x256 real monochrome images. Of course, for standard TV images the results can be improved by increasing the number of subbands. The systems with 5-th order 2-D separable and nonseparable wave digital filters have been considered. The 1-D lattice prototype IIR WDF's [15]were designed as equivalents to the FIR QMF's of the type 24C [5]. The 2-D WDF's were obtained using the spectral transformation (9).

As an example, the image "ULA" (cf. Fig. 5) has been coded using the 5-bit DPCM technique for the lowest subband and the masking technique for all the other subbands. For bit-rates 0.56, 0.50, and 0.44, the reconstructed images have been shown in Fig. 6, 7, and 8, respectively. In order to compare objective quality of the reconstructed images, the signal-to-noise ratio is computed:

$$SNR = 10 \log_{10} \frac{(255)^2}{e_{ms}^2} \tag{10}$$

where e_{ms}^2 - the average mean square error.

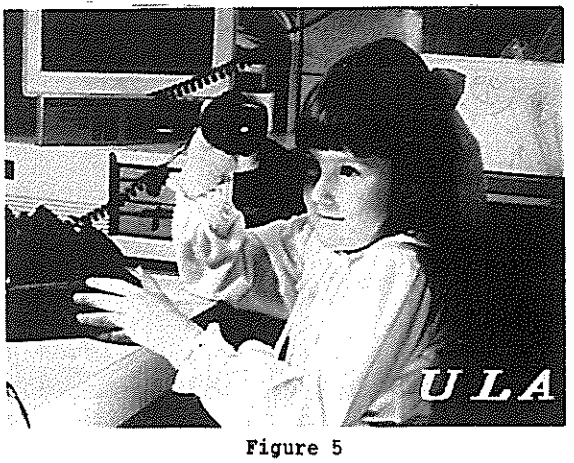


Figure 5

Figure 6

Figure 7

Figure 8

compression ratio	bit/pxl	SNR	threshold d
14.06	0.56	26.84	3
16.02	0.50	27.18	5
17.87	0.44	27.05	7

REFERENCES

[1] Kronander T., A new approach to recursive mirror filters with a special applications in subband coding of images, IEEE Trans. ASSP-36, (1988), 1496.

[2] Ramstad T., IIR filterbank for subband coding of images, in: ISCAS 88 (IEEE, Helsinki 1988) pp. 827-830.

[3] Ramstad T., Husoy J., Parallel complex filterbank for subband coding of images, in: 6th Int. Symp. Networks, Systems, Signal Proc. (Zagreb 1989) pp. 162-165.

[4] Domański M., Cyfrowe filtry rekursywne dla subzakresowego kodowania obrazów, in: KST 88, Bydgoszcz 1988, pp. 254-262, in Polish.

[5] Domański M.: Efficient wave filter banks for sub-band coding of images, to be published.

[6] Bleja M., Domański M., Results in subband coding of monochrome images, in: Int. Symp. Signals, Systems, Electronics (URSI, Erlangen 1989) pp. 314-317.

[7] Domański M., Cyfrowe lp-pasywne filtry dwuwymiarowe (Technical University of Poznań, 1990), in Polish.

[8] Fettweis A., Wave digital filters: theory and practice, Proc. of the IEEE, 74 (1986) 270.

[9] Basu S., Fettweis A., On the synthesiability of multidimensional lossless two-ports, in: Proc. IEEE 1987 Int. Symp. on Circuits and Syst. (IEEE, Philadelphia 1987) pp. 690-693.

[10] Basu S., Fettweis A., On synthesiable multidimensional lossless two-ports, IEEE Trans., CAS-35, (1988) 1478.

[11] Dudgeon D., Mersereau R., Multidimensional digital signal processing (Prentice-Hall, Englewood Cliffs, 1984).

[12] Jayant N., Noll P., Digital coding of waveforms (Prentice-Hall, Englewood Cliffs, 1984).

[13] Bleja M., Domański M., Image data compression using subband coding, to be published.

[14] Westerink et al., Quantization error analysis of image sub-band filter banks, in: Proc. ISCAS 88 (IEEE, Helsinki 1988) pp. 819-822.

[15] Gazsi L., Explicit formulas for lattice wave digital filters, IEEE Trans., CAS-32 (1985) 68.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

IMPLEMENTATION OF BLOCK-ADAPTIVE SUBBAND CODING OF IMAGES ON A TRANSPUTER ARRAY.

Chaouki DIAB, Rémy PROST (member EUSIPCO), Robert GOUTTE (member EUSIPCO).

INSTITUT NATIONAL DES SCIENCES APPLIQUEES de LYON.
Laboratoire de Traitement du Signal et Ultrasons, URA CNRS 1216, bât. 502,
69621 VILLEURBANNE CEDEX, FRANCE.

In a recent paper an error-free image decomposition/ reconstruction method for blockwise subband coding schemes has been proposed. Blockwise subband coding allows both adaptation of the subband number and bits allocation to local changes in image statistics. We investigate here the implementation of this method on a transputer array in order to archive medicine images. The proposed architecture is massively parallel and requires that the transputer number must be multiple of 4.

1. INTRODUCTION

The medicine image archiving is one among the domains where the images compression is very needed. The practical use of the archived images requests that both compression and decompression processes are sufficiently fast.

Subband coding of images is a technique widely used to reduce the images data amount. The image subband coding consists on the splitting of the image into some subimages corresponding to a set of oriented frequency bands. Each subimage is then quantized and coded using an appropriate coder matched to its statistics.The classical approach uses the Quadrature Mirror Filters (QMF's) to eliminate the aliasing that can appear in the splitting process. Unfortunately, the use of finite impulse response filters induces a border effect in the reconstructed image.This phenomenon forbids the coding of the image by blocks, and then deprives the technique of the advantages of an adaptivity of the process to the local statistic changes in the image.

In recent works, we have proposed an image decomposition/ reconstruction method free of border-effects [1] and a block-adaptive subband coding algorithm [2]. In the present work, our objective is the implementation of these algorithms on a transputer array in order to archive medicine images. In section 2, we recall the error-free method. In section 3, we present the block-adaptive algorithm. Finally, we developp the implementation aspect where we propose a massively parallel architecture to get the performances to be required for the application.

2. ERROR-FREE IMAGE DECOMPOSITION/ RECONSTRUCTION METHOD FOR THE SUBBAND CODING SCHEME

2.1. Four-band decomposition process

In order to reduce the complexity of presentation , we consider first, the one dimensional case. Here, the image is reduced to a signal of N-samples.The idea is to use half-band ideal filters implemented in the frequency domain by the Discrete Fourier Transform (DFT).The implementation steps of this technique are illustrated in Fig.1.

The discrete Fourier transform of the ideal half-band filter pair (h_a , g_a) where h_a and g_a are the low-pass and high-pass filters respectively, is defined as follows:

$$(F_k^N(h_a),\ F_k^N(g_a)) = (1,0) \qquad k=0,\ldots,N/4-1$$
$$= (0,1) \qquad k=N/4,\ldots,3N/4-1$$
$$= (1,0) \qquad k=3N/4,\ldots,N-1 \qquad (1)$$

where $F_k^N(x)$ denotes the component k of the DFT of the discrete signal x of length N.The DFT of the original signal s_n^0 is

$$F_k^N(s^0) = \sum_{n=0}^{N-1} s_n^0 \exp(-2\pi ikn/N) \qquad k=0,\ldots,N-1 \qquad (2)$$

Thus, the low-frequency components of the signal are, for k=0,...,N/4-1 and k=3N/4,...,N-1:

$$F_k^{N/2}(s^{-1}) = F_k^N(s^0) \qquad k=0,\ldots,N/4-1$$
$$= F_{k+N/2}^N(s^0) \qquad k=N/4,\ldots,N/2-1 \qquad (3)$$

The high-frequency components of the signal are, for k=N/4,...,3N/4-1:

$$F_k^{N/2}(d^{-1}) = F_{k+N/4}^N(s^0) \qquad k=0,\ldots,N/2-1 \qquad (4)$$

The decimated low- and high-frequency signals S_{-1} and D_{-1} are computed by Inverse Discrete Fourier Transform (IDFT) to give:

$$s_n^{-1} = \frac{1}{N/2} \sum_{k=0}^{N/2-1} F_k^{N/2}(s^{-1}) \exp(2\pi ink/(N/2)) \qquad (5)$$

$$d_n^{-1} = \frac{1}{N/2} \sum_{k=0}^{N/2-1} F_k^{N/2}(d^{-1}) \exp(2\pi ink/(N/2)) \qquad (6)$$

for n=0,...,N/2-1

It is well known that the Fourier transform of a real signal has an even real part and an odd imaginary part. For the DFT, it follows that the first and middle samples of the imaginary part are zeros. In our case, we have:

$$\mathrm{Im}\{F_k^{N/2}(s^{-1})\} \neq 0 \qquad \text{for } k=N/4 \qquad (7)$$

$$\mathrm{Im}\{F_k^{N/2}(d^{-1})\} \neq 0 \qquad \text{for } k=0 \qquad (8)$$

where Im{.} denotes imaginary part. It implies that s_n^{-1} and d_n^{-1} are not real signals.

To preserve the amount of data, we retain only the real parts of s_n^{-1} and d_n^{-1}. This implies that the two imaginary parts of (7) and (8) are set to zero. It can be noted that, from (3) and (4),we have:

$$\mathrm{Im}\{F_0^{N/2}(d^{-1})\} = -\,\mathrm{Im}\{F_{N/4}^{N/2}(s^{-1})\} \qquad (9)$$

The exact reconstruction of the original signal is achieved when we have the complete signals s_n^{-1} and d_n^{-1}. If we transmit only the real part of s_n^{-1} and d_n^{-1}, then a reconstruction error results.This error is evaluated in [1]:

$$\varepsilon_{2p} = 0$$

$$\varepsilon_{2p+1} = \frac{2}{N}(-1)^{p+1}\,\mathrm{Im}\{F_{N/4}^{N}(s^0)\} \quad p=0,\ldots,N/2-1 \qquad (10)$$

Only the even samples are reconstructed without error.The error can be eliminated when we transmit one supplementary sample: $\mathrm{Im}\{F_{N/4}^{N}(s^0)\}$, using (10).

The principle of the reconstruction process is based on the evaluation of the DFT of s_n^{-1} and d_n^{-1} and then upsampling and recombining them to obtain the DFT of the original signal (see Fig.1).

In the decomposition of an MxN image, the decomposition of rows generates two MxN/2 sub-images and M additional samples.The decomposition of columns generates four M/2xN/2 sub-images and N additional samples.The total excess of data is M+N.

2.2. One step multi-band decomposition

The discrete Fourier transform of a row (resp. a column) can be split into $B=2^b$ subbands by one decomposition step.Thus, the decomposition hardware is reduced to one stage using the Fast Fourier Transform (FFT) and the Inverse FFT (IFFT). In this case, both M and N must be powers of 2.

It is proved in [1] that this method needs less multiplications per pixel than the FIR QF's method when $n_0 \geq 16$ and $B \geq 4$ which is the usual case in image data compression, where $2n_0$ is the filters length.

3. BLOCK-ADAPTIVE SUBBAND CODING OF IMAGES

3.1. Block-partitioning of the image

The goal of partitioning is the adaptation of the subband number and bit allocation to local changes in image statistic. Taking into account the local image correlation distance,we define a minimum block size of $N_0 x N_0$ with N_0 is power of 2 and we use the following criterion to merge adjacent blocks of similar statistics:

Step 1: partition the NxN-image into constant size blocks $N_0 x N_0$.The resulting number of blocks is $N_{b0} = (N/N_0)^2$.

Step 2: compute the mean value m_k and the variance σ_k^2 of each block k, and then centre it. Compute the maximum

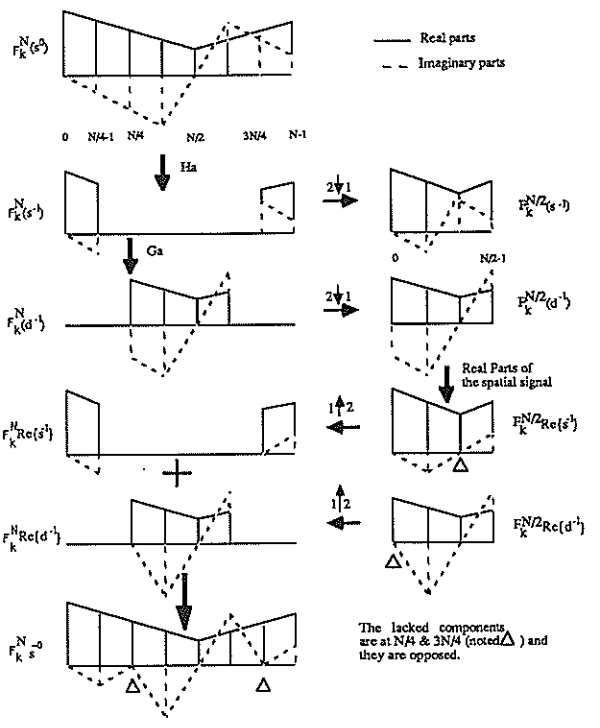


Fig.1 -Illustrative scheme of the two-band analysis/reconstruction of the proposed method. The lacked imaginary components are coded separately.

(σ_{max}^2) and minimum (σ_{min}^2) values of σ_k^2 for k=1,...,N_{b0}.

Step 3: merge adjacent blocks k_1 and k_2 when their variances are matched by the following criterion:

$$\log\left(\frac{\sigma_{k_1}^2}{\sigma_{k_2}^2}\right) \leq \frac{1}{N_{b0}}\log\left(\frac{\sigma_{max}^2}{\sigma_{min}^2}\right) \qquad (11)$$

The resulting number of blocks will be N_b.The dimension of the final block k is $N_k x M_k$ with the property

$$N^2 = \sum_{k=0}^{N_b} N_k M_k \qquad (12).$$

3.2. Subband number allocation for each block

Clearly, the minimum number of subbands in each block is four.To allocate more subbands for blocks having a large activity (i.e. large frequency bandwidth) and large dimension (to reduce the computational cost) we use the following heuristic criterion:

$$B_{xk} = 2^{\text{Integer part of } \log_2\left(\frac{\sigma_k^2 N_k}{\sigma^2 N} B\right)}$$

$$B_{yk} = 2^{\text{Integer part of } \log_2\left(\frac{\sigma_k^2 M_k}{\sigma^2 N} B\right)} \qquad (13)$$

where

B_{xk} , B_{yk} are the numbers of subbands in the block k in horizontal and vertical directions, respectively.

σ^2 is the variance of the entire image.

B^2 is the number of subbands used when we process the entire image.

3.2. Subband bit allocation

In order to minimize the m.s.e. of the reconstructed image we distribute the total number of bits allocated to the coded image into the subbands of each block subjected to the following constraint:

$$S=\frac{1}{N^2}\sum_{k=1}^{N_b}\sum_{b=1}^{B_{xk}B_{yk}}\beta_{kb}\frac{N_kM_k}{B_{xk}B_{yk}} \qquad (14)$$

where β_{kb} is the bit rate (bits/pixel) in the subband b of the block k.

The variance $\sigma^2_{e_{kb}}$ of the quantization error of the subband kb is given by [3]

$$\sigma^2_{e_{kb}} = g_{kb}(\beta_{kb})\,\sigma^2_{kb}\,2^{-2\beta_{kb}} \qquad (15)$$

where

σ^2_{kb} is the variance of the subband kb and,

$g_{kb}(\beta_{kb})$ is a slowly varying function of the number of bits assigned to the subband kb.

In our evaluation, we consider that $g_{kb}(\beta_{kb})=g_{kb}$ is constant. Its value depends on the histogram of the subband.The variance of the quantization error of the entire image is

$$\sigma^2_e=\frac{1}{N^2}\sum_{k=1}^{N_b}N_kM_k\sigma^2_{e_k} \qquad (16)$$

where

$$\sigma^2_{e_k}=\frac{1}{B_{xk}B_{yk}}\sum_{b=1}^{B_{xk}B_{yk}}\sigma^2_{e_{kb}} \qquad (17)$$

given that the subbands of each block are orthogonal. Thus,we must minimize (16) subject to the constraint of (14). This problem has been solved for the entire image (Nb=1). However, the block partitioning needs a new solution.This standard problem is solved using Lagrange multipliers.The m.s.e. optimal bit assignment then is

$$\beta_{kb}= S+\frac{1}{2}\log_2\left(\frac{g_{kb}\sigma^2_{kb}}{\bar{\sigma}}\right) \qquad (18)$$

where

$$\bar{\sigma}=\left\{\prod_{i=1}^{N_b}\prod_{b=1}^{B_{xi}B_{yi}}\left(g_{ib}\sigma^2_{ib}\frac{N_iM_i}{B_{xi}B_{yi}}\right)\right\}^{\frac{1}{N^2}} \qquad (19)$$

By incorporating (18) into (15), the variance $\sigma^2_{e_{kb}}$ becomes

$$\sigma^2_{e_{kb}}=\bar{\sigma}\,2^{-2S} \qquad (20)$$

Incorporating (20) into (16) and (17) results in

$$\sigma^2_e=\sigma^2_{e_k}=\sigma^2_{e_{kb}}=\bar{\sigma}\,2^{-2S} \qquad \text{for each k and b}$$

The quantization error is uniformly distributed in the subbands and blocks.

The evaluation of g_{kb} from the histogram of the subband kb results in excessive computational load. So, we use the following procedure:

Step 1: We use initial values of g_{kb} based on the assumption that, generally, the low-frequency band has a Gaussian law, and the high-frequency bands have a Laplacian law. Hence, we compute the bit rates β_{kb} and we quantize the subband kb. From the resulting error $\sigma^2_{e_{kb}}$,we compute g_{kb} using (15).

Step 2: We achieve final coding using the previously computed g_{kb}.

4. IMPLEMENTATION

4.1. Technical constraints

The transputer is designed to be used in an array of devices to run Occam programs in parallel. Four interprocessor bidirectionnal serial links or channels are available on the chip. This array is controled by a master transputer which communicate with the host computer. Only two links of the master can be used to communicate with the array.

The master has large external RAM memory whereas the others have a smaller one to reduces the cost of the machine.

4.2. The parallelization of the algorithm tasks

The block-partitioning of the image can be done by a square transputer array where each transputer processes the square block of pixels corresponding to its physical position. Then, the merging of blocks stored into neighbouring transputers must be done by the master. The subband decomposition or reconstruc-tion of blocks and the subband variances computation can be acheived in parallel using an array of processing elements consisting of four connected transputers. The bit allocation in each subband cannot be done in parallel because all the subband variances are needed to compute $\bar{\sigma}$. The last step that can be parallelized is the quantization task.

Taking into account the technical constraints and the above discussion, we propose the architecture of Fig.2, where we assume that the entire image is stored in the external RAM memory of the master1.This architecture is composed by four processing elements (as defined above) full connected in a square array of level 1. We consider a processing element as a transputer array of level 2.

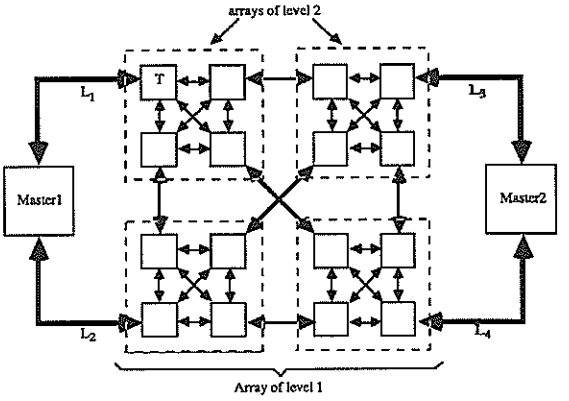


Fig. 2- The proposed architecture.

4.3. Computation steps and channels data transfer for subband coding image decomposition

A. Loading the image into the transputer array using links L_1 and L_2. Each transputer receive the block of pixels correspon-ding to its physical position.This is a loading into arrays of level 2.The loading scheme is described in Fig.3.

B. Computing by each transputer the m_k's and σ_k^2's corres-ponding to its own constant size blocks and then merging them.

C. Transfering to the master1 the m_k's and σ_k^2's of merged blocks as well as their coordinates XL_k, YL_k ; XR_k, YR_k of the top left and the bottom right corners, respectively.

D. Merging by the master1,the border blocks of adjacent transputers, updating their m_k and σ_k^2, centring the resulting blocks and computing their subband numbers Bx and By.Then, each block k is caracterized in the master1 by the vector C_k

$$C_k = (m_k , \sigma_k^2, Bx_k , By_k, XL_k, YL_k , XR_k, YR_k)$$

E. Loading through links L_1 and L_2 resulting blocks into array of level 1, where the four connected transputers have to process one block simultaneously.

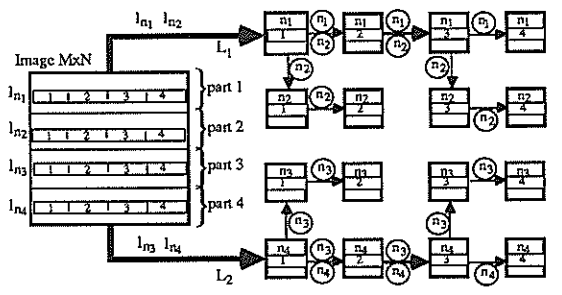
F. The subband decomposition of a block is done by one processing element according to the configuration described in Fig. 4. After decomposition, the transputers of the processing element compute individually the resulting subband variances and transmit them to the master1. The decomposed blocks were keeped in the processing elements memories.

G. Computing all subband bit allocation tables by the master1.

H. Loading bit allocation tables of decomposed blocks into array of level 1. The processing element acheives the quantization and coding task for all subbands of its own blocks, and transmits the final results to the master2 through links L_3 and L_4.

4.4. Computation steps and channels data transfer for image reconstruction

Only three steps are needed to the reconstruction process:



- The image is split into four horizontal parts (part1 to part4).
- The lines of part1 and part2 are transmitted via L_1 whereas those of part3 and part4 are transmitted via L_2.
- l_{ni} is the line n of part i for n=1,..., N/4. Each quarter of pixels of line l_{ni} is loaded into the transputer array according to the number written in each transputer boxes.

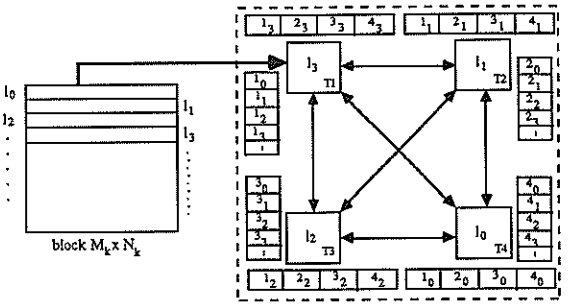
(n_i) denotes the transfer of some quarters of pixels of line l_{ni}.

Fig. 3- The loading configuration of the entire image into level 2 array.

A'. Loading from master2 through L_3 and L_4 , the coded blocks into level one array.

B'.These blocks are decoded and reconstructed by the pro-cessing elements using a similar procedure to decomposition (F.)

C'. Then, the reconstructed blocks are transmitted to master1 where the reconstructed image will be stored.



: Horizontal boxes illustrate that the line l_n after decomposition into Bx_k subbands by the transputer T_i is divided into four parts m_n (m=1,2,3,4). $T_i$keeps the part i_n and transmits the others m_n to T_m's.

: Vertical boxes illustrate that after the above exchanges, each transputer T_i stores the ith part of all processed lines. When the all $M_k$lines are processed, each transputer decomposes his $N_k/4$ columns into By_k subbands. The decomposition procedure of lines and columns is identical.

Fig.4- illustration of the block subband decomposition task into one processing element.

4.5. Hardware requirement

Clearly, the number of transputers must be multiple of 4.The level one array is made by pxq processing elements, then the image dimension (MxN) and the minimal block size (N_0xN_0) must verify the following equations:

$$\frac{M}{2q} = KN_0 \quad \text{and} \quad \frac{N}{2p} = LN_0 \qquad (21)$$

where K and L are integers. As M,N and N_0 are power of 2 then p,q,K and L must be also power of 2.

The local memory for image data of each transputer must be greater than $\frac{MN}{4pq}$. For example, if M=N=1024 and p=q=2 then the required local memory is 64 Kbytes and the memory capacity of the masters must be 1 Mbytes.

5. CONCLUSIONS

We have proposed a parallel implementation of an error-free decomposition/reconstruction method of subband coding images by blocks. We have show that this method can be massively parallelized on a transputer array and then permits a fast medicine images archiving and retrieving.

REFERENCES:

[1] Diab, Ch., Prost, R. and Goutte, R., Error-Free Image Decomposition/Reconstruction For Subband Coding Schemes, in Signal Processing: Image Communication, Vol. 2, N° 1, April 1990.

[2] Diab, Ch., Prost, R. and Goutte, R., Block-Adaptive Subband Coding Of Images, in Proc.IEEE, Inter. Conf. of Acoust., Speech, Signal Processing: ICASSP-90, April 3-6, 1990, Albuquerque, USA.

[3] Berger, T., Rate Distortion Theory (Englewood Cliffs, NJ:Prentice-Hall, 1971).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

MULTI RESOLUTION IMAGE CODING : A SOLUTION TO COMPATIBLE CODING

Michel PECOT, Philippe TOURTIER, Yann THOMAS

THOMSON CSF / LER, Ave Belle Fontaine, 35510 Cesson-Sevigne, FRANCE

In this paper an efficient multi resolution coding scheme based on subband techniques is presented. It allows compatible coding between TV and High Definition TV applications. Introduction of inter image coding must however be dealt carefully to preserve compatibility capabilities.

1. INTRODUCTION

Introduction of digital high definition television (HDTV) and improvement of network transmission capabilities make image data compression a more and more important research area. This topic has been studied for several years and some techniques such as predictive coding and transform coding (mainly Discrete Cosine Transform) are now well established. However, as lower and lower bit rates are looked for, their intrinsic drawbacks are becoming more and more critical.
Besides, they are not very well suited to compatible coding, even though already used in this context. Compatibility between the various services which will reach customers (HDTV, TV, Videotelephone applications) is clearly requested [1]. These applications correspond to a hierarchy of video scanning formats. Thus the compatibility issue is mainly concerned with the design of a multi resolution coding system: reconstruction of images with less resolution than the original should be possible. Naturally, this operation must be done without decoding the full transmitted bit stream. As an example, a TV set should be able to decode and display in its own format a HDTV broadcast. Upward compatibility which is concerned with the opposite problem, i.e. the ability of a receiver to decode and display a signal of lower resolution than its own working format, is another issue of interest in network applications. However, using the same algorithm and the same data organization for each picture format directly answers this problem.
A rather new technique in the image field, called subband coding, is able to answer this point since it allows a multi resolution analysis of input images. Thus, information available in subbands is well suited for partial reconstruction of the input signal. Besides this technique generalizes the concept of transform to the overlapping window case while avoiding pixel based processing of predictive methods and may thus yield better visual quality at low bit rates.
Up to now, subband techniques have been used mainly as a preprocessing prior to DPCM, DCT or even Vector Quantization in overall coding schemes [2], [3], requiring thus a small number of subbands. Our own interest here lies in developing a stand alone coding method based on this technique by use of the simplest coding for each band, i.e. PCM coding.

2. THE FULL SUBBAND APPROACH

The basic idea of subband splitting is to concentrate the signal energy in a few subbands. Coding can then take advantage of this property by allocating most of the bit rate to high energy bands. Besides, to make simple PCM encoding of each band close to optimal, subband signals as white as possible (i.e. with spectrum as flat as possible) and independent from one another must be obtained. Band splitting and decomposition filters that will lead to such an optimal scheme have thus to be found.
It is well known [4] that any analysis-synthesis system may introduce three types of distortion: interband aliasing (due to maximal subsampling), phase and amplitude distortions. However interband aliasing can always be totally removed which makes the overall linear system shift invariant. Perfect reconstruction of the input signal is then stated in terms of constraints, unfortunately not always realizable, on the analysis filter bank. Besides, decomposition of the two-dimensional image spectrum may be done in several ways according to the desired shape of the subbands (rectangular, quincunxial ...), the structure of the filter bank (parallel or hierarchical) and the decomposition filters themselves (finite or infinite impulse response filters, separable or non separable filters).

2.1. Filter bank structure

The hierarchical structure is naturally better suited in order to find out the 'best' decomposition since it allows to choose any spectrum splitting based on a unique elementary cell. Besides, this approach is the right framework to fulfil compatible coding constraints.
The elementary cell splits its input signal into a few basic subbands, typically 2 or 4. Such cells are then arranged in a tree structure and some parts of the spectrum may be more or less split according to their informational content. Such a modularity does not hold in a parallel system unless different subsampling order are chosen for each channel, which would lead to quite an impossible derivation of corresponding perfect reconstruction filters. Furthermore, the simplest way of designing this basic cell is to consider a separable approach: rows and columns are treated separatly and only monodimensional filters are used. The input image is thus split in 4 rectangular bands as sketched in figure 1. This figure shows off the corresponding lower resolution image (band 1) of interest when looking for compatible coding.
As a consequence, the design of the overall analysis-synthesis system is reduced to that of a two-band monodimensional cell for which general classes of solutions exist.

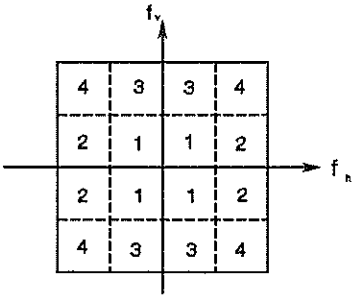


Figure 1: separable 4-band spectrum splitting

2.2. Basic cell filters

This monodimensional elementary cell is described in figure 2. As well known, perfect reconstruction of the original signal is obtained with the following choices:

$$G0(z) = 2.H1(-z)/D(z)$$

$$G1(z) = -2.H0(-z)/D(z) \quad (1)$$

where

$$D(z) = [H0(z)H1(-z)-H0(-z)H1(z)] \quad (2)$$

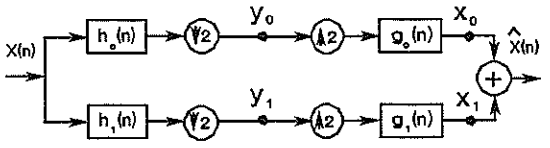


Figure 2 : Two-channel 1D basic cell

However, in such a scheme, use of Finite Impulse Response (FIR) filters is preferable in order to avoid numerical problems with finite precision arithmetic. Moreover, FIR filters may have linear phase which is generally preferred in image processing, especially when filtered images are of interest for compatibility purposes. Reconstruction filters G0 and G1 have then to be chosen as:

$$G0(z) = 2.H1(-z)$$

$$G1(z) = -2.H0(-z) \quad (3)$$

Both conditions cancel inter-band aliasing and perfect reconstruction property is then stated in terms of D(z) as:

$$D(z) = z^{L} \quad (4)$$

Each pair of solutions (H0,H1) is thus deduced from the design of a product filter P(z) = H0(z).H1(-z) and its subsequent factorization. However, extra constraints, due to coding purpose, reduce the number of admissible solutions.
The first constraint is closely related to the quantization of subband signals y0 and y1. Assuming that both signal are independently quantized, it is desired that quantization error variances add together independently at the reconstruction stage. It can be shown [5] that this condition is equivalent to the orthogonality of signals x0 and x1 and may be stated in terms of the intercorrelation function of analysis filters h0 and h1 as follows:

$$Rh01(2k) = 0 \text{ for all } k$$

Furthermore, it is shown that the only FIR filters satisfying both perfect reconstruction and orthogonality conditions are conjugate quadrature filters (CQF). In other words, both conditions are equivalent to:

$$h1(n) = -(-1)^{K-n}.h0(K-n) \quad (5)$$

for some odd integer K, and

$$|H0(f)|^2 + |H0(f+1/2)|^2 = 1 \quad (6)$$

Such filters, which have necessarily an even length, have been extensively studied in [6] and may be designed with almost any given shape. Their major drawback is that they cannot have linear phase as soon as

their length is greater than 2. Nevertheless, clever pairing of the zeros when factorizing the autocorrelation function Rh0(z) may lead quasi-linear phase provided that long enough filters are used. A second requirement that is needed for the basic cell is the statistical decorrelation of subband signals y0 and y1. This will in fact allow to encode each band independently from one another with close to optimal performance. It is shown in [5] that if h0 and h1 satisfy the CQF relation (5), have linear phase and even length, then pointwise decorrelation of subband signals is obtained, that is to say:

$$E[y0(n)y1(n)] = 0 \text{ for all } n.$$

This property allows to obtain decorrelated quantization errors whatever quantization coarseness. Remaining inter-band correlation is furthermore determined by the spectral overlapping of filters h0 and h1. Such linear phase filters are well known as Quadrature Mirror Filters (QMF) and cannot yield exact reconstruction of the input signal, except for the trivial 2-tap filters. Some optimization procedure has thus to be used when designing them in order for (6) to be approximatively satisfied. Now, a reconstruction signal to noise ratio greater than 50 dB is high enough for coding purpose; unfortunately, 12-tap or more filters are needed to reach such performances.
Two families of 'admissible' filters are thus availabable. Choice of a 'good' filter relies on a compromise between the approximate decorrelation property of CQF's and the approximate reconstruction property of QMF's. This compromise essentially affects hardware complexity of resulting filters. Besides, visual quality of low-pass subsampled images (compatible constraint) will take a prominent part in this choice.

2.3. Decomposition tree

As seen above, inter-band decorrelation essentially relies on the filter choice. Intra-band decorrelation depends on the decomposition tree. It consists of a hierarchical arrangement of several two-dimensional four-band basic cells. From a theoretical point of view, it is clear that intra-band correlation decreases as the number of splitting increases, i.e. as the decomposition tree goes deeper and deeper. In fact, further splitting of a subband and then PCM encoding of the four resulting bands always yields a theoretical coding gain greater than one - compared with direct PCM encoding of the original band - except if that subband has already a flat spectrum. Two types of subband images have however to be distinguished: the low frequency band, which has always been low-pass filtered in both directions, and the other ones. Some correlation will always remain in the former, due to its visual content, even if highly split. Nevertheless, this subimage becomes smaller and smaller and thus rapidly accounts for a negligible share of the total cost of the original image. Besides, the more this band is split, the longer its corresponding reconstruction filter, which may cause a larger spreading of quantization noise in the reconstructed image. Practically, 3 or 4 decompositions of this band yield quite an optimal compromise. Considering high-pass filtered subimages, it is clear that correlation quickly decreases; splitting of such bands must however be limited to reduce quantization noise propagation at the reconstruction stage, all the more so as it occurs near contours. Simulations have shown that the best compromise between intra-band decorrelation and quantization noise propagation is given by the 'optimal' tree sketched in figure 3.

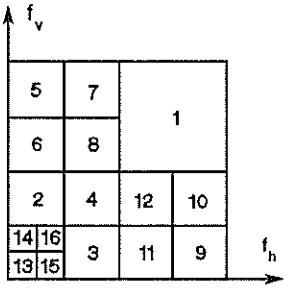


Figure 3 : 'Optimal' 16 band decomposition tree

2.4. Subband encoding

After band splitting each subband is independently PCM encoded. Quantization of each band could be done using optimal Llyod-Max quantizers followed by fixed length encoding, since statistics of high frequency bands are quite stationary over input images. Now, it is well known that linear quantization followed by entropy coding performs generally better and is easier to design. This latter solution has thus been preferred.
Let qi the quantization step of subband i. Optimal qi values have thus to be found which minimize the reconstruction error variance subject to a given average bit rate D. The orthogonality relation (5) ensures that the reconstruction error variance equals the sum of subband quantization error variances Vi; since linear quantization is used and assuming that small enough quantization steps are used, Vi may be expressed as:

$$Vi = qi^2/12 \quad (7)$$

Furthermore, the total bit rate may be written as the sum over all the bands of intermediate rates Di:

$$Di = 4^{-Ki}.Hi \quad (8)$$

where Ki represents the number of time band i has undergone a 4-band split and Hi is the entropy of the corresponding quantized band:

$$Hi = h(Xi) - \log(qi) \quad (9)$$

h(Xi) being the differential entropy of band i before quantization. (9) also assumes small enough quantization steps. It is then easily shown that optimal quantization steps satisfy:

$$qi = Q.4^{-Ki} \quad (10)$$

Thus any further 4-band split divides quantization step by 2. Once quantized, each band is encoded using variable length codes (VLC's) and run length descriptions. Efficient VLC's may be easily designed for high frequency bands, since they are generally symmetrically exponential-like distributed whatever the input image. Concerning the low frequency band, fixed length codes may perform better since its statistics are highly dependent on the input distribution.
The intra coding scheme described above is however not efficient enough to meet CCIR quality requirements for high definition television images, at bit rates around 0.8 bit per pel. Temporal correlation has to be taken into account in an inter image scheme.

3. COMPATIBLE INTER SCHEME

The proposed hybrid scheme is based on motion estimation and compensation. In order to fulfil compatibility constraints and avoid drift problems between the coder and associated compatible decoders, motion compensation must be done at the subband level. Besides, motion estimation is performed using a block matching algorithm and may be done either at the image level or at the subband level without loss of compatibility capabilities. In the former case, motion vectors have to be scaled down according to the decimation factor for use in each band. Sub-pel motion compensation has then to be used in bands if the downscaling process is performed without loss of accuracy and some interpolation is thus needed. However, since each subband suffers from aliasing, no interpolation filter can recover the image information that was used for motion estimation, which limits the efficiency of the compensation process. To avoid such problems, motion estimation should be done at the subband level. A hierarchical estimation algorithm is then well suited to reduce the overhead of motion vectors. In a first step, motion vectors are estimated in the base-band (band 13); these vectors are used for motion compensation in bands belonging to the same resolution level (bands 13 to 16); furthermore, they serve as initialization vectors for the estimation at the next resolution level which is done on the low-pass image reconstructed from bands 13 to 16. This process is then iterated until the image level. Such an approach naturally leads to a differential coding of resulting vectors.

4. RESULTS

An efficient compatible coding algorithm has been presented. Compatibility is met by using the same subband splitting for input images whatever their resolution, and decoding only useful part of the spectrum at the decoder side. This scheme gives good image quality at 0.8 bit per pel for the original image as well as for its compatible parts and may thus be used for HDTV transmission on a 140 Mbit/s channel.

REFERENCES

[1] Y.M Le Pannerer, Compatible solutions for TV and HDTV transmission, International TV Symposium, Montreux France, 1989

[2] P.H Westerink, D.E Boekee, J Biemond and J.W Woods, Subband coding of images using vector quantization,IEEE Trans. Comm., Vol 36, No 6, June 1988

[3] J.W Woods and S.D O'Neil, Subband coding of images, IEEE Trans. ASSP, Vol 34, No 5, Oct 1986

[4] M Vetterli, Analysis, synthesis and computational complexity of digital filter banks, Ph.D dissertation No 617, Ecole Polytechnique Federale de Lausanne, Switzerland, April 1986

[5] M Pecot, P Tourtier and Y Thomas, Compatible coding of television images, to be published in Image Communication, Special Issue on HDTV

[6] M.J Smith and T.P Barnwell, Exact reconstruction techniques for tree-structured subband coders, IEEE Trans. ASSP, Vol 34, No 3, June 1986

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Transmission of images over a bursty and random channels

K.FAZEL. J.J.LHUILLIER
Laboratoires d'Electronique PHILIPS
3 Av Descartes.Limeil-Brévannes. FRANCE

ABSTARCT : This paper investigates a combined source-channel encoding for random and bursty channels where the source encoder is based on the DCT and VLC technique. The channel encoder is principaly based on the uneqaul-error-protection principle. To measure the importance of each bit in the transmetted frame after VLC, a factor of sensitivity for each bit to channel error is defined. Then, by utilizing this factor, the optimal error rate for each bit that minimaizes the effects of channel noise, is estimated. The solution discussed here is firstly, to reduce the BER after correction by exploiting the characteristics of the channel to some moderate values, and secondly, to adjust the BER of the transmetted frame bits, according to their importances.

1. Introduction

To reduce the redundancy of digital TV images, different coding algorithms have been studied. The performance of any algorithm is characterized by its compression factor and by the SNR at the decoder side. The higher the compression factor at the source, the greater is the significance of the transmitted information, and hence the greater the effects of channels errors.

As a result, for noisy channels applications, it is necessary to correct the channel errors. In the field of error control, many coding techniques have been studied. These techniques consist of coding the images without taking into account their characteristics, or on the other hand combining the source and channel coding by utilizing the properties of the images. This last technique permits selective protection of the most sensitive part of the information against channel errors. The efforts carried out by Modestino et al [1], Gomstock and Gibson [2] and others showed the interest of using combined source-channel coding of images. Generally they use orthogonal transform with a fixed number of bits allocated to the transformed block, as a source encoder. Then by estimating the most important bits of the transformed blocks, they evaluate the performance of some block and convolutional codes.

The recent application of variable-length-coding (VLC) in source encoders, improves noticeably the performance of this last technique. These codes minimize the output bit rate without loss of information. Hence, the transmitted information become more vulnerable to the channel errors.

In the case where the source encoder is based on the Discrete-Cosine-Transform (DCT), followed by a VLC, a combined source-channel coding scheme has been studied [3] . In this study the channel encoder is based on the unequal-error-protection principle. In order to measure the importance of each bit in the transmitted frame after VLC, a factor of sensitivity for each bit to channel error is defined. Then by utilizing this factor, the optimal error rate for each bit in the frame, that minimizes the effect of channel errors, is estimated. This study [3] shows that the design of an UEP code constructed by the Blokh-Zyablov (UEP-BZ) code construction method with 11-10% of redundancy requires some complex hardware.

The purpose of this paper is to introduce an alternative solution which is to concatenate two simple codes where one is based on the equal-error-protection (E.E.P) principle, and the other one is based on the unequal-error-protection (U.E.P) principle combined with source-encoder. The inner-code (EEP), by exploiting the characteristics of the channel errors reduces the bit error rate to some moderate values while the outer code (UEP) adjusts the bit error rate required for the transmitted frame bits, according to their importance. These two concatenated codes provide 10% of redundancy and simulations showed that the best choice (trade-off complexity performances) is to devote 4% of redundancy for the inner code and 6.25% for the outer code.

This paper is organized as follows : in section 2, after a brief review of the source-encoder employed, a factor of sensitivity for each bit is derived. Then by utilizing this factor, the optimal error rates required for each bit in the transmitted frame is estimated (section 2). Section 3 is devoted to the performance evaluations of the inner codes (for random and bursty channels) and of the outer Blokh-Zyablov UEP code. Two sequences of images transmitted in a bursty and random channels are considered, and the complexity of the system is discussed.

2. Source coding algorithm, bit error sensitivity, required bit error rates

a. Source coding algorithm, influence of channel errors

One of the most attractive source coding scheme possessing a high performance is based on the Discrete-Cosine-Transform (DCT) [4] . Let's briefly look at the more important functions of this scheme :

- The picture is divided into small blocks of (NxN) pixels, and each block is submitted to DCT.
- These transformed samples are quantized. The quantization step varies according to the activity of the block and to the frequency in order to limit the visibility of the coding noise.
- Finally the variable-length-coding (VLC) is performed to the quantized values (except DC components) of several adjacent blocks, multiplexed on a coefficient basis. This VLC minimizes the output bit rate without loss of information. In addition, performing this coding on several adjacent blocks permits to take benefit of the spatial luminosity correlation existing in a picture, in order to reduce the number of bits devoted to the DC component of each block.

As this process generates a variable output bit rate, a buffer memory is required, combined with a regulation mechanism, reacting on the quantizer (adapting the quantization step) to match the fixed bit rate, necessary for the channel.

The variable length coded frames by combining four adjacent blocks have the following form :

SW	DC	AC	V.L.

where DC represents the continue component of the firt block followed by 4x2 bits the activities (AC) of 4 blocks (fixed length part), and completed by the variable length part of the quadblock which represents the difference between the remaining DC

component and the 1st DC component and the remaining other coefficients, previously scanned in a zig-zag manner , multiplexed and coded by a variable length coding. Since the number of non zero coefficient of a quadblock is variable, every frame is ended by special pattern (end of block : EOB), that belongs to the VLC table.

The occurrence of errors in these frames causes the following events :

- If the errors occur in the fixed length part of the frame, the visibility of the corresponding defects is proportional to the significance of the corrupted information (an error in the most significant bits of the DC component would dramatically modify the whole block).
- And if an error occur in the variable length part of the stream, a correct segmentation is no longer guaranteed. It may cause the apparition of false patterns or even a spatial shift of images. The apparition of false patterns depends strongly on the position of errors in the variable-length-part of the frame [3].

Assuming a good synchronization, the importance of each bit in the transmitted frame can be quantized by its sensitivity to channel errors.

b. Error sensitivity of each bit

In order to protect each bit of the bit stream proportionally to its influence on the SNR when corrupted by an error, a factor of sensitivity for each bit is defined [5].

Let denote P_N the noise power after source decoding in the presence of channel errors. Then P_N is given by :

$$P_N = \sum_{i=1}^{M} \sum_{k=1}^{4N^2} (x^i(k) - \hat{x}^i(k)_e)^2$$

where M and N^2 are respectively the number of quadblock (4 adjacent blocks) in the image and the number of pixels in each block. And $x^i(k)$ represents the k^{th} pixel amplitude of the i^{th} block before encoding and $\hat{x}^i(k)_e$ the k^{th} pixel amplitude of the i^{th} block after decoding, in the presence of channel errors . If there is no channel errors $\hat{x}^i(k)_e = \hat{x}^i(k)$ and $(x^i(k) - \hat{x}^i(k))$ represents the source coding noise.

Let li be the length of the frame i, then the number of error patterns will be 2^{li}. Assuming independence of error patterns across the frame, P_N can be expressed as : $P_N = P_{Nq} + P_{Nc}$; where

$$P_{Nq} = \sum_{i=1}^{M} \sum_{k=1}^{4N^2} \varepsilon_q^2(k,i) \qquad \text{which is the source coding noise and :}$$

$$P_{Nc} = \sum_{i=1}^{M} \sum_{e=1}^{2^{li}} Pe A_e^i; \qquad A_e^i = \sum_{k=1}^{4N^2} \varepsilon_c^2(k,i) + 2\varepsilon_c(k,i) \times \varepsilon_q(k,i)$$

where : $\varepsilon_q(k,i) = x^i(x) - \hat{x}^i(k)$
$\varepsilon_c(k,i) = \hat{x}^i(k) - \hat{x}^i(k)_e$

Pe denotes the probability of the e^{th} error pattern,

P_{NC} depends on the channel errors, and it can be considered as the sum of channel noise and the interaction of source coding-channel noises.

By assuming that we can neglect the effect of error pattern of weight $e > 1$ after correction, then P_{Nc} can be written as :

$$P_{NC} = \sum_{i=1}^{M} \sum_{j=1}^{l_i} Pj A_j^i \simeq \sum_{j=1}^{lmax} Pj \sum_{i=1}^{M} A_j^i = \sum_{j=1}^{lmax} Pj \overline{Aj}$$

where Pj is the probability of error in the j^{th}bit (Pj doesn't depend on the frame) and Aj^i represents the noise power due to an error on bit j of block i. And lmax is the maximal frame length value and $\overline{Aj}$ the average contribution of an error on bit j on any block.

From the previous formula, the sensitivity of bit j can be derived :

$$(SNR)_j = 10 \log \left(\frac{P_{SM}}{P_{Nq} + \overline{Aj}} \right) ; \quad j = 1,...lmax$$

where P_{SM} is defined as the theoretical maximal power of the image. So $(SNR)_j$ is obtained by systematically putting an error at the j^{th} bit position of all frames. A curve of sensitivity has been plotted for "Girl" TV image sequence in fig.2

c. Optimal bit error rate required

As was discussed before, the SNR of an image in the presence of errors depends strongly on the bit error rate for each bit. To determine the optimal BER for each bit that achieves a high SNR, it is desirable to fulfill the following condition : $P_{Nc} = K << P_{Nq}$. Then the set of optimal Pj provides that all contributions to noise power are equal [5]. It yields :

$$P_{Nc} = lmax\ Pj\ \overline{Aj} = K, \quad j = 1,...lmax$$

or $Pj = \frac{k}{\overline{Aj}}$; $\quad k = \frac{K}{lmax}$ where K is a predefined constant

So to respect the above condition, we have to equalize the curve of sensitivity factor. To this end, an unequal-error-protection code is desirable:

3. Application of a concatenated UEP code with an EEP code to the image transmission

In [3] it has been showed that the application of an UEP code with three levels of protection has a high performance. However the complex hardware of such a system leads us to find an alternative solution for this problem. The transmission channel considered in [3] is a memoryless Binary-Symmetric-Channel (BSC). Hence for more realistic channels i.e. bursty channels, a "Bit-Interleaver" in order to disperse the gathered errors, is required. So the system described in [3] doesn't exploit the presence of memory in real channels.

The solution discussed here is to concatenate two simple codes. This scheme by providing a reasonable hardware complexity, permets to take benefit of the presence of memory in the channel where its block-diagram is given in fig. 1. The inner code, adapted to the channel errors characteristics is based on the equal-error-protection principle, while the outer code, combined with the source encoder is based on the unequal-error-protection principle. The inner code reduces the bit error rate to some moderate values. Then, the bit-error rate required for each bit in the VLC frame, is adjusted by the outer code.

The UEP code is applied to each VLC frame independently. Hence the redundancy added by this code is variable i.e. the data rate after the coding process is not fixed. The solution adopted here is the same as the one proposed in [3]. This consists of incorporate a single regulation system for source and channel encoders. The feedback controls the source quantizer step in order to obtain a constant data rate for the inner code.

It should be noted that the other solution proposed in [3] assumes a good synchronization between the VLC frames, is also adopted here. This is to transmit the VLC frame length l by controlling it using a powerful code.

After a brief recall of the definition of EEP and UEP codes by the "separation vector" of a linear code, we will examine the performance of these two codes separately.

Definition[3] : for a linear code C(n,k) over the alphabet GF(q) the separation vector S(G) = $(S(G)_1,...S(G)_k)$ with respect to a

generator matrix (or generator polynomial) of C is defined by :

$$S(G)_i = \min \{wt(mG) \mid m \in GF(q)^k; mi \neq 0\}$$

where Wt(.) denotes the Hamming weight function, and m the message word.

From the above definition the UEP and EEP codes are defined as :

- If $S(G)_i = S(G)_j$, $\forall i,j = 1,...k$. Then the code is an equal-error-protection i.e. the code provides the same protection level for all the message component m_i.
- Otherwise ($S(G)_i \neq S(G)_j$, $\forall$ i,j = 1,...k). The code is considered as an unequal-error-protection.

The error correction capability of these codes in a q-ary symmetric channel is given :

Theorem [3] : a linear code C over GF(q), with generator matrix G and complete nearest-neighbour-decoding guarantees the correct interpretation of the i^{th} message digit, whenever the error pattern weight is less or equal to :

$$\left[\frac{S(G)i - 1}{2} \right]$$

a. **Performance of the inner code (EEP codes)**

The choice of the inner EEP code depends on the transmission environment:

1) **Random channel**

For a BSC, a binary cyclic C(n,k) code (such as BCH) with t bits error correction capability, is the best choice. Let consider p the probability of channel errors, and assume that the decoder doesn't add more than t errors, then the BER after correction is bounded by :

$$BER \leq \frac{1}{n} \sum_{i=t+1}^{n} (i+t)\, p^i (1-p)^{n-i}$$

2) **Bursty channels**

Bursty channels are often modeled by a Markov chain [6] with M states, split into M_1 good states (error free states) and $M_2 = M\text{-}M_1$ bad states (erroneous states) with defined transition probabilities matrix .Generally the transition probability matrix coefficients are the transition probabilities between bits, so far a non-binary code (such as RS codes) ; the model has to be extended [6]. The extended model consists of q + 1 states (q is the number of bit making up the non-binary code symbol's), including one good state and q bad states. The state i (i = 0,1...q) represents a symbol with i errors. Then the probabilities between symbol states are derived from the binary model parameters [6] Then the bit error rate after decoding for a RS(N,k,t) code is bounded by :

$$BER \leq \sum_{m=t+1}^{N} \sum_{w=m}^{mq} \frac{(w+qt)}{Nq} P(m,w,N)$$

where t is the number of correctable symbol errors and P(m,w,N) denotes the probability that m error symbols and w error bits occur in N symbols . This last terme is derived from the model parameters [6]

b. **Performance of the outer code (UEP code)**

This code being suitable for source encoder [3] imposes that errors in the input of decoder have to be independent. Hence a bit interleaver is required after the inner decoder to disperse the errors. This code is constructed by the Blokh-Zyablov [7] code construction method :

1) **Blokh-Zyablov UEP code construction**

Let Bi be a (n,k_i, d_i) linear code over $GF(2^{ai})$ with generator matrix G_i and minimum distance d_i, for $i = 1...r$. Let G be a generator matrix of a C(N,K) code with $K = \sum_i a_i$ and let $m = (m_1,...m_r)$; $m_i \in GF(2^{ai})^{ki}$, be the i^{th} component of the message m. Then the BZ coding process is as follows :

- the m_i ; i = 1,...r is coded by the code B_i : $C_i = m_i G_i$.
- Then C_i is arranged in the i^{th} block (a_i lines and n column) of Kn matrix M, where each column of the i_{th} block is the binary representation of symbols of C_i.
- Finally the complete matrix M is coded by G and gives the BZ codewords : $C = G^T M$.

The resulting code is a two dimension (N,n) code of length nN and rate $R = \frac{\sum a_i k_i}{nN}$.

Let e_i (i = 1,...r) denotes the minimum distance of the binary code generated by the last (r-i+1) blocks of G, then the UEP capability of the BZ is derived as follows :

Theorem [7] : the BZ code has a separation vector $(S_1,...S_r)$ for the message space $(m_1,...m_r)$, that satisfies the following inequalities :

$$S_i \geq e_i d_i \text{ ,if } d_1 e_1 \geq d_2 e_2 ... \geq d_r e_r$$

For our specific application i.e. to combine source-channel coding, since the length l_j of each frame is variable then it imposes the constraint to the UEP-BZ that of shortening the codewords (to adapte the length of the code to l_j) if $l_j <$ codeword length. Then UEP-BZ can be shortened if the matrix G has the following form :

$$G = \begin{bmatrix} I & 0 \\ 110....0 \\ ..... \\ 100...01 \end{bmatrix}$$

where $e_1 = 1$, $e_2 = ... = e_r = 2$ and I is the a_1 x a_1 identity matrix.

Let consider $P(l_j)$ the probability of occurrence of a frame of length l_j(this probalility is measured for two sequences of images) , and assume that the generator matrix G has the above from, then the BER after decoding for the protection level i is given by :

$$BER_i \leq \frac{1}{na_i} \sum_{l_j = l_{min}}^{l_{max}} P(l_j) \sum_{w > \frac{d_i e_i - 1}{2}}^{\min(l_j + r_j, nN)} (w + a_i \frac{(e_i d_i - 1)}{2}) P(w, \min(l_j + r_j, nN))$$

where rj are the redundancy bits of the UEP-BZ codes , and P(w, min(l_j+ rj,nN)) is the probability of w errors occuring in a block of min(l_i + rj,nN) bits and is given by :

$$P(w, \min(l_j + r_j, nN)) = C_w^{\min(l_j + r_j, nN)} (1-p)^{\min(l_j + r_j - w, nN)}$$

p denotes the bit error rate after the inner decoder. In the same manner :

$$BER_i \leq \frac{1}{na_i} \sum_{l_j > k_1 + ..k_{i-1}}^{l_{max}} P(l_j) \sum_{w > \frac{d_i e_i - 1}{2}}^{\min(l_j + r_j, nN)} (w + a_i(d_i - 1)) P(w, \min(l_j + r_j, nN)); \; i \geq 2$$

4. **Examples and simulation results**

Two sequences of images : "Girl" and "Baltimore" are coded (by taking a block of 8x8 pixels for the DCT) with the source encoder described in section 2 (the bit rate is fixed to 1 bit/pixel). The error sensitivity factor of each bit in the coded VLC frame is measured

for each sequence and is represented in fig.2. The optimal error rate for each bit P_j (assuming a good synchronization) by respecting the constraint in section 2 ($K=10^{-5}$) is estimated and reordered in an increasing manner (fig.3). These curves show that an unequal-error-protection code is desirable.

The frame lengths which are the most important information are assembled (by four length) and coded by a binary C(52,40) shortened cyclic code.

For a random channel a binary shortened BCH (200,192) code which corrects one error, as an inner code is taken. Whereas for the bursty channel considered here a shortened RS(100,96) code over $GF(2^7)$ is suggested.

The bursty channel model considered is the model of errors issued from a Viterbi decoder with a 4-state convolutional encoding combined with a 16-PSK modulation [8] . This model consists of 3 states : 2 error free states and one error state .

The UEP code, as an outer code is constructed by the Blokh-Zyablov code construction method. The Bi sub-codes used to construct this code are summerized in table.1

The inner and outer codes have a coding rate $R=0.9$. The performance of this concatenated code for an input $BER=10^{-4}$ is presented for a random and the bursty channels in Fig-4. These results are compared with a three levels UEP code described in [3].These curves show that the presence of a memory effect gives a higher performance with respect a random channel. However in all cases , this concatenated scheme has higher performance than a unique UEP-BZ code .

The complexity of the system depends on the complexity of the inner and outer decoders . The inner decoders (BCH or RS) has a reasonnable complexity (only one or two error correction). The complexity of the outer code is especially related to the complexity of the B1 code . The other decoders (Bi, i=2 ,3, 4) are only erasable-decoders . For the UEP code considered , B1 corrects only two errors , which has a very reasonnable complexity.

Hence the described system , by having a very high performance provides an acceptable hardware complexity

5. Conclusion

In this paper the problem of designing a combined source-channel encoding of images for random and bursty channels has been considered . The channel encoder is based on a concatenated scheme : the outer code is an UEP code and the inner code is an EEP code . The inner code , adapted to the channel characteristics reduces the bit error rate to some moderate values , while the outer code adjusts the required BER for the transmitted frames bit , according to their importance . The study showed that this concatenated scheme has higher performance and lesse complexity than a single UEP code .

6. References

[1] J.W. Modestino, D.G. Daut and A.L. Vickers - "Combined source-channel coding of images using the block cosines transform" - IEEE Trans. on Com., vol. 29, pp. 1261-1273, September 1981.

[2] D.R. Comstock and J.D. Gibson - "Hamming coding of DCT compressed images over noise channels" - IEEE Trans. on Com., vol. 32, pp. 856-861, July 1984.

[3] K.Fazel.and J.J. Lhuillier - "Application of unequal- error-protection codes on combined source channel coding of images" SUPERCOMM.90-Atlanta-Georgia .U.S.A

[4] W.H. Chen and W.K. Prat - "Scene adaptive coder" - IEEE Trans. on Com., vol. 32, n° 3, March 1984.

[5] J. Hagenauer, N. Seshadri and C.E.W. Sundberg - "Variable-rate sub-band speech coding and matched channel coding for mobile radio channels" - IEEE Vehicular Technology Conference, VTC-88, Philadelphia, Penn. June 1988.

[6] A.I. Drukarev - "Performance of Error Correcting Codes on Channels with Memory" - IEEE Transactions on Communic., vol. 34, pp. 513-521, June 1986.

[7] E.L. Blokh, V.V. Zyablov - "Coding of generalized cascade codes" - Problemy Peredachi Informatsii, vol. 10, n° 3, 1974.

[8] K.Fazel and Ph.Salembier- "Application of error modeling at the output of maximum likelihood decoder to concatenated coded 16-PSK".GLOBECOM 89 .Dallas.U.S.A

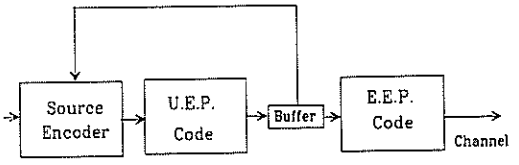


Fig1-A concatenated EEP code and UEP combined with the source encoder

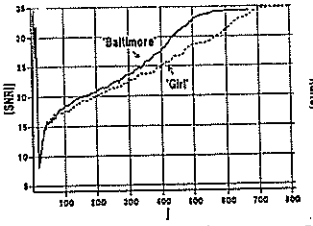


Fig2a Factor of bit error sensitivity

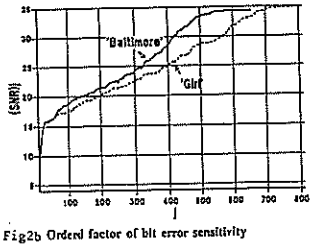


Fig2b Ordeal factor of bit error sensitivity

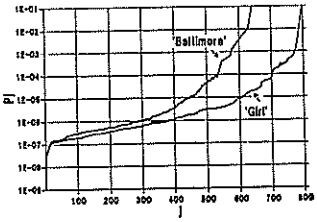


Fig3 Optimal error rate for each bit

Table.1 The U.E.P & U.E.P + E.E.P subcodes

U.E.P	U.E.P + E.E.P
The subcodes on U.E.P B1(63,45) over GF(2) B2(63,61) over GF(2⁶) B3(63,62) over GF(2⁶)	The subcodes of -E.E.P RS / BCH - U.E.P B1(127,113) over GF(2) B2(127,125) over GF(2) B3(127,125) over GF(2) B2(127,126) over GF(2)

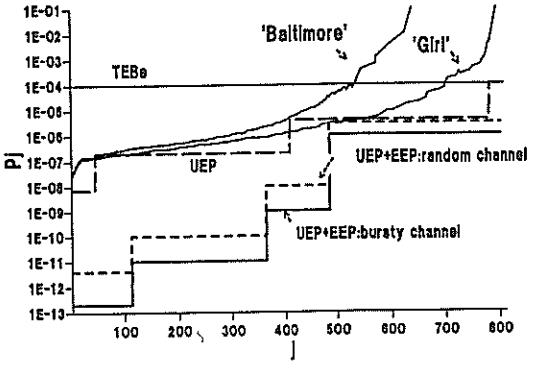


Fig4 The Performance of UEP + EEP codes

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

An Experiment on Buffer Occupancy Control in Video Coding for Several Bitrates*

Antonio Ortega, Narciso García, and Guillermo Cisneros

Grupo de Tratamiento de Imágenes, E.T.S. Ingenieros Telecomunicación
Universidad Politécnica de Madrid, E-28040 Madrid, Spain

A buffer control design method is described. This method is based on the statistical modelling of the video source, and the definition of cost functions, associated to several coder parameters. A practical cost function, which reflects the buffer occupancy state (A high cost is associated to overflow), is defined, and the buffer control strategy consists of choosing the minimal-cost coder working mode. The experimental results, obtained for a typical videophony sequence, prove that this is workable design method. Its main advantage over classical methods is the provision of a general design method, which can be applied to the buffer control of different coders, and can be used under a wide range of transmission bitrates.

1. Introduction

The perspective of service integration within the ISDN frame has led to wider studies of image services (image transmission, videotelephony, videoconferencing) and, therefore, has generated research in video transmision and coding.

When designing a video codec, two main factors have to be considered, which will be given a different emphasis, depending on the type of service the codec is intended for:

- Image quality (which should be kept constant, at a level consistent with the users' expectations, for a given service).
- Output bitrate (which should be kept as low as possible within a certain range, depending on the service).

In a packet video environment, the Asynchronous Transfer Mode (ATM) and the sharing of the channel among several sources (video or others) will result in the possibility of accessing to a variable transmission capacity [4], [5], [6], [7], [8]. Nevertheless, in the following work this option has not been considered, and a DCT-based variable bitrate coder connected to a Fixed Bit Rate (FBR) transmission channel has been used. In this last case, a buffer is needed to adapt the variable output of the coder to the fixed bitrate of the channel.

This paper proposes an approach to handle the aforementioned output buffer, minimizing the overflow probability, thus avoiding the quality loss. If the buffer becomes full, at least one frame could be removed, originating a decrease in subjective quality at the receiver end.

This approach is based on:

- Statistical knowledge of the video source that will produce an acceptable *model* for it.
- Definition of *cost functions* associated to the buffer occupancy level, the current quality, and the changes in quality level from the current status.

In order to simplify the model, the cost functions have been, in the experiments, associated only to the buffer occupancy level. A statistical model for the video source will be evaluated. This model provides a statistical knowledge of the video source behaviour for each coder working mode (each working mode is uniquely defined by a parameter called transmission factor, as defined in [1], [2], [3]). The most suitable choice for this transmission factor will the one that minimizes the cost. A general procedure for designing this buffer control strategy, independent from the channel bitrate and the output of the coder, will be presented. The results and the graphics shown were obtained from the application of this method to the sequence internationally known as "salesman" (Only the luminance was used).

2. The Source Model

2.1. Coder choice

For the work presented here, a Hybrid DCT-codec simulator loosely based on the CMTT-2 specifications

This work has been done within Eureka-256: "Bit-Rate Reduction System for HDTV Digital Transmission". It has been partially supported by the Plan Electrónico e Informático Nacional and the Comisión Interministerial de Ciencia y Tecnología of the Spanish Government.

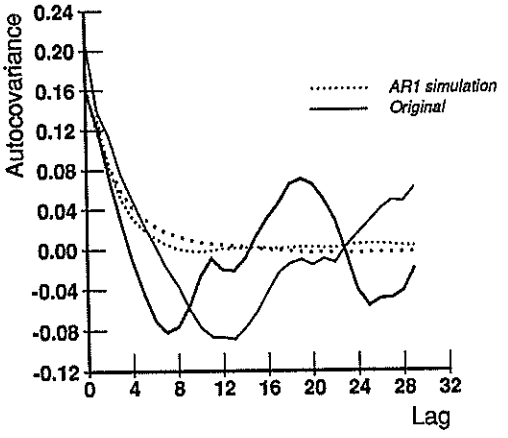


Figure 1: AR1 adjustment of source autocovariance: at 10 (Top) and 5 images/sec

has been written [1], [2], [3]. The coder is controlled by means of a variable parameter called transmission factor ([1], [2], [3]), which affects coding in the following way:

- A single quantizer is used for all the image blocks, and all the DCT coefficients within a block.
- A Variable Length Coding (VLC) is applied, so the quantized levels most often transmitted (the levels closest to zero) are coded with a smaller number of bits (Run-lengths of zeros are also coded).
- Each DCT coefficient is scaled by a certain, psychophysically computed, weighting factor before quantization. The scaling factor depends on the coefficient being quantized, so that higher order coefficients are more likely not to be transmitted (quantized to zero). As each coefficient is divided by the scaling factor, the latter represents the transmission threshold.
- Additionaly, for each frame or image, a transmission factor is calculated. A higher value for the transmission factor means that the scaling factor, and therefore the transmission threshold, is bigger.

In this work, the transmission factor has been considered as the coder' s sole variable parameter (It is assumed that this does not imply a loss of generality). Therefore, the control strategy will produce a function that relates buffer occupancy, and other parameters, to the transmission factor (i.e. the coder mode of operation).

2.2. Modelling

The modelling has consisted of:

- Obtaining experimental bit per pixel data for each frame of the original sequence, coded using several transmission factor values.
- Finding a statistical model that best fits the bit per pixel sequence.

Modelling by means of an AR process has been tested for both DPCM and Hybrid DCT codecs ([6], [7], [8]). An AR1 process was adjusted for the experimental sequence. Figure 1 presents the results obtained for the "salesman" sequence.

The purpose of this work was not the video source modelling, but rather the study of buffer control strategies. Therefore, although having a good model for the source was essential for the control strategy here presented, it was left for further studies the refining and assessing of such models. For the purpose of the work it was considered that the results were not good enough to use the AR1 process hypotesis in what followed.

Instead, it was assumed that the video sources, pending further study, could be considered as white noise sources, as far as bit per pixel sequences were concerned. This assumption does not necessarily reflect the actual video source behaviour, which may be different for each type of video service, but rather, at least from the control viewpoint, it is equivalent to a worst-case situation, i.e. the past behaviour of the source does not carry information on the future.

3. The cost functions

Suppose a good model for the source has been found, that is, for any given image, which has been coded with x_n bits per pixel, the conditional probability density function (pdf) $f(x_{n+1}|x_n)$ is known and accurately describes the statistical behaviour of the next image [†].

This pdf would depend, in the case that has been considered, on the coder mode of operation. The *mode of operation* could be defined as the combination of coder variables' values, that define the coder behaviour and are dynamically changed during coding (e.g. as a function of buffer occupancy). Each type of coder allows a different parameter to be changed (e.g. quantizer step in certain DCT coders and bands to be transmitted in sub-band coding). In what follows, this parameter, regardless of what it actually is, will be called *transmission factor* (tf) the name it receives in the DCT coder design that has been considered [1], [2], [3].

For each image, the expected number of bits per pixel will depend on the previous image, the sequence statistics and the mode of operation. Thus, the pdf for the bits per pixel rate can be written as $f_i(x_{n+1}|x_n)$, for each transmission factor, tf_i.

Additionally, suppose that a cost function, c_o, is defined to evaluate, according to the coder design criteria, the effect of the coder producing a certain number, x_{n+1}, of bit per pixel in the next image of the sequence.

Let $C_q(tf_i)$ and $C_x(tf_i, tf_j)$ be, respectively, costs associated to the quality and the change of transmission factor. Let $C_o(tf_i)$, as defined below, be the total cost

[†]Although bits per pixel per image are mentioned, the same would apply to a smaller control unit, such as a image stripe in the case of TV coding

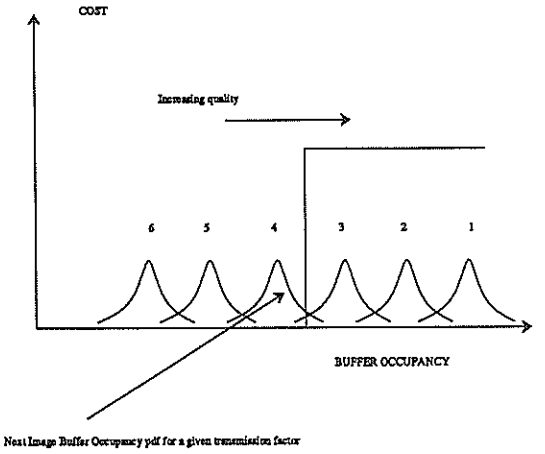


Figure 2: Cost function and next image buffer occupancy pdf for several transmission factors

associated to the buffer occupancy, for the choice of a transmission factor tf_i.

$$C_o(ft_i) = \int_0^\infty c_o(x_{k+1}) f_i(x_{k+1}|x_k) dx_{k+1} \qquad (1)$$

Then, the total cost for each transmission factor could be written as:

$$C(tf_i, k+1) = C_q(tf_i) + C_x(tf_i, tf_j) + C_o(tf_i) \qquad (2)$$

This cost function reflects the two main parameters to evaluate coder performance,

- Distortion, considered here as, both fixed image and sequence, subjective quality.
- Bit-rate, reflected in the cost function that favors maximum buffer occupancy without incurring in buffer overflow (i.e. maximizing the use of the transmission line).

In a real application of the aforementioned cost funtion, different weights should be applied to its 3 components, depending on which is the most important aspect, for the considered service (TV, videophony, videoconference). For instance, a cost function designed for a TV codec could include a $C_x(tf_i, tf_j)$ cost function that would minimize the changes in subjective quality.

The experiments were performed on videophony sequences. The target, in this case, is the maximum reduction in bitrate (minimum bitrate required) to make possible transmitting through a low bitrate channel. In that regard, image quality criteria are secondary and, therefore, a single cost function associated to buffer occupancy is considered. Figure 2 shows the applied cost function for the experiment

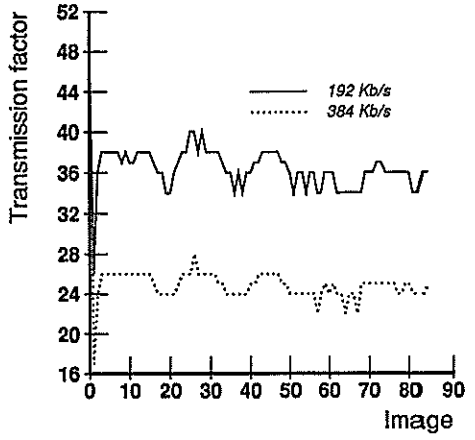


Figure 3: 5 images per sec. simulation: Transmission factor evolution

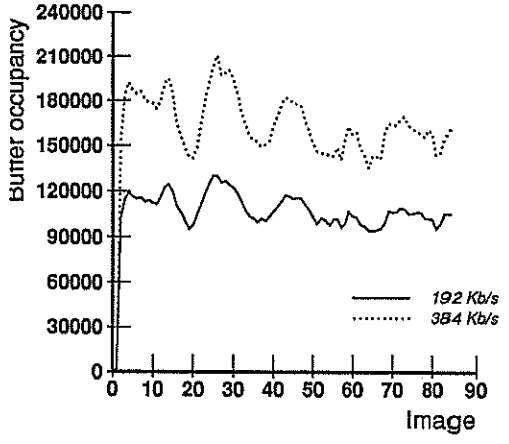


Figure 4: 5 images per sec. simulation: Buffer occupancy evolution

The best choice, given the statistical knowledge of the source and the cost functions, will be the transmission factor that minimizes the costs. In the experiments here presented, a non-zero cost is associated only to buffer overflow situations. The cost, as defined in Equation 1, is calculated, for each transmission factor. The lowest (maximum quality) transmission factor to have a given (and small) cost is chosen for the forthcoming image. For instance, in Figure 2 the costs would be evaluated for transmission factors 1 to 6 and transmission factor number 4 would be chosen.

4. Experiments and Results

To prove the efficacy of the control strategy previously described, a simulation was carried on. In order to simplify the method, the source was taken to be a white noise generator, considering this as a worst-case situation. The noise generator has a different mean and

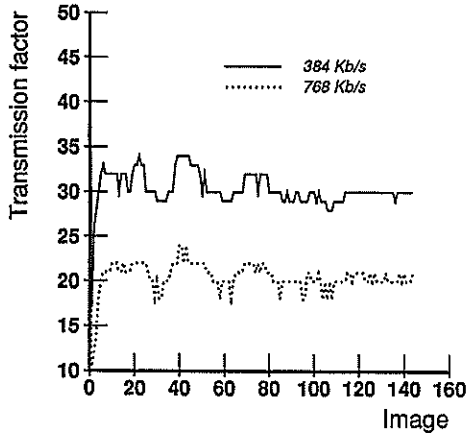


Figure 5: 10 images per sec. simulation: Transmission factor evolution

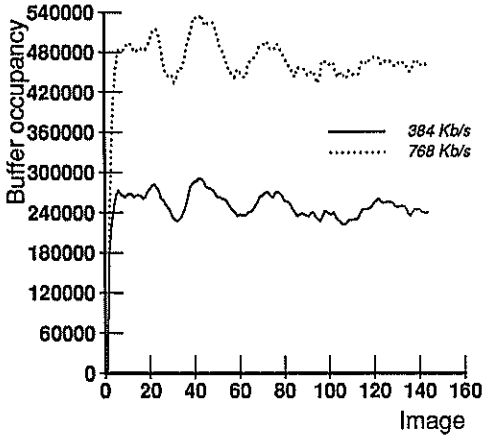


Figure 6: 10 images per sec. simulation: Buffer occupancy evolution

variance for each transmission factor. The data for the video source mean were obtained by coding the "salesman" sequence.

Figures 3, 4, 5, and 6 show that this is a workable control strategy. Figures 3 and 5 represent the transmission factor evolution throughout the simulations, at sample rates of 5 and 10 images per second, respectively. The buffer occupancy, as reflected in Figures 4 and 6, is kept well below overflow levels (Buffers are 1 second long for each bitrate, thus having different physical lengths). As can be seen, the buffer never approaches the overflow level, due to the conservative criterium used for the transmission factor choice (see Figure 2). This could be easily dealt with, by choosing a higher minimum cost in the decision.

5. Conclusions

The main interest of this work is that it provides a general way to tackle the buffer control problem independently of the video service (videophony, videoconference, TV), while classical methods tended to be application-specific.

The work presented here has been devoted so far to the definition of simple costs functions. These cost functions and the knowledge of the video source statistics are the basis of the transmission factor choice, for the next image in the sequence.

In the future, to assess the potential benefits of this approach, an effort has to be made to produce better statistical models and cost functions more adapted to the design criteria.

References

[1] "Draft New Recommendation: Transmission of Component-coded Digital Video Signals for Contribution-quality Applications at Third Hierarchical level of CCITT, Recommendation G.702" Document CMTT/303 October 1989.

[2] "Proposed Modifications to Report AD/CMTT: Digital Transmission of Component-coded Television Signals at 30-34 Mbit/s and 45 Mbit/s", and 45 Mbit/s using Discrete Cosine Transform" Document CMTT/321 October 1989.

[3] N. García, F. Jaureguizar, J. I. Ronda and A. Sanz, "HDTV Parallel Codec Simulator ", Proc. Third International Workshop on HDTV, Torino, August 1989.

[4] G. Karlsson and M. Vetterli, "Packet Video and Its Integration into the Network Architecture " IEEE Journal on Selected Areas in Communications, Vol. 7, pp. 739-751, June 1989.

[5] W. Verbiest, L. Pinnoo and B. Voeten, "The impact of the ATM concept on video coding" IEEE Journal on Selected Areas in Communications, Vol. 6, pp. 1623-1632, December 1988.

[6] M. Nomura, T. Fujii and N. Ohta, "Basic characteristics of variable rate video coding in ATM environment" IEEE Journal on Selected Areas in Communications, Vol.7, pp. 752-760, June 1989.

[7] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson and J.D. Robbins, "Performance models of statistical multiplexing in packet video communications " IEEE Transactions on Communications, Vol. 36, pp. 834-843, July 1988.

[8] P. Sen, B. Maglaris, N.E. Rikli and D. Anastassiou, "Models for packet switching of Variable-Bit-Rate video sources" IEEE Journal on Selected Areas in Communications, Vol. 7, pp. 865-869, June 1989.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Improving the Performance of a Low-rate Image Coder Connected to a Noisy Gaussian Channel

Thomas Woerz Michael G. Perkins
The German Aerospace Research Establishment (DLR)
NE-NT-T
D-8031 Oberpfaffenhofen
West Germany
Tel. *8153/28804

Abstract

We investigate the problem of communicating the output of a vector quantizer image coder over a Gaussian channel. We assume that a fixed bandwidth is available and that the energy per modulator symbol is fixed. Our reference system is a QPSK modulator with a gray code mapping between bits and modulator symbols. Channel signal-to-noise ratios corresponding to QPSK bit error rates of 10^{-1} and 10^{-2} are considered. The techniques investigated include Hamming coding, convolutional coding, 8-PSK trellis-coded modulation, and methods based on simulated annealing coupled with quantizer optimization.

1 Introduction

In this paper we address the problem of communicating the output of a low-rate vector quantizer image coder over a noisy Gaussian channel. The monochrome images to be communicated are 512-by-512 pixels in size with 8 bits of intensity information per pixel. The vector quantizer operates on 4-by-4 blocks of pixels, and the codebook consists of 16 vectors. Four bits are therefore required to communicate each source vector, implying a compression ratio of 32:1. A full search vector quantizer designed using the LBG algorithm is employed [1].

We assume that a fixed bandwidth is available for transmission, and that this bandwidth limits the number of modulator symbols that can be transmitted per second to the rate R. We further assume that the energy per modulator symbol, E_s, is fixed. The performance criterion chosen is the mean-square error between the original picture and the picture reconstructed by the receiver. Subjective evaluations of the image quality are also made because of the known limitations of the mean-square error distortion measure in a human vision context [2].

Our reference system is a QPSK modulator with a gray code mapping between bits and modulator symbols. The decoder makes a maximum-likelihood decision on each received symbol, and outputs the corresponding bits. Note that two QPSK symbols are required to communicate a single source vector, and that two bits are communicated per QPSK symbol. Channel signal-to-noise ratios corresponding to QPSK bit error rates of 10^{-1} and 10^{-2} are investigated.

The performance of a variety of systems is compared to that of the reference system. The techniques to be compared can be classified into two categories: (1) those that reduce the number of source bits communicated per modulator symbol to one, and (2) those that keep the number of source bits communicated per modulator symbol at two. Category 1 techniques require twice as many channel uses per picture as the reference system, while Category 2 techniques require the same number of channel uses per picture as the reference system.

The Category 1 techniques investigated were Hamming coding and convolutional coding; various decoding methods were investigated in conjunction with these two techniques. The QPSK modulator of the reference system was employed. The Category 2 techniques investigated were 8-PSK trellis-coded modulation and methods based on simulated annealing coupled with quantizer optimization.

Section 2 discusses in detail the reference system and the other systems investigated. Section 3 presents the simulated performance of the various systems, and presents a few conclusions that can be drawn from our investigations.

2 Systems Investigated

Figure 1 shows a baseband system diagram general enough to represent all the systems we considered. Referring to this diagram, we assume that the monochrome image source outputs a sequence of pixels from a raster-scanned image where each pixel assumes a

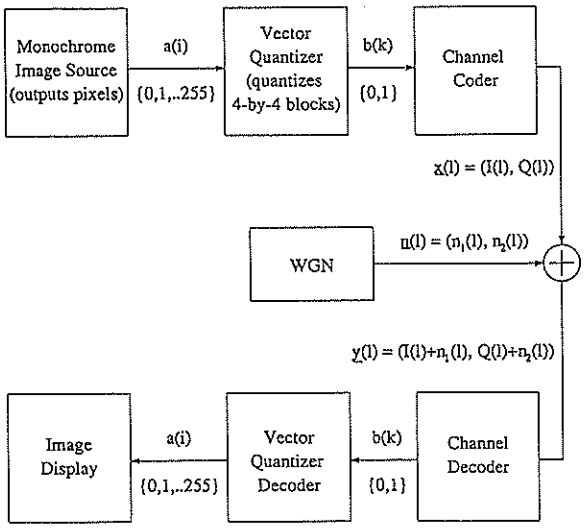


Figure 1: System Investigated

value from the set {0,1,2,...255}. The vector quantizer groups the pixels into 4-by-4 blocks, quantizes each block into one of 16 codevectors, and outputs a corresponding 4 bit sequence for each block. The channel coder processes the bits from the quantizer and outputs a stream of (I, Q) pairs for transmission over the channel. The channel decoder receives the (I, Q) pairs corrupted by additive white Gaussian noise, and outputs an estimate of the bit stream input to the channel coder. Finally, the vector quantizer decoder processes this bit stream to obtain an approximation of the original pixels.

In all the systems investigated we assume that the energy per modulator symbol, E_s, is fixed; in other words, the transmitted (I, Q) pairs must satisfy

$$E_s^2 = I^2 + Q^2. \tag{1}$$

For convenience we choose $E_s = \sqrt{2}$ and adjust the noise variance as required to achieve different SNRs. Below we describe the systems investigated.

2.1 The Reference System

For the reference system the (I, Q) pairs are restricted to the set $\{(\pm 1, \pm 1)\}$; *i.e.* QPSK modulation is employed. The channel coder box in Figure 1 maps quantizer bits onto modulator symbols according to the following gray code rule: $(0,0) \rightarrow (-1,-1)$, $(0,1) \rightarrow (-1,1)$, $(1,0) \rightarrow (1,-1)$, $(1,1) \rightarrow (1,1)$. The channel decoder box makes a maximum-likelihood (ML) decision on each received (I, Q) pair, and outputs the corresponding bits. Note that because of the gray code mapping the most probable symbol errors only cause a single bit error to occur.

2.2 Category One Techniques

As mentioned in the introduction, the techniques in this category reduce the number of quantizer bits transmitted per (I, Q) pair to one. Five category one techniques were investigated:

- **1.1** The channel coder applies the extended (8,4) Hamming code to the quantizer bits, then uses the QPSK gray code mapping employed by the reference system to transmit the coded bits across the channel. The channel decoder makes an ML decision on each received (I, Q) pair until eight bits have been accumulated. It then searches through the codewords of the (8,4) code for the nearest (minimum Hamming distance) codeword to the received sequence; once found, the corresponding quantizer bits are output to the source decoder. Note that for the (8,4) code there is no guarantee of a unique minimum-distance codeword; ties are resolved randomly.

- **1.2** The channel coder applies the extended (8,4) Hamming code to the quantizer bits, then uses the QPSK gray code mapping employed by the reference system to transmit the coded bits across the channel. The channel decoder accumulates four received (I, Q) pairs in a buffer, then makes an ML decision as to which codeword was transmitted. Note that the full power of the (8,4) code is now exploited because soft-decision channel outputs are used in finding the nearest (8,4) codeword.

- **1.3** This system is identical to that of **1.2**, except the channel decoder now makes a maximum-aposteriori (MAP) decision instead of an ML decision.

- **1.4** The channel coder applies a constraint length 3, rate 1/2, convolutional code to the quantizer bits, then uses the QPSK gray code mapping employed by the reference system to transmit the coded bits across the channel. The channel decoder makes an ML decision on each received (I, Q) pair, and outputs the corresponding bits to a Viterbi decoder. The output of the Viterbi decoder is decoded by the source decoder. Decoding in this manner is called hard-decision Viterbi decoding.

- **1.5** The channel coder applies the same code as used in **1.4** to the quantizer bits, then uses the QPSK gray code mapping employed by the reference system to transmit the coded bits across the channel. The channel decoder outputs each received (I, Q) pair directly to a Viterbi decoder. The output of the Viterbi decoder is subsequently decoded by the source decoder. Decoding in this manner is called soft-decision Viterbi decoding.

2.3 Category Two Techniques

Like the reference system, the techniques in this category communicate two quantizer bits per (I, Q) pair. A total of three techniques from this category were investigated. Before listing these techniques, however, we briefly discuss some of the ideas used in these systems.

Let the cost of decoding quantizer vector i as vector j be given by

$$D_{ij} = \|\mathbf{q}_i - \mathbf{q}_j\|^2, \tag{2}$$

where $\|\mathbf{q}_i - \mathbf{q}_j\|$ is the Euclidean distance between vectors i and j. Then the expected distortion introduced by the channel is given by

$$E(D) = \sum_{i=1}^{N} p(i) \sum_{j \neq i} D_{ij} p(j \mid i), \tag{3}$$

where $p(i)$ is the probability of the ith vector and $p(j \mid i)$ is the probability that $\mathbf{q}_j$ is decoded given that $\mathbf{q}_i$ was transmitted. Because the computation of $p(j \mid i)$ is often intractable, it is frequently easier to work with the upper bound

$$E(D) \leq \sum_{i=1}^{N} p(i) \sum_{j \neq i} D_{ij} Q(\gamma_{ij}), \tag{4}$$

where

$$\gamma_{ij} = \frac{d_{ij}^2 + N_0 \ln(p(i)/p(j))}{d_{ij}\sqrt{2N_0}}, \tag{5}$$

$$d_{ij} = \|\mathbf{c}_i - \mathbf{c}_j\|, \tag{6}$$

$$\tag{7}$$

and $\mathbf{c}_i$ is the vector formed by concatenating the sequence of (I, Q) pairs assigned to $\mathbf{q}_i$ into a single vector. We will refer to the set of vectors, $\{\mathbf{c}_i\}$, as the signal vectors. This upper bound is appropriate for a MAP decoder such as the one employed in system 1.3 [3]. For Category 2 techniques, two (I, Q) pairs are used to communicate each quantizer vector, so the signal vectors are four-dimensional.

For a given modulator signal set, $\{(I, Q)\}$, we can use simulated annealing to find the assignment of quantizer vectors to signal vectors that minimizes the bound in Equ 4 [4]. Note that if the cardinality of the modulator signal set is greater than four, then the set of possible signal vectors has a cardinality greater than 16. Once the simulated annealing has selected an optimal or nearly optimal assignment, it is possible to determine through simulation the channel transition probabilities $p(j \mid i)$. Finally, given these probabilities a modified version of the LBG algorithm can be used to optimize the quantizer codebook for the noisy channel [5].

Three Category 2 techniques were investigated:

- **2.1** The channel coder of the reference system is replaced by a 16-by-16 symbol interleaver, followed by an 8-PSK trellis-coded modulator; an Ungerboeck code with memory two was used [6]. The channel decoder outputs each received (I, Q) pair directly to a Viterbi decoder. The output of the Viterbi decoder is subsequently deinterleaved then decoded by the source decoder. Note that quantizer *symbols* were interleaved: *i.e.* groups of four bits were treated as a single entity by the interleaver.
- **2.2** The modulator signal set is chosen to be that of the reference system and simulated annealing is used to find a good assignment of quantizer vectors to signal vectors. The decoder accumulates two received (I, Q) pairs in a buffer, then makes a MAP decision as to which signal vector was transmitted. The quantizer codebook has been optimized for the known channel transition matrix.
- **2.3** This system is identical to that of **2.2** except that the modulator signal set was chosen to be 8-PSK.

3 Results

The results of our investigations are summarized in Table 3. In the table the value $E_s/N_0 = 2.15$dB corresponds to a reference system bit error rate of 10^{-1} while the value $E_s/N_0 = 7.33$dB corresponds to a reference system bit error rate of 10^{-2}. Results for the well-known picture "Lena" from the USC image processing library are reported. The figure of merit tabulated is the PSNR between the original picture and the picture decoded by the receiver. If $I(m, n)$ is the intensity of the pixel in row m column n of the original picture, and $\hat{I}(m, n)$ is the intensity of the corresponding pixel in the reconstructed picture, then the PSNR is defined by

$$\text{PSNR} = 10 \log_{10} \left[\frac{255^2}{(I(m,n) - \hat{I}(m,n))^2} \right]. \tag{8}$$

The number of errors shown in the table is the number of quantizer symbols decoded incorrectly. All tabulated values are averages computed over five noise patterns; five patterns was deemed sufficient because the computed standard deviations were always a small fraction of the average values. The largest PSNR standard deviation, for example, was only 0.17 dB.

At a reference system bit error rate of 10^{-1}, only the Category one techniques were useful. Of these techniques, those exploiting soft-decision decoding significantly outperformed those exploiting hard-decision decoding. Convolutional coding yielded the best performance both in terms of PSNR and visual impression.

One easily visible characteristic of convolutional coding is an error streaking effect in the decoded image. This results from channel disturbances serious enough for the viterbi decoder to select a trellis path that extends over several source symbols before it remerges with the correct path. The subjective quality of the pictures produced by ML and MAP decoding of the Hamming coded pictures was identical, despite the fewer number of errors produced by the MAP technique and its superior PSNR. Although the Hamming code based techniques do not exhibit any error streaking effects, they are still perceptually inferior to the pictures produced by convolutional coding.

At a reference system bit error rate of 10^{-2} it was easy to obtain error-free performance with several of the category one techniques; unfortunately, these techniques require twice as many channel uses as the Category two techniques. The Category two techniques still have noticeable degradations due to the channel errors, however in the case of Technique 2.3 the subjective quality was surprisingly good. Recall that for this technique the signal set was expanded by a factor of two from QPSK to 8-PSK prior to performing simulated annealing and reoptimizing the quantizer codebook. The simulated annealing was able to successfully exploit the increased number of assignment possibilities to signficantly improve performance with respect to simulated annealing on the QPSK signal set. The errors that did occurr tended to be soft in nature; by this we mean that in regions where the original picture is mostly dark, channel errors causing white pixels to appear were rare. It is interesting to note that although simulated annealing using the QPSK signal set resulted in fewer channel errors, worse performance both in terms of PSNR and subjective impression was obtained.

Ungerboeck coding performed worse than the reference system in the 10^{-1} case, however was significantly better than the reference system in the 10^{-2} case. Although Ungerboeck coding resulted in about half as many errors as Technique 2.3 in the 10^{-2} case, the errors that occurred tended to be hard and very noticeable. Of the Category two methods investigated, Technique 2.3 was preferable overall both in terms of PSNR and subjective impression.

It should be mentioned that in the 10^{-1} case the performance advantage of convolutional coding with respect to Hamming coding could have been eliminated by using simulated annealing to find a good assignment of quantizer vectors to channel vectors. We suspect, in fact, that the resulting pictures would look significantly better than those produced by convolutional coding. On the other hand, a soft-decision MAP decoder for a Hamming code must perform sixteen dimension-eight dot products and sixteen additions per quantizer symbol. The viterbi algorithm for the 4-state convolutional code investigated requires only sixteen dimension two dot products and sixteen additions per quantizer symbol: a significant savings. Finally, the performance of convolutional coding could have been improved by using a symbol interleaver to reduce streaking effects in the decoded image.

Tech.	$E_S/N_0 = 2.15dB$		$E_S/N_0 = 7.33dB$	
label	PSNR (dB)	Err.	PSNR (dB)	Err.
Ref.	16.42	5601	23.82	628
1.1	18.81	2580	26.87	34
1.2	22.78	772	27.20	0
1.3	23.37	606	27.20	0
1.4	18.74	2605	27.18	3
1.5	24.52	367	27.20	0
2.1	13.34	11015	25.00	300
2.2	20.59	3924	25.77	478
2.3	21.39	5054	26.11	580

Table 1: Simulation results for the image "Lena".

References

[1] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, January 1980.

[2] D. Sakrison, "On the role of the observer and a distortion measure in image transmission," *IEEE Trans. Commun.*, vol. COM-25, pp. 1251–1267, November 1977.

[3] M. Perkins, "Optimizing signal constellations for the output of a quantizer," in *Proc. of IEEE Globecom Conf., Dallas*, pp. 1895–1900, November 1989.

[4] S. Kirkpatrick, C. Gelatt, and M. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, May 13 1983.

[5] K. Zeger and A. Gersho, "Vector quantizer design for memoryless noisy channels," in *Proc. IEEE Conf. on Commun., Philadelphia*, pp. 1593–1597, June 1988.

[6] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 55–67, January 1982.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Combined Source-Channel DCT Image Coding for the Gaussian Channel

Michael G. Perkins * Tom Lookabaugh †

1 Introduction

We address the problem of combined source-channel DCT image coding for the AWGN channel. The usual approach to this problem is to adopt a fixed modulation format, for example BPSK, QPSK, or 16-QAM, and then to design *bit-oriented* source and channel coders for use with this modulator. For such systems the channel is generally modeled as being binary-symmetric with a bit error probability that depends on the signal constellation and the SNR. Several researchers have investigated such systems [1], [2], [3].

We emphasize a different approach and consider the signal constellation to be an integral part of the system and subject to optimization. We derive locally optimal quantizer/signal-constellation pairs for each DCT transform coefficient using a recently introduced algorithm for joint quantizer/signal-constellation design [4]. Since we do not restrict ourselves to dealing with bits, the signal constellations employed do not necessarily comprise 2^n points for some integer n. Furthermore, the signal constellations are generally asymmetrical. In the following the coder based on these optimized pairs will be referred to as the optimized coder.

We compare the performance of the optimized coder to two reference coders. Reference coder one, referred to as the basic coder, is similar to the optimized coder, however the number of quantizer reconstruction levels is restricted to be a power of two, and the modulation format is assumed to be BPSK. No channel coding is employed. Reference coder two is referred to as the JPEG coder; it, or a derivative, will soon become an international standard (JPEG), and will appear in commercial products. It is a highly efficient variable-rate DCT-based image coder. Error protection is obtained by inserting synchronization information into the source coder's output so that the source decoder can recover from a bit error. BPSK modulation is assumed.

The comparison is made for SNR's corresponding to BPSK error rates of 10^{-2}, 10^{-4}, and 10^{-6}. The comparison between the JPEG coder and the optimized coder is especially interesting as the JPEG coder is an excellent example of the traditional approach to combined source-channel coding.

It is assumed throughout that coherent modulation with perfect transmitter/receiver carrier-phase synchronization is employed. For the two reference coders we assume maximum-likelihood decoding of each received bit; for the optimized coder we assume that the peak transmitted power per signalling dimension is limited to that of the BPSK modulator and that the receiver makes maximum-aposteriori decisions on an appropriate sequence of channel outputs. Subjective evaluation and the peak signal-to-noise ratio between the original picture and the picture decoded by the receiver are used to compare the systems.

These systems are described in more detail in Section 2 and the results of the comparisons are presented in Section 3.

2 Description of the Coders

2.1 The Basic Coder

The basic DCT coder divides the monochrome picture to be transmitted into blocks of size 16-by-16 pixels and then transforms each block. Bits are allocated among the transform coefficients in a manner designed to minimize the mean-square error between the original picture and its coded version; each block uses the same bit allocation. Lloyd-max quantizers optimized for the gamma distribution are used to quantize the coefficients. The natural binary counting assignment of quantizer levels to quantizer bits is used.

The bit allocation algorithm works as follows. Let $d(n)$ be the mean-square error introduced by using n bits to quantize a gamma distributed random variable with mean zero and variance one. Assume that B bits are to be allocated among the 256 transform coefficients. Let n_{ij} be the number of bits assigned to the ith coefficient when the jth bit is to be assigned. For $i = 1$ to 256 compute

$$\Delta_i = [d(n_{ij}) - d(n_{ij}+1)]\sigma_i^2 \qquad (1)$$

where σ_i is the variance of the ith coefficient. The quantity Δ_i is the marginal decrease in the mean-square error that would result from assigning the jth bit to the ith coefficient. Assign the jth bit to the

*The German Aerospace Research Establishment (DLR), NE-NT-T, D-8031 Oberpfaffenhofen, West Germany
†Compression Labs, Inc., 2860 Junction Avenue, San Jose, CA 95134, USA

coefficient with the largest Δ_i value, increment j, and repeat until all bits have been assigned.

Note that since no channel coding is employed, all the available bits are used by the source coder.

2.2 The Optimized Coder

This coder, like the basic coder, divides the monochrome picture to be coded into blocks of size 16-by-16 pixels and then transforms each block. However, instead of allocating bits, the available *signalling dimensions* are allocated among the coefficients.

In order to allocate the available signalling dimensions it is necessary to know the function $D(n_d, \text{SNR})$. This function yields the expected squared error between a sample of a mean zero variance one gamma distributed random variable and the approximating value reconstructed by the receiver when n_d signalling dimensions are used to communicate the sample. It is clear that $D(\cdot)$ depends on the SNR, the quantizer, the signal constellation, and the mapping of quantizer levels to signal constellation points. The optimal quantizer/signal-constellation pair (QS pair) is the one that minimizes $D(\cdot)$.

We found locally optimal QS pairs for given values of (n_d, SNR) by using a QS pair design algorithm [4]. For a fixed number of quantizer levels, l, the algorithm works by iteratively applying two sub-algorithms: one for optimizing a signal constellation for a given quantizer, and one for optimizing a quantizer for a given signal constellation. Using this algorithm, a $D(n_d, \text{SNR}, l)$ curve was generated. The QS pair that minimized $D(n_d, \text{SNR}, l)$ was retained for later use. Note that the number of quantizer levels is *not* limited to be a power of two by this approach. Rather, for given values of (n_d, SNR) an attempt is made to use the optimum number of quantizer levels and an optimum signal constellation. Since a block channel code can be viewed as a mapping between quantizer output bits and signal constellation points, we can view the above optimization procedure as simultaneous design of the quantizer, channel code, and modulator.

Figure 1 shows the performance of the optimized QS pairs found by the above procedure. Curves corresponding to the channel SNRs required for a BPSK modulator to achieve bit error rates of 10^{-2}, 10^{-4}, and 10^{-6} are shown. The performance of the reference coder's quantizers across a noiseless channel is also shown. As expected, as the SNR increases, the performance across the channel for a fixed number of signalling dimensions improves.

In the 10^{-2} case when four or more dimensions are used, the number of quantizer levels, l, in the optimized QS pair is less than 2^{n_d}. This effectively means that the output of the quantizer is block coded by a rate $\log_2(l)/n_d$ channel code with respect to BPSK modulation. When $l < 2^{n_d}$, the signal constellation points

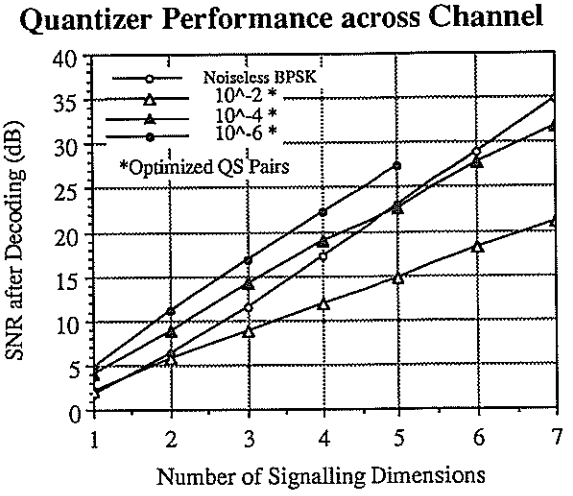


Figure 1: Optimized QS Pair Performance

found by the algorithm always lie on the vertices of an n_d-dimensional hypercube.

Figure 2 illustrates the 2-D signal constellation used by the optimized QS pair for the 10^{-4} case. The corresponding quantizer reconstruction values and their probabilities are shown to the right of the signal constellation. It is interesting to note the large separation between levels 1 and 7; this is reasonable as the largest error occurs when 1 is transmitted and 7 is decoded.

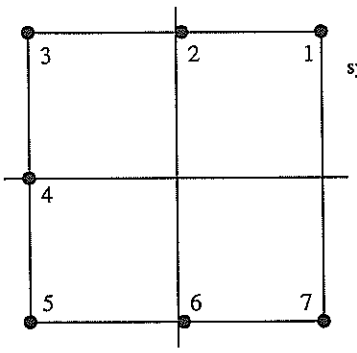
The algorithm for allocating signalling dimensions is identical to that used for allocating bits in the basic coder, except that the expression for Δ_i is changed. For a given SNR, Let n_{ij} be the number of *dimensions* assigned to the ith coefficient when the jth dimension is to be assigned. Then the marginal decrease in mean-square error is now given by

$$\Delta_i = [D(n_{ij}, \text{SNR}) - D(n_{ij} + 1, \text{SNR})]\sigma_i^2. \quad (2)$$

2.3 The JPEG Coder

The Joint Photographic Experts Group (JPEG) has been working for several years under the auspices of ISO and CCITT to develop a versatile high performance image compression algorithm for use as an international standard. The technical portion of the standard is near completion [5]. We refer to the coder based on this standard as the JPEG coder.

The baseline algorithm performs a discrete cosine transform on each 8 by 8 block of the image. The transform domain coefficients are quantized using uniform stepsize quantizers with stepsize depending on the location of the coefficient. The quantized DC coefficient is predicted by the quantized DC coefficient of the previous block in raster scan format. The differ-



symbol	value	prob
1	-3.59	0.016
2	-1.74	0.060
3	-0.65	0.153
4	0.00	0.540
5	0.64	0.154
6	1.73	0.061
7	3.59	0.016

Figure 2: Optimized 2-D Signal Constellation, 10^{-4} case

ence between predicted and actual coefficient is Huffman coded. The remaining quantized coefficients are converted to a one-dimensional array by zig-zag scanning and then coded using a combination of run-length and Huffman coding. In essence, the array is parsed into runs consisting of a sequence of zero or more zeros followed by a single non-zero quantized coefficient, and each such run is coded as a single Huffman codeword. Following the last non-zero coefficient in a block, a reserved Huffman codeword indicating "end of block" is transmitted.

A bit error in the JPEG encoded bit stream can have several effects. If the error is in the header and control information (containing information on frame size, number of color components, etc.), a catastrophic failure to decode is likely to result. If the error occurs within a Huffman codeword, that codeword will certainly be incorrectly decoded. A more serious problem is that the decoder may lose synchronization since the incorrectly decoded Huffman codeword may have a different total length compared to the original codeword. Decoding will proceed in a completely erroneous fashion until the length of incorrectly decoded Huffman codewords equals exactly the length of some sequence of original codewords. If this serendipitous resynchronization occurs relatively quickly, then the error effects will be spatially limited, although usually quite visible. However, if resynchronization does not occur until, say, more than one 8 by 8 block's worth of bits later, then the DC coefficient prediction is likely to be corrupted and serious and continuing error propagation will result. Most extended unsynchronized decoding episodes can be detected by syntax violations such as decoding more than 64 coefficients in a block.

Recognizing the seriousness of error propagation during decoding, the JPEG committee has provided for the insertion of resynchronization codes after each n blocks, where n is chosen by the encoder and transmitted in the header information. The resynchronization codes have a three bit sequence number attached which helps recovery from damaged resynchronization codes (although errors in sequence numbers may themselves cause serious error propagation). After each resynchronization code, the DC prediction is set to zero to control DC prediction error propagation. Since resynchronization codes cost bits, the cost in overhead must be weighed against the value of the additional error protection.

We implemented a JPEG encoder and decoder based on [5]. The decoder included extensive syntax checking to detect unsynchronized decoding. Upon detecting an error, the decoder replaces blocks in the interval bracketed by resynchronization codes and containing the detected error. The visual effect of block replacement is reduced by predicting the DC value of the replaced block based on surrounding blocks, but no AC information is used.

3 Results and Conclusions

Figure 3 compares the performance of the optimized coder to the basic coder at different BPSK bit error rates for the Lena picture. An average of 0.5 signalling dimensions per pixel was used; this translates to 0.5 bpp for the reference coder. The ordinate is the PSNR between the original Lena picture and the picture reconstructed by the receiver. The gain is most dramatic at low bit error rates, but still significant at a bit error rate of 10^{-6}. It is interesting to note that on channels correpsonding to BERs of 10^{-4} and 10^{-6} the optimized coder outperforms what the basic coder achieves on a noiseless channel. At a BER of 10^{-2}, the subjective gain is quite dramatic: bit errors in the basic coder frequently cause whole blocks to be dramatically disturbed, whereas symbol errors in the optimized system are frequently unobserved or only mildly objectionable.

The performance of the JPEG algorithm is shown in Table 1. At each error rate and resynchronization interval, we generated twenty different error patterns for the Lena image and then attempted to decode the image. We declared a catastrophic failure to decode if the header information was sufficiently damaged to prevent decoding or if more than 50% of the decoded image area was replaced due to detected errors. For successfully decoded images, we determined the average percentage of area obscured by replacement and a PSNR figure. In all cases, the quantization stepsizes were linearly scaled to achieve 0.5 bpp including the resynchronization code overhead.

The JPEG algorithm completely fails to decode at a BER of 10^{-2} regardless of resynchronization inter-

Resynch Interval	Performance criterion	Bit Error Rate 0	10^{-6}	10^{-4}	10^{-2}
1 block	PSNR % cat % obsc	28.9	28.9 0% <1%	21.9 40% <1%	100%
4 blocks	PSNR % cat % obsc	34.9	34.9 0% 0%	31.0 15% 1%	100%
64 blocks	PSNR % cat % obsc	35.7	35.7 0% 0%	23.4 5% 7%	100%
1 frame	PSNR %cat %obsc	35.8	35.8 0% 0%	100%	100%

Table 1: Performance of the JPEG Coder

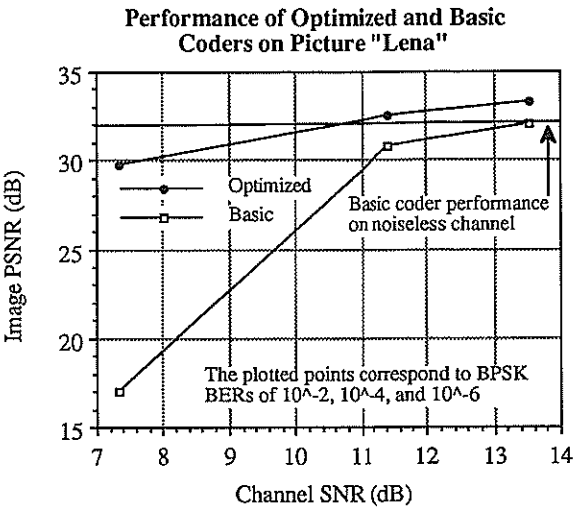


Figure 3: The Optimized Coder Compared to the Basic Coder

val. At 10^{-4}, a four block resynchronization interval provided the best PSNR but 15% of the images catastrophically failed to decode. Even though successfully decoded images had a relatively high PSNR, the visual effect of errors was quite annoying (clearly visible distorted blocks for undetected errors and lost information for detected errors). At 10^{-6}, the algorithm behaved quite well and a 64 block resynchronization interval would control the infrequent errors with a small (0.1 dB) decrease in performance due to resynchronization overhead.

We would typically expect the JPEG algorithm to be combined with a channel error control code. Given the goal of both power and bandwidth efficiency, the best candidates for such codes are relatively recent coded modulation schemes such as Ungerboeck's trellis coded modulation [6]. Such codes seem quite capable of correcting to an error event probability of 10^{-6} or below on an AWGN channel which supports a BER of 10^{-4} with uncoded BPSK with no expansion in bandwidth. However, it seems much more difficult to achieve this type of performance on a channel which supports a BER of 10^{-2} with uncoded BPSK, even for codes with high asymptotic coding gains (and hence high complexity), so that it will be difficult to successfully use the JPEG algorithm on such high-noise channels.

Of the systems compared in this paper, the optimized coder is preferable for channels supporting BPSK bit error rates of 10^{-2}, while, coupled with trellis-coded modulation, the JPEG coder is preferable for channels supporting BPSK bit error rates of 10^{-4} or less.

References

[1] V. Vaishampayan and N. Farvardin, "Optimal block cosine transform image coding for noisy channels," Tech. Rep. TR-86-37, University of Maryland, Systems Research Center, August 1986.

[2] J. Modestino, D. Daut, and A. Vickers, "Combined source-channel coding of images using the block cosine transform," *IEEE Trans. Commun.*, vol. COM-29, pp. 1261–1274, September 1981.

[3] D. Comstock and J. Gibson, "Hamming coding of dct-compressed images over noisy channels," *IEEE Trans. Commun.*, vol. COM-32, pp. 856–861, July 1984.

[4] M. Perkins, "Joint vector quantizer and signal constellation design for the Gaussian channel," in *Abstracts of Papers, International Symposium on Information Theory, San Diego*, p. 35, January 1989.

[5] "JPEG technical specification, revision 5, JPEG-8-R5, ISO/IEC JTC1/SC2/WG8, January 2, 1990." Available from X3 Secretariat, 311 First Street NW, Suite 500, Washington DC 20001-2178, USA.

[6] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 55–67, January 1982.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Performance Evaluation of High Resolution Image Compression Algorithms in Presence of Transmission Noise

L. Alparone, G. Benelli, F. Fabbri

Dipartimento di Ingegneria Elettronica, Università di Firenze,
via S.Marta 3, 50139 Firenze (Italy)

High Definition Television (HDTV) images are generally compressed in order to reduce the bandwidth requirements for transmission. However, compressed images are more sensitive to noise of the communication channel. In this paper the effect of channel errors on HDTV images compressed through techniques based on Discrete Cosine Transform (DCT) and Vector Quantization (VQ) is analyzed. Channel coding techniques suitable for HDTV images are also presented. In particular, algebraic and soft decoding algorithms are described. The performance of the described schemes is determined trough simulations.

1. INTRODUCTION

Many compression techniques have been developed to reduce the bit rate for digital images transmission, maintaining, at the same time, a good fidelity between the original and the compressed image. Some of the most powerful compression algorithms are based on the Discrete Cosine Transform (DCT) [1],[2] and Vector Quantization (VQ) [3],[4]. In these techniques compression is attained exploiting the spatial or temporal correlation of the original image pixels.
Data compression techniques will be extremely important in the development of High Definition Television (HDTV). In this case the information needs a very high transmission rate; further, good quality is a fundamental request for HDTV images, hence sophisticate compression algorhitms are required to avoid a significant distortion in the compressed images.
Further degradation is inevitably added during the transmission over a noisy channel. This degradation is more evident in compressed images, since compressed data are generally much more sensitive to noise than correlated source data. Decorrelation of original data, made during the compression step, introduces a high correlation between noise samples and pixels of the received and decompressed image. Hence, errors in few compressed data spread over many pixels of the reconstructed image. Degradation will be as more evident as more complex is the compression technique. Therefore, some form of error control protection needs be applied if high quality reconstruction is requested.
Research in this sense has been carried on for what concern DCT-processed images by Comstock and Gibson [5], Modestino, Daut and Vickers [6]. Many other works are dedicated to image coding using Vector Quantization but with no considerations about the effects of channel errors.
In this paper the effect of channel errors in the transmission of images compressed by DCT and VQ is investigated and some techniques of channel coding are proposed and performances are evaluated. Image transmission with PSK modulation over a AWGN channel has been simulated.

2. DISCRETE COSINE TRANSFORM AND VECTOR QUANTIZATION

The two dimensional DCT (2D-DCT) is a unitary transform commonly used in image compression. The 2D-DCT of the sequence $f(i,j)$, for $0 \le i,j \le N-1$ is defined as:

$$F(h,k) = (2/N)c(h)c(k)\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} f(i,j)\cos\left[\frac{(2i+1)h\pi}{2N}\right]\cos\left[\frac{(2j+1)k\pi}{2N}\right] \qquad (1)$$

for $0 \le h,k \le N-1$

being $c(0)=1/\sqrt{2}$, $c(h)=1$ for $h=1,2,\ldots N-1$
The inverse 2D-DCT transform is given by:

$$f(i,j) = (2/N)\sum_{h=0}^{N-1}\sum_{k=0}^{N-1} c(h)c(k)F(h,k)\cos\left[\frac{(2i+1)h\pi}{2N}\right]\cos\left[\frac{(2j+1)k\pi}{2N}\right] \qquad (2)$$

for $0 \le i,j \le N-1$
Like other unitary transforms, DCT has the property of compacting image energy

in few coefficients corresponding to the lowest spatial frequencies. The coefficient F(0,0) is called d.c. coefficient because is related to the average gray level of the image. In a typical compression technique based on DCT, original image is partioned in square blocks of NxN pixels (N=16 in our case) [1]. A bidimensional transformation is applied to each block. Compression is achieved by discarding the lowest energy coefficients. The selected coefficients are quantized, properly coded and transmitted. The number of quantization bits of each coefficient is calculated according to a scheme in which the number of bits is assigned proportionally to the variance of the coefficient. The more important coefficients are quantized with a number of bits greater than lower energy coefficients.
In this technique, transformed coefficients are scalar quantized. Better performances are achievable by quantizing and coding vectors instead of scalar values in a set of correlated data, like image data. Compression methods using vector quantization are based on this assumption. A N level vector quantizer is a function that maps each input vector $x=(x_1.....x_n)$ into a representative vector $y=q(x)$, (codeword), belonging to a finite set $Y=(y_1.....y_N)$, (codebook).
The quantizer is defined by Y and by the set of partitions $S=(S_1.....S_N)$ with

$$S_i = \{x:q(x)=y_i\} \qquad (3)$$

The partition S_i is called the i-th Voronoi region. The first step in a compression technique based on VQ is the decomposition of original image in a set of vectors. Many kinds of decomposition have been proposed [3]. We have considered small blocks of 4x4 adjacent pixels as vectors. A scansion per columns of these vectors generates a sequence X called "training set" (step 2). This sequence of vectors is used to generate the codebook Y with an iterative algorithm due to Linde, Buzo and Gray [4] (step 3). Let us assume as distortion criterion the Minimum Square Error (MSE). In this case the optimal codebook satisfies two necessary conditions. First, the input set X is partioned in a set of N Voronoi regions:

$$S_i = \{x:|x-y_i| \leq |x-y_j|, i \neq j\} \qquad (4)$$

$i=1,2,....,N$
Then, the codeword correspondent to the region S_i is the mean value of the region:

$$y_i = E\{x|x \in S_i\} \quad i=1,2,...,N \qquad (5)$$

After this, each input vector is quantized with the closest codeword. Compression is achieved by replacing the rapresentative codeword with a label. Reconstruction of the image can be achieved by using the label as an address in a table containing the codebook.

3. EFFECTS OF CHANNEL NOISE ON HDTV IMAGES COMPRESSED BY DCT AND VQ

The transmission channel is assumed to introduce an Additive White and Gaussian Noise (AWGN). A binary Phase Shift Keying modulation (PSK) is also assumed. In this case the bit error rate (BER) of the demodulator output is given by:

$$p = q(\sqrt{[2E/No]}) \qquad (6)$$

being E the energy of the transmitted signal, No/2 the double-sided noise spectral density and q(x) the Q-function. In this section a qualitative description of the noise effect is given. Results of some simulations giving the trend of the signal-to-noise ratio of the reconstructed image (SNR) varying with the signal-to-noise ratio of the channel (E/No) are reported.
By denoting with σ^2 the variance of the original image and e^2 the mean squared error between the original $(f(i,j))$ and the reconstructed image $(f\tilde{}(i,j))$ the signal-to-noise ratio (in dB) of the reconstructed image is defined as:

$$SNR = 10Log\frac{\sigma^2}{e^2} \qquad (7)$$

In an image compressed through the DCT algorithm a channel error in a single coefficient determines a reconstruction over all the block containing this coefficient. Errors in the d.c. coefficients are more evident since they cause a change in the average gray level of the block. In order to visualize the effects of the errors in the reconstructed images, we have performed a computer simulation. The monochrome image used in the simulations is a 512x512 digital image derived from an HDTV signal and is shown in Fig.1(a). Fig.2(a) shows the reconstructed image by assuming a bit error probability of 10^{-3} in the transmission channel. Fig.3 shows SNR versus E/No in the channel.
Let us consider now the effect of the transmission noise on the VQ-compressed image. A reconstructed image is shown in Fig.1(b) for a $BER=10^{-2}$. However in this case two different sets of data are transmitted: labels and codewords. From a

qualitative point of view, errors in the two sets cause different degradations in reconstructed images at the receiver.
To receive a wrong label causes a decision for an incorrect codeword. To the vector corresponding to that label will be not associated the closest codeword, i.e. the codeword introducing the minimum distortion. A wrong bit in a label, hence, causes an error in every pixel forming the codeword relative to that label. In the received image blocks of 4x4 pixels clearly wrong are visible. Fig.1(c) shows the reconstructed image by assuming that the transmission channel introduces errors only on the labels.
Let us consider the effects of the channel noise on the transmission of the codebook. An error in a codeword propagates over more pixels of the reconstructed image because more training vectors can be quantized with the wrong codeword. The fault pixel is visible every times the codeword is used during image reconstruction procedure. In other words, an error in the i-th codeword will be propagated over the whole i-th Voronoi region of the reconstructed image. This propagation is more evident if the compression ratio increases. Fig.1(d) shows the reconstructed image when the channel introduces errors only on the codebook. Fig.4 gives SNR in the reconstructed image when a VQ alghoritm is used.

4. CHANNEL CODING

In the previous paragraph we have seen how much seriously DCT and VQ compressed images are degraded by channel noise. Hence, to obtain satisfactory images some form of error control coding is necessary. Redundant bits must be applied to the source coding bits to correct some of the bit errors occurring during the transmission. Channel coding techniques involve inevitably an expansion of the transmission bandwidth and this is clearly against the goals of data compression operation. Hence, we have applied channel coding techniques in a way requiring no bandwidth expansion. The adopted strategy is to use some of the source coding bits as redundant bits for the channel coding, [5],[6].
For images compressed with the DCT-based technique single error correcting Hamming codes (7,4) (see Fig.2(b)), (15,11) and (31,26) have been used. Performances have been evaluated by reporting SNR varying with (br*n/k*E/No), where br is the bit rate of the compressed image, n/k is the inverse of the code rate (Fig.5). To avoid bandwidth expansion comparison between channel coded case and uncoded case is made at the same value of br*n/k. In this way the transmission bandwidth is assumed equal for all reconstructed images.
Images compressed with the VQ-based technique are more heavily degraded by channel errors. Hence, more powerful coding techniques have been used to obtain good quality images at the receiver. A very powerful code is the Golay (23,12) code, which corrects every combination of three or less errors in 23 bits. The uncoded image is compressed using 256 codewords (br=0.625), while the coded image using 64 codewords (br=0.4); results are shown in Fig.2(c).
The sensitiveness to noise of VQ processed images suggests the use of very efficient coding techniques. It is well known that the performances of a channel coding system can be improved by using soft-decision decoders (Fig.2(d)). In this case, with the same redundancy, it is possible to achieve better performances in terms of coding gain. Fig.6 shows the results attainable using soft-decision techniques in decoding of VQ compressed images. It is evident the gain in the value of SNR, especially for high BER, obtained without further redundance in the transmitted bits. The code used in this case is Hamming (31,26). The uncoded image is compressed with 256 codewords, the coded image with 128 (br=0.5).

ACKNOWLEDGMENT

We thanks the RAI, Centro Studi, in Turin which furnished the original picture used in this work. The work has been carried out under the financial support of the National Research Council (CNR) in the frame of the Telecommunication Project.

REFERENCES

[1] W.Chen, H.Smith: "Adaptive Coding of Monochrome and Color Images", IEEE Trans., 1977, COM-25, n. 11

[2] A.K.Jain: "Image Data Compression: a Review", Proc. of IEEE, vol.69, n.3, mar.81

[3] N.M.Nasrabadi, R.King: "Image Coding Using Vector Quantization: a Review", IEEE Trans., 1988, COM-36, n.8

[4] Y.Linde, A.Buzo, R.M.Gray: "An Algorhitm for Vector Quantizer Design", IEEE Trans., 1986, COM-34, n.7

[5] D.R.Comstock, J.D.Gibson: "Hamming Coding of DCT-compressed Images Over Noisy Channel", IEEE Trans., 1984, COM-32, n.7

[6] J.W.Modestino, D.G. Daut, A.L. Vickers: "Combined Source-Channel Coding of Images Using the Block Cosine Transform", IEEE Trans., 1982, COM-29, n.9

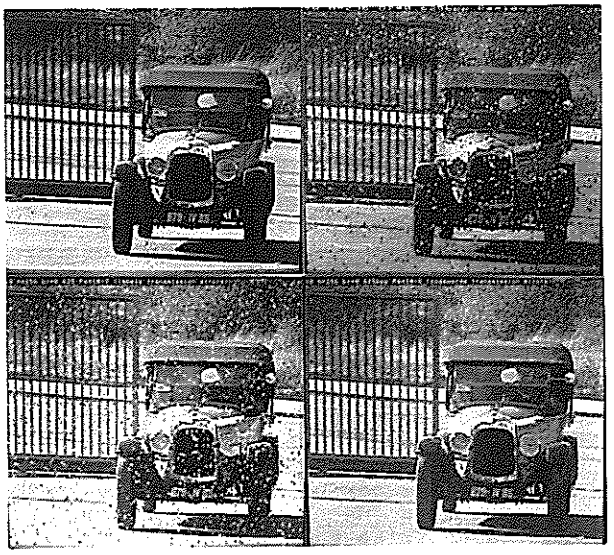


Fig.1 a) b)
c) d)

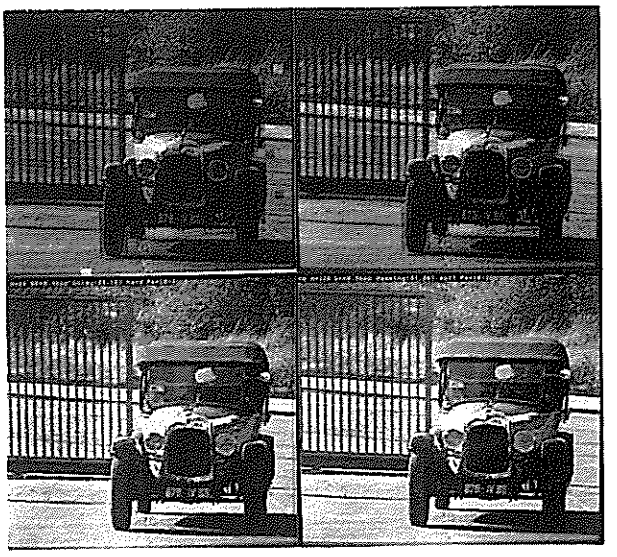


Fig.2 a) b)
c) d)

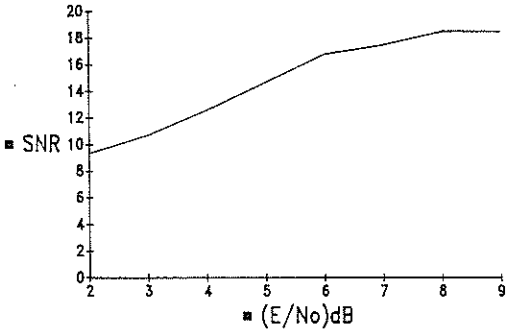


Fig.3

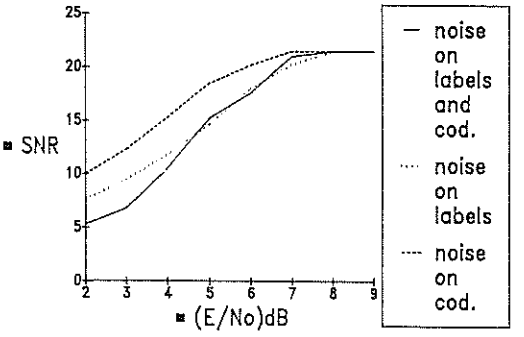


Fig.4

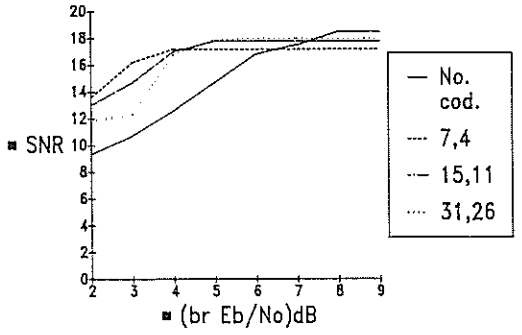


Fig.5

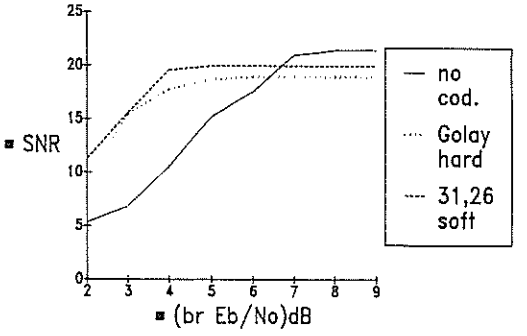


Fig.6

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Quantization Algorithm and Buffer Regulation for Universal Video Codec in the ATM Belgian Broadband Experiment *

Leduc J.P. - Poncin O.
Laboratoire de Télécommunications et d'Hyperfréquences
Université Catholique de Louvain
Bâtiment Maxwell, B-1348 Louvain-la-Neuve, Belgium

Abstract

This paper deals with the coding of digital television, the TV codec aims to reduce the incoming TV bit rate of 216 Mbits/s with a factor of 6 to 100 to eliminate all the redondant information. All the main 'know how' stays in the quantization algorithm which has to preserve the useful information and in the regulation which maintains the trade-off between constant image quality and level of the output bit rate. The goal of the paper is to show one of the multiple way to tackle this challenge in the case of the flexible codec built for the ATM Belgian Broadband Experiment [5].

1 Introduction

This paper intends to describe the core of a digital TV codec based on DCT transform computation i.e. the algorithm of both the quantization and the buffer regulation. It is designed to fit for either variable or fixed bit rate channels.
When the transmission channel is of fixed bit rate (STM channel), the image quality is forced to fluctuate according to the informational content of the sequences; a buffer storage and a regulation have to be implemented in the codec to efficiently deal with those fluctuations. When intrinsically variable bit rate channels are considered (ATM networks), the main advantage to be gained from such a variability is an image quality almost invariable in time and in space (i.e. within the image) except in fall-back modes when the information content has such a high density that a temporary degradation of quality is required to comply with the bit rate specifications negotiated at call setup between the codec and the network.
The quantization algorithm allows to realize bit rates ranging from 2 Mbits/s to 40 Mbits/s with a corresponding quasi-constant picture quality; 2Mbits/s is related in our case to videoconference of high quality and 40 Mbits/s to the standard contribution application.
A constant quality in space is achieved locally by an adaptive quantization based on the computation of a local block criticality.
A constant quality in time could be obtained with a buffer regulation built with a PID taking into account further additional requirements.

*The following text presents research results realised with ALCATEL-BELL-SDT company in the framework of the Belgian Broadband Experiment Project. The scientific responsability is assumed by its authors.

2 Adaptive quantization

The research presented here has been performed for DCT codecs and more specially codecs without motion compensation. Adaptive quantization is more efficient when a motion compensation is not performed, this is the case for VBR codecs where a good quality can be obtained at the expense of higher peak bit rate when needed.

2.1 Weighting of transform coefficients : coefficient by coefficient adaptation

The DCT performs a spectral discrimination of the picture; the high order coefficients represent, at least to some extent, the high spatial frequency contents of the picture. Due to the fact that high-frequency noise is less visible than low-frequency one, high order coefficients can be encoded more coarsely.

Therefore, a first step consists by multiplying each transformed coefficient by a "*visibility weighting factor*" according to its order. More details about this weighting of DCT coefficients can be found in [1], [2].

2.2 Block by block adaptation

Many simulations have proved some theorical considerations about visual impact of quantization noise : some blocks are more sensitive to quantization noise

than others. In this paper, those blocks are called *"critical"*.

Therefore, an uniform quantization applied to each block leads to a constant quantization noise but also to a variable quality picture. The aim of this chapter is to determine some criticality classes; an appropriate quantization according to each block being the final goal.

2.2.1 Criticality notion

The criticality of a block is an image of the block resistance to quantization noise : the more a block is critical, the more quantization noise is visible there. Such a notion is different from the activity of a block, the activity expresses the energy contained in the transformed coefficients. A block may be active without being critical-as the foliage of a tree-, critical but not active -as oblique edges- and so on. Due to the nature of the DCT, blocks with oblique and fine detailed structures are the most sensitive to quantization; on the contrary, horizontal, vertical and random structures are the most robust ones. Therefore we want to reduce quantization noise (by an adapted quantization) for critical blocks to the detriment of less critical ones. So such a quantization is called *"adaptive"*. Quantization noise will be variable within one picture but the subjective quality will be uniform there.

2.2.2 Criticality computation

Several criteria for criticality determination can be found in pel domain or in DCT domain. We have chosen the second option, it provides best results mainly because of the source of the criticality is to be found in the DCT definition.

Only AC coefficients contain some information about the block criticality, therefore only AC coefficients will be examinated for criticality determination.

It has appeared necessary to split up 63 AC coefficients into three regions; this splitting allows one to distinguish structures with vertical or horizontal edges, oblique edges, fine details or random structures. Three main regions have to be examinated as illustrated in figure 1.

The first segmentation of the transform domain requires to discriminate an area containing the first row and the first column from all the other AC coefficients. In fact the DCT concentrates the energy of horizontal and vertical structures on only a few AC coefficients respectively in the first row and in the first column because of the horizontal and vertical definition of DCT basis functions.

Areas II et III contain respectively middle diagonal and high diagonal frequencies.

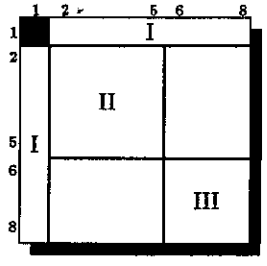


Figure 1 - DCT domain splitting

When processing oblique structures, the transformation spreads the energy over the whole set of AC coefficients; their amplitude is inversely proportional to the frequency. This spreading of energy has at least two main explanations :

1. the DCT basis functions are defined according to a rectangular sampling mesh;
2. as computing a DCT transformation is equivalent to a two-dimensional symetrization in the pel domain followed by a DFT computation. Spectral interpretation in the DCT domain has to be considered taking into account this symetrization : oblique structures generate AC coefficients of high magnitude in both vertical and horizontal frequencies as well as in diagonal frequencies.

The following step is to define a parameter characterizing each area. The best parameter is the maximum of absolute value of the region coefficients. Therefore, we have three local maximum AC values : one for each area.

Ten tests performed on those maximum AC values allow to classify blocks into four criticality classes. Only two bits per block are needed to indicate its class (only 1.6 % of the bitrate for a transmission at 20 Mbit/sec).

2.2.3 Quantization

For each class, we choose a matrix of visibility weighting factors and we adapt the average level of quantization according to the buffer regulation.

Therefore we have :

- an adaptation block by block -*by using criticality information*-,
- a regulation of the average level of quantization stripe by stripe -*as it will be explained in the next chapter*-.

3 Buffer regulation

The buffer regulation is based on models for both the TV incoming information and the codecs. It aims to work for either fixed or variable bit rate transmissions.

The incoming TV information has been modeled with a cyclo-stationary stochastic process with a constant or slowly varying mean during each particular sequence. Consequently, the decorreleted information is composed of (1) a concatenation of steps (or slope) characterized by random height and duration (the height of the information step depends mainly on the intrinsic sequence complexity; this long term process implies a long term regulation) and of (2) a cyclo-stationary stochastic process with pseudo-cycles displaying a period of a field or a frame (short term process).

The codec regulation is designed to cope with this previous modeling: a buffer smooths the short term variations and a feedback regulation of the the buffer occupancy to the quantization step monitors the response to the incoming steps (or slope) of information density; the use of a PID regulator adds some properties to the regulation: it monitors the transient behaviour and the steady-state error and, moreover, it prevents the buffer from overflowing.

3.1 Codec at fixed bit rate.

The quantizer has been modelized [3] with a non-linear law. In term of entropy bit rate averaged on the duration of 8 video lines (termed a stripe), the quantizer output D_q is a function of the input D_{in} expressed as follows:

$$\begin{aligned} D_q(n) &= D_{in}(n) \exp\left(-k \times (\delta - \delta_0)\right) \\ &= D_{in}(n) \exp\left(-k \times TF(n)\right) \end{aligned}$$

where δ is the number of suppressed bits then with q being the quantizing step: $q = 2^{\delta}$ and δ_0 corresponds to the finest step of quantization; n is the sampling time (the sampling period=one stripe). TF is the quality control command called the transmission factor fixed constant during a whole stripe. This model is valid for pictures of standard and high definition. A "quasi-time constant" δ induces a "quasi-time constant" picture quality when corrected locally by the criticallity measurement.

The control loop is built as shown in figure 2. The quantizer adapts its quantizing step according to the quality level imposed by the transmission factor. The quantized output stream is then entropy coded by a variable length coder [4]. This process involves a delay of one sampling period: it needs the whole quantized stripe before starting to encode all the bits; its output is a variable length code which contributes stripe by stripe to the fixed channel bit rate D_{can}. The exceeding stripe bit stream eps(n) is routed to a buffer. At the end of each stripe, the buffer occupancy BF(n) is sized and compared to its reference BF_{ref}. The result is applied to a PID regulator which performs a scaled addition of the summator I(n), the differentiator D(n), and, the instantaneous buffer occupancy $(BF(n) - BF_i)$. The output of the PID regulator constitutes the new computed version of the transmission factor and is applied as TF(n) during the following stripe (i.e. delayed by one sampling period).

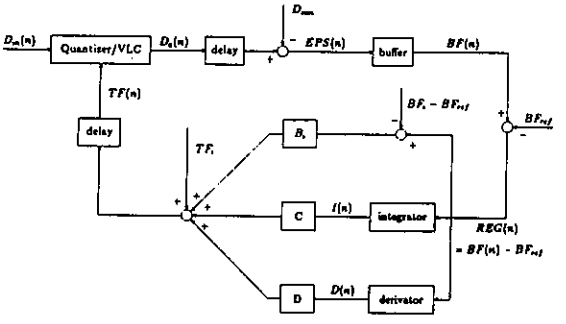


Figure 2 - Control loop model with PID regulator

The complete loop study leads to discrete non-linear difference equations of the fourth order for either $TF(n)$ or $BF(n)$. The loop stability has been studied with several tools: for the approximated linear model, by Routh-Hurwirth and Schur-Cohn criteria and eventually by the Nyquist Theorem; for the non-linear model, by the second theorem of Ljapounov, the Banach contraction theorem, the Schröder functional equation and by the phase portraits. In case of stability, even more stringent requirements are to be fulfilled to generate acceptable image quality: preference is stressed on monotone decreasing solutions which forbid stripe by stripe oscillatory evolution of the transmission factor and therefore of the local image quality. Complex solutions in conjugated pairs have to be avoided; for the same reasons, real negative root have to be excluded from the set of the solutions. Acceptable solutions are either critical damping provided by two pairs of equal real and positive roots inferior to unity or any heavier damped response.

3.2 Codec regulation for variable bit rate.

Codecs transmitting on variable bit rate channels are provided with some kind of police function. The part

of a police function is to negociate at call setup a bit rate characteristics and to monitor the output bit rate to ensure that it complies with the precise agreed limits.
In this case of variable bit rate channels, the quality can be maintained constant (TF(n)=constant) except when the output bit rate exceeds the negociated profil in this case a regulation is switched on. The reaction loop is therefore somewhat rearranged allowing two working modes:

- with images of low and middle density of information no regulation is required except a buffer to achieve some profil of the output bit rate. The quantizing step is monitored by a constant value of TF imposed by the broadcaster leading with the use of the criticality to a constant quality level. All bit rate beyond the profil is accumulated into the buffer. The codec spends a major part of its time in this nominal mode.

- with sequences of too high density of information, a fall-back working mode must be implemented to comply with the negociated bit rate characteristics and to avoid buffer overflow. A regulation identical to that previously described is activated to insure a return to an acceptable bit rate within a specified transient time. In this case, the value of TF(n) is increased to force the bit rate to decrease, consequently, the quality is no more constant but slightly degraded until the decrease of the input density of information.

Fixed bit rate transmissions can be considered as a particular case with constant profil. Discussion of police functions is not the field of this paper, nethertheless, let us notice that the regulation proposed here is independant of that function.

4 BBE Codec

Those adaptive quantization and buffer regulation are implemented in a TV codec used in *"BBE"* -the Belgian Broadband Experiment-.[5]

Such a codec covers a bitrate range between 2 Mbit/sec (High Quality VideoTelephony) and 40 Mbit/sec (Contribution TV - Standard channel in a compatible approach of HDTV).

DCT calculation is performed in blocks within a frame; this codec uses a hybrid loop without any motion compensation. The main option of this codec is to avoid the cost of a motion compensation by taking care specially over the *"after-DCT"*.

The VLC insered after the quantizer is discribed in [4]. The size of the buffer following the VLC is one Mbyte.

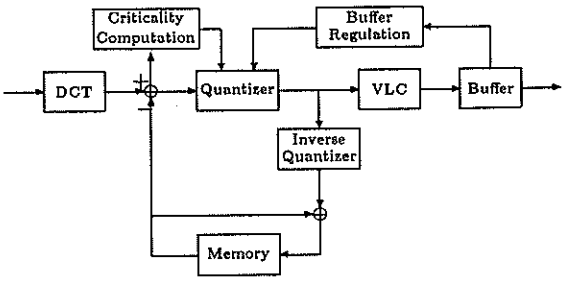


Figure 3 - BBE codec

5 Conclusion

Such a codec is a very good compromise between good quality and low complexity of hardware. Let us recall it is coding pictures in a bitrate range between 2Mbit/sec and 40 Mbit/sec. The higher is the transmission factor, the more adaptive quantization is efficient.

Simulations performed on the "BBE - codec" will be presented at several qualities : videoconference, distribution, contribution.

References

[1] B. Macq "Perceptual Transforms and Universal Entropy Coding for an Integrated Approach to picture coding" *thesis submitted for the Degree of Docteur en Sciences Appliquées - september 1989.*

[2] B. Macq - P. Delogne " Weighted optimum bit allocation to transform coefficients for video codec" *submitted to IEEE transactions on communication.*

[3] J.P. Choffray - P. Delogne - O. Poncin - B. Van Caillie "Bit rate regulation for digital television coding" *presented at PCS 90.*

[4] B. Macq " An universal entropy coder for transform or hybrid coding" *presented at PCS 90.*

[5] S. D'Agostino and all "Universal Video Codec in the Belgium Broadband Experiment" *presented at Packet Video 90.*

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

An Orthogonal Image Transform based on QMF filters

T. George Campbell, Todd R. Reed, and Murat Kunt

Signal Processing Laboratory, Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland

An orthogonal transform based on polar separable QMF filters is introduced. The conditions for perfect reconstruction properties to hold will be discussed. A method for constructing orthogonal pairs of QMF operators is derived. An example of a two band orthogonal fan filter pair is presented.

Introduction

Traditional methods of coding images have taken advantage of many tools from signal processing and information theory. One of the important models of this type is the Quadrature Mirror Filters (QMF) filter bank [1]. This type of system has the property that it can be sub-sampled without aliasing. This property comes from the fact that the unit sample response of the individual subfilters are orthogonal. Recently there has been much interest in visual models in image processing. Much of this work is motivated by resent developments in the study of the mammalian visual system [2] [3] [4].

In this paper we wish to generalize the concept of QMF filter systems to include arbitrary two dimensional orthogonal filter banks with the property of perfect reconstruction, without directly addressing the problem of subsampling. In this context, an orthogonal perfect reconstruction filter bank can be viewed as a projection from the image space onto a set of orthogonal basis functions. The basis functions being the impulse responses of the filters.

Orthogonality is defined such that the inner product of basis functions is zero when the basis functions are not the same:

$$< h_k, h_l >= \sum_{\mathbf{r}} h_k(\mathbf{r}) h_l(\mathbf{r}) = 0, \quad k \neq l \qquad (1)$$

where $\mathbf{r}$ is the region of support of the basis function.

Orthogonality is important because it offers the possibility to reduce the redundancy in the representation of the image (I). With an orthogonal decomposition, the weighting coefficients (c_k) can be found directly by using the inner product:

$$< h_k, I >= \sum_{\mathbf{r}} h_k(\mathbf{r}) I(\mathbf{r}) = c_k. \qquad (2)$$

This also leads to an inverse transform that is very straightforward:

$$< h_k, I > h_k = c_k h_k. \qquad (3)$$

We form orthogonal pairs of filters out of a properly selected prototype filter $H(z)$ with the property that

$$h(0) = \sum_{k=-M}^{M} h(k)^2, \quad (k \neq 0). \qquad (4)$$

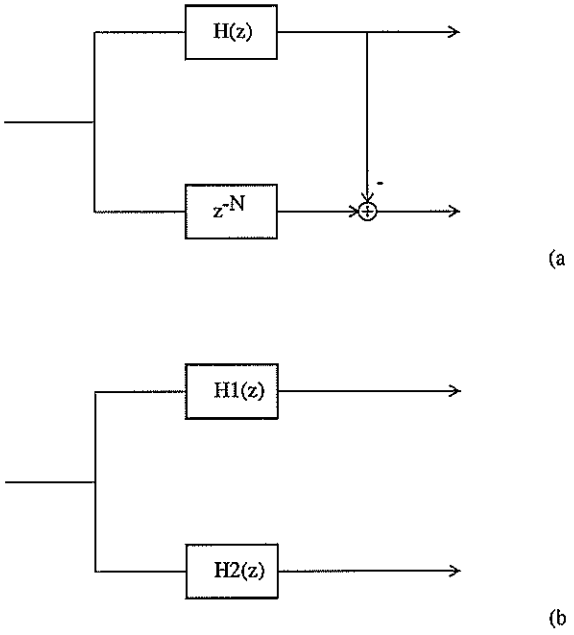


Figure 1. Two equivalent representations of the filter pairs.

The two channels can be seen to be orthogonal and representable as in Figure 1 (a. or equivalently as in 1. (b. Part (a shows that we can obtain perfect reconstruction by summing the outputs.

In [4] fan filters were used for directional decomposition in image coding. Watson [2] used ideal fan filters combined with ideal lowpass filters as a multiscale model of the representation of human vision. In this paper we show the design of a two band fan filter pair which is an orthogonal perfect reconstruction filter bank. This represents a step toward a full orthogonal decomposition as shown in Figure 2. To get a full orthogonal decomposition of this type, we need analytic responses of the filters in the spatial domain. In [5] it is shown that analytic responses can be found for general fan filters. These responses are infinite in extent, but can be approximated by windowed FIR filters. We do this using an analytic approximation of the prolate window function.

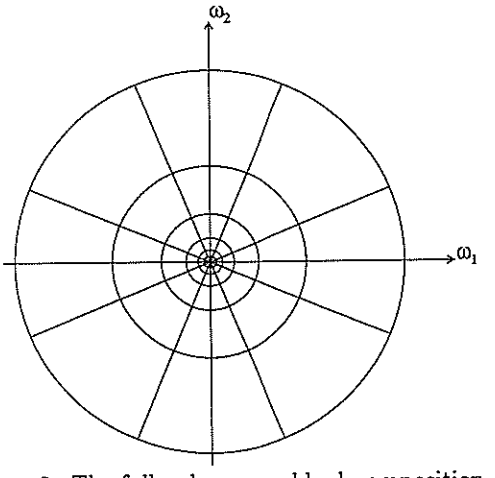


Figure 2. The full polar separable decomposition.

Fan filters

The ideal 2-D fan filter in frequency domain is a division of the frequency plane into wedges starting at the origin. In [5] it is shown that a fan filter,

$$H_0(\omega_1,\omega_2) = \begin{cases} 1, & \text{if } \theta_2 < \tan^{-1}(\frac{\omega_2}{\omega_1}) < \theta_1 \\ 0, & \text{otherwise} \end{cases}, \quad (5)$$

has the impulse response

$$\begin{aligned} &h_0(t_1,t_2) = \\ &u\left(\frac{t_2\cos\theta_2 + t_1\sin\theta_2}{\sin(\theta_1-\theta_2)}\right) u\left(\frac{t_2\cos\theta_1 + t_1\sin\theta_1}{\sin(\theta_1-\theta_2)}\right) \\ &+u\left(\frac{t_2\cos(\theta_2+\pi) + t_1\sin(\theta_2+\pi)}{\sin((\theta_1+\pi)-(\theta_2+\pi))}\right) \\ &u\left(\frac{t_2\cos(\theta_1+\pi) + t_1\sin(\theta_1+\pi)}{\sin((\theta_1+\pi)-(\theta_2+\pi))}\right) \end{aligned}, \quad (6)$$

where u is the inverse Fourier transform,

$$u(t) = 1/t + i\delta(t) \quad (7)$$

of the ideal step function

$$U(\omega) = \begin{cases} 1, & \text{if } \omega > 0 \\ 0, & \text{if } \omega < 0 \end{cases}. \quad (8)$$

The window function

The Saramäki window function [6] is an analytic approxamition of the discrete prolate spheroidal window. This window function was chosen because it offers an approximation of the ideal response with minimum out of band energy for a given spatial extent.

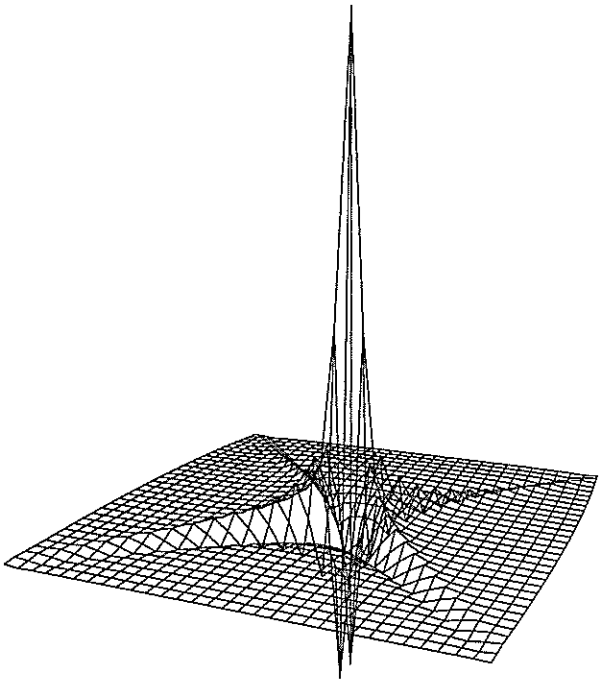


Figure 3. The impulse response of a windowed fan filter.

The windowed filter function is given by

$$h(n_0, n_1) = w(n_0, n_1) h_0(n_0, n_1) \tag{9}$$

where $h_0(n_0, n_1)$ is the discretized response of the ideal filter and

$$w(n_0, n_1) = w(n_0) w(n_1) \tag{10}$$

is an M by M separable window. The coefficients of the window function are given by

$$w(n) = \begin{cases} \widehat{w}(n)/\widehat{w}(0), & -M \le n \le M \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where

$$\widehat{w}(n) = \widehat{w}_0(n) + 2 \sum_{k=1}^{M} \widehat{w}_k(n). \tag{12}$$

The $\widehat{w}_k(n)$'s can be calculated by the recursion relations:

$$\widehat{w}_0(n) = \begin{cases} 1, & n = 0 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

$$\widehat{w}_1(n) = \begin{cases} \gamma - 1, & n = 0 \\ \gamma/2, & |n| = 1 \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

$$\widehat{w}_k(n) = \begin{cases} 2(\gamma - 1)\widehat{w}_{k-1}(n) - \widehat{w}_{k-2}(n) \\ +\gamma[\widehat{w}_{k-1}(n-1) + \widehat{w}_{k-1}(n+1)], & -k \le n \le k \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

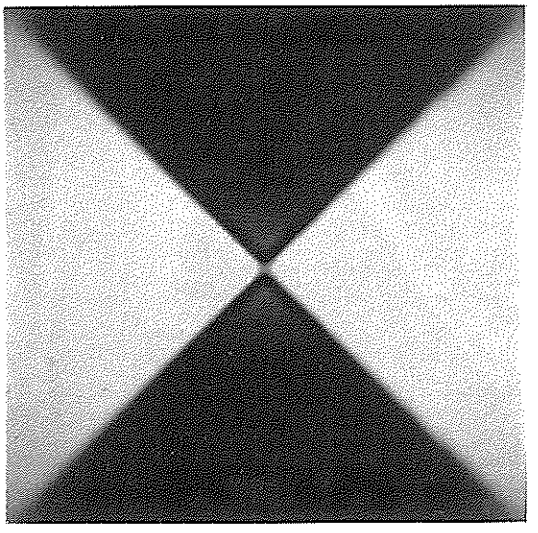


Figure 4. An image of the frequency response of a windowed fan filter.

where

$$\gamma = \frac{1 + \cos\dfrac{2\pi}{2M+1}}{1 + \cos\dfrac{2\beta\pi}{2M+1}}. \tag{16}$$

The parameter β determines the main lobe width of the frequency response. In the following example we use the value of $\beta = 2.56$

Example

The impulse response of a filter constructed by combining (6) and (9) is shown in Figure 3, this is the vertical filter with $\theta_1 = \pi/4$ and $\theta_2 = -\pi/4$. The frequency response (as an image) is shown in Figure 4. Figure 6 shows and image filtered with the horizontal filter. Figure 7 shows the same image filtered with the vertical filter. Though in this case the vertical filter was formed from (6) and (9), with $\theta_1 = 3\pi/4$ and $\theta_2 = \pi/4$ it is exactly equivalent to the application of (4) to the horizontal filter. The original image and the image reconstructed from the two subimages are shown in Figure 5 and Figure 8 respectively. The filtering was done using floating point arithmetic on a Silicon Graphics 4D25 with 16 MIPS processor and took about 2.5 minutes. The reconstruction error after quantization is zero.

Figure 5. Original Leena image.

Figure 8. Reconstructed Leena.

Figure 6. Leena filtered with the horizontal filter.

Figure 7. Leena filtered with the vertical filter.

Acknowledgment

We would like to thank Tapio Saramäki and Ferran Marqués for useful discussions during the preparation of this manuscript.

References

[1] J. W. Woods and S. D. O'Neil, "Subband coding of images," IEEE Trans. Acoust. Speech Signal Process. ASSP-43, 1278-1288 (1986).

[2] A. B. Watson, "Efficiency of a model human image code," Journal of the Optical Society of America, Vol. 4, No. 12, Dec. 1987.

[3] M. Porat and Y. Y. Zeevi, "The Generalized Gabor Scheme of Image Representation in Biological and Machine Vision," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 10, No. 4, July, 1988.

[4] M Kunt, A. Ikonomopoulus, and M. Kocher, "Second-generation Image-coding Techniques," Proceedings of the IEEE, Vol. 73, No. 4, April 1985.

[5] T. G. Campbell and A. M. Geurtz, *Analytic Formulation of Fan Filters in the Spatial Domain*, in Proceedings of the Latvian Signal Processing International Conference, Riga, Latvia, April 23-27, 1990.

[6] T. Saramäki, *A Class of Window Functions with Nearly Minimum Sidelobe Energy for Designing FIR Filters*, 1989 IEEE International Symposium on Circuits and Systems, Portland, May 8-11, 1989.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Digital Transmission of Component Coded HDTV Signals using the Discrete Cosine Transform: DESIGN OF A VISIBILITY THRESHOLD MATRIX *)

J.Oest, F.J.Guirao

Retevisión
Subdir. de Innovación Tecnológica
Lab. de Proceso de Imágenes
28002 MADRID - SPAIN

N.García Santos

Gr.de Tratamiento de Imágenes
E.T.S. Ing.de Telecomunicación
Univ. Politécnica de Madrid
E-28040 Madrid, SPAIN

It is becoming very popular to take into account the frecuency response of the human observer when considering transform coding schemes for images. In a coding system based on the Discrete Cosine Transform (DCT), this is achieved by a multiplicative visibility threshold matrix (V.T.M.) applied to the transform coefficients in order to code more accurately those spatial components to which the eye is more sensitive.
This document presents an analysis of the computation of the luminance and crominance visibility threshold matrices for HDTV images giving two families of tables based on the following approaches:
- A theoretical one, based on previous works (luminanance) ([1],[2],[3])
- A practical one, based on subjective assessment tests (lum & cro).

1 V.T.M. BY A THEORETICAL APPROACH

1.1 INTRODUCTION

Different studies ([1],[2],[3],[4]) have demonstrated that the quantization noise of each transform coefficient does not equally affect the subjective quality of the picture. Improved coding systems will result if the visual response of the observer is taken into account. Therefore, for a fixed output bit-rate, each coefficient has to be coded as accurately as the human visual system is sensitive to its distortion. The minimum amplitude of one coefficient that makes the corresponding basis function visible will therefore be its visibility threshold.

Hall ([1], [2]) took these considerations into account and introduced an optical spatial response curve measured by Mannos and Sakrinson ([3]), to weight the *FOURIER* transform coefficients prior to quantization, bit allocation and coding.

Clarke ([4]) demonstrated afterwords, that an appropriate weighting could also be applied to the *DCT* coefficients in the manner reported by Hall using a sufficiently slowly varying weighting function of spatial frequency in order to allow improved coding efficiency.

*) This work has been done within EUREKA-256: "Bit-Rate Reduction system for HDTV Digital Transmission"

1.2 METHODOLOGY

Based on these results, the method that has been used in this document in order to obtain the weighting factors associated to each one of the DCT coefficients of an 8x8 image block is represented in FIGURE 1:

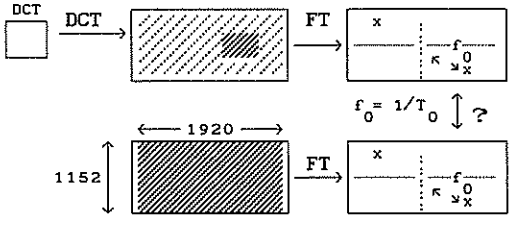


FIG.1

For each coefficient of the block we will:
A) obtain its inverse DCT
B) expand the result to the whole screen (1920 pixels * 1152 lines)
C) obtain the Fourier transform of the whole screen
D) associate a weighting factor to the resulting frequencies by comparison with the frequencies corresponding to the FT of an image with the same characteristics as the one obtained by expanding the DCT^{-1} of the coefficient (FIG.1). This will be the weighting factor to be applied to the coefficient.

Steps A), B) and C) can be performed in an easy way considering the general formulation of the Inverse Cosine Transform and the Discrete Fourier Transform. Taking into account the viewing distance between the observer and the screen (for HDTV D=3H, where D is the viewing distance and H is screen height), and the aspect rate of the screen (16/9), the module of the absolute spatial frequency in $(^{c}/_{o})$ corresponding to a coefficient $X(k,l)$ (obtained from the DCT of an 8x8 image block) will be:

$$f_{c/o} = \frac{2\pi}{16}\sqrt{(4,8k)^2 + (9l)^2} \qquad D=3H$$

If we directly applied the optical response weighting function obtained by Mannos and Sakrinson [3] to the DCT coefficient of an 8x8 image block we would associate small weighting factors to the low frequency coefficients. This would result in a greater quantization error which will not be perceptible on the block, but over the whole screen as a block effect.
In this way, we can't use the original visual response curve, but a modification of it which smooths the curve in the low frequencies, in order to consider higher weighting factors for the low order coefficients of the 8x8 image block, and eliminate in this way the block effect [6].

1.3 RESULTS

Table (1) shows the frequencies and weighting factors associated to each DCT coefficient of a 8x8 block for D=3H (these factors have to multiply the DCT coefficients and will be considered only for the *luminance* component).

$$W = 2,6\ (0,0192 + 0,114f)\ \exp\left(-(0,114f)^{1,1}\right)$$

k / l	0	1	2	3	4	5	6	7
0	1	1	1	0.93	0.78	0.61	0.45	0.32
1	1	1	0.98	0.93	0.78	0.61	0.45	0.32
2	1	1	0.98	0.91	0.76	0.59	0.44	0.32
3	1	1	0.97	0.88	0.73	0.57	0.42	0.30
4	0.98	0.98	0.94	0.84	0.69	0.54	0.40	0.29
5	0.96	0.95	0.89	0.78	0.64	0.50	0.38	0.27
6	0.91	0.89	0.82	0.72	0.59	0.46	0.35	0.25
7	0.83	0.80	0.74	0.65	0.53	0.42	0.32	0.23

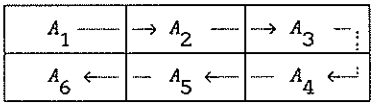
TABLE 1 D=3H

2 VTM's BASED ON SUBJECTIVE ASSESSMENT

2.1 VTM / QUANTIZATION NOISE

As we mentioned earlier, the different coefficients will be quantized with different precision, depending on the visibility threshold.
This quantization will produce a noise with mean square error $e=a^2/12$ where a is the quantization step, supposing an uniform quantizer. So we generated an error function that simulated the value of the quantization noise of each coefficient and considered this function as the coefficient value in order to apply the DCT^{-1} over it. Observing when the generated functions were visible with different values of the amplitude for each coefficient, a valid visibility threshold matrix was obtained [6].

Six squares were presented on the screen (FIG.2) and in each square a different multiple of the image function of one coefficient was included, increasing from the left upper corner to the right side, then down and left, finishing in the left bottom corner.



$A_1 < A_2 < A_3 < A_4 < A_5 < A_6$

FIG. 2

The viewers were asked to mark how many of these squares they could see anything in. In this way, we could deduce which was the minor value of the amplitude each person could see. For example, if a person marked 4 squares, the minor value of the amplitude would be A_3 and A_3 was noted as the visibility threshold seen by this person for the coefficient analyzed. Doing the same process for each one of the luminance and crominance coefficients, the luminance and crominance visibility threshold matrices were obtained and are shown in tables 2 and 3.

3. CONCLUSIONS

Comparing the two matrices (TABLES 1,2 & ANNEX 1) obtained in this paper we can observe that the practical matrix has a much sharper shape than the theoretical one, which probably stems from the fact that the theoretical study has been based on an approach of the spatial response of the human visual system.

k l	0	1	2	3	4	5	6	7
0	1	1	*0.89*	0.37	*0.24*	0.13	*0.09*	0.07
1	1	1	0.65	0.31	0.20	0.09	0.06	0.04
2	*0.89*	0.65	*0.52*	0.25	*0.17*	0.07	*0.05*	0.03
3	0.70	0.51	0.40	0.21	0.14	0.06	0.04	0.03
4	*0.57*	0.41	*0.33*	0.18	*0.12*	0.05	*0.04*	0.03
5	0.48	0.33	0.25	0.13	0.09	0.05	0.03	0.02
6	*0.41*	0.27	*0.20*	0.10	*0.07*	0.04	*0.03*	0.02
7	0.37	0.23	0.18	0.08	0.05	0.04	0.03	0.02

TABLE 2 D=3H LUMINANCE

In order to make a comparative evaluation of the efficiency of both matrices, subjective assessment tests will be done on different HDTV sequences, using the two developed matrices and the matrices recommended by CMTT/2 for 4:2:2 sequences (Annexe 3 of CMTT/2-66), as no HDTV matrix has yet been recommended. The shape of these matrices is shown in Annexe 1.

The following are the features considered for the subjective assessment tests:
* The output bitrates will be 136 Mbit/s (18% of the original bitrate), 60 Mbit/s (9% of the original bitrate) and 40 Mbit/s (6% of the original birate)
* The double stimulus impairment scale method as described in CCIR rec. 500-3 will be used.
*The viewing distance will be D=3H.
* A grey level of 128 (scale between 1-255) will be considered as the mean value of the luminance in an image and will be therefore used as the continuos level in the second way of obtaining the visibility threshold matrix. In this way, the matrix obtained can be assumed as a valid one if the eye is considered as a lineal system.

k l	0	1	2	3	4	5	6	7
0	1	1	0.77	0.71	0.65	0.50	0.50	0.39
1	1	0.92	0.84	0.77	0.60	0.50	0.46	0.32
2	0.84	0.84	0.77	0.71	0.55	0.46	0.42	0.25
3	0.77	0.71	0.65	0.65	0.50	0.42	0.35	0.25
4	0.65	0.60	0.60	0.55	0.46	0.39	0.32	0.23
5	0.60	0.55	0.46	0.46	0.39	0.30	0.25	0.16
6	0.42	0.39	0.39	0.35	0.30	0.25	0.20	0.13
7	0.35	0.35	0.35	0.35	0.23	0.21	0.17	0.11

TABLE 3 D=3H CROMINANCE

REFERENCES

- [1] Hall,C.F. and Andrews,H.C. : "Digital color image compression in a perceptual space" Proc. SPIE 1978, 149, pp 182-188

- [2] Hall,C.F. : "Perceptual coding in the Fourier transform domain", NTC convention record, 1980, pp 36.1.1-36.1.7

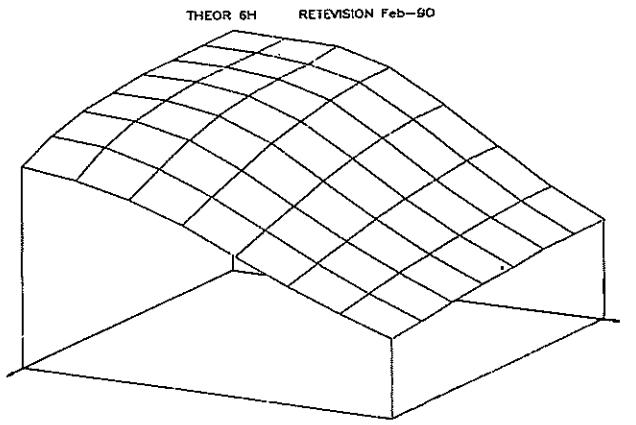
- [3] Mannos,J. and Sakrinson,D. : "The effects of a visual fidelity criterion on the encoding of images", IEEE Trans. 1974, IT-20, pp 525-236

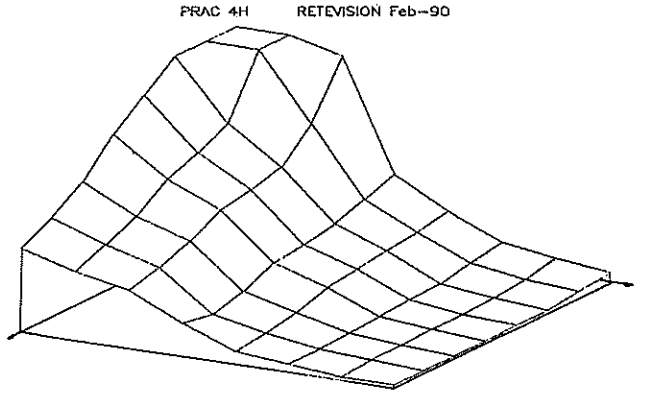
- [4] Clarke, R.J. and Tech,B. : "Spectral response of the Discrete Cosine and Walsh Hadamard transforms", IEEE Proc., Vol 130, N.4, June 1983, pp 309-313

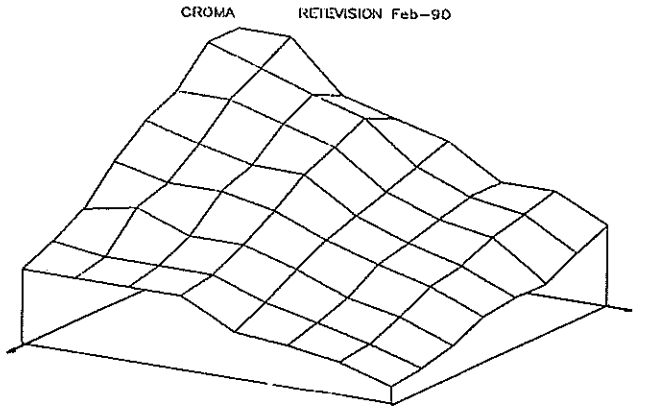
- [5] M.J. Knee and N.D. Wells, BBC: "Designing Quantizers for Transform Coding", CMTT/2 - DCT - 25

- [6] F.J. Guirao and J.A. Oest, RTVE: "Optimum Visibility Threshold Matrix for D.C.T. Coefficients", 3rd International Workshop on HDTV, Volume III

ANNEX 1


THEOR 6H
RETEVISION Feb-90


PRAC 4H
RETEVISION Feb-90


CROMA
RETEVISION Feb-90

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Digital Transmission of Component Coded HDTV Signals using the Discrete Cosine Transform: "REDUCED NUMBER OF CODING MODES" *)

F.J. Guirao, J.A. Oest
Lab. de Proceso de Imágenes
Retevisión
P. de la Castellana 83/85
28046 Madrid - SPAIN

N. García.
Gr. de Tratamiento de Imágenes
E.T.S.I. de Telecomunicación
Univ. Politécnica d Madrid
28040 Madrid - SPAIN

Hybrid transform coding schemes exploit the temporal redundancy as well as the spatial redundancy in video signals by applying a prediction along the temporal axis and a transform encoding along the spatial axis on a block by block basis .Depending on the image contents, three different coding modes are used: **intra-field**, **inter-field** or **inter-frame** in order to achieve the minimum output bit-rate using a Variable Length Coding.
The information of each 8x8 pixel block mode must be sent but as blocks are grouped in quadblocks (2 luminance and 2 crominance blocks) corresponding to the same area of pixels, an optimization over the number of coding modes within a quadblock can be made.

1. INTRODUCTION

Hybrid transform coding schemes exploit the temporal redundancy as well as the spatial redundancy in video signals by applying a prediction along the temporal axis and a transform encoding along the spatial axis on a block by block basis .If the differences between the pixels and the corresponding prediction values are very little, the transform coefficients obtained by coding these differences will generally have little amplitude, suitable to be coded using VLC words.

Three different coding modes can be considered:
- in the case of static areas, the optimum prediction value is that of the co-sited pixel belonging to the previous frame (**inter-frame** mode)
- if slight movements are present, the optimum prediction value can be obtained from the previous field, because the temporal distance is smaller (**inter-field** mode)
- in other cases no prediction is made and the transform is applied directly over the pixels (**intra-frame** mode).

The optimum choice within the three coding modes (intra-field, inter-field or inter-frame) should be determined "a posteriori" for each 8x8 block, on the basis of the minimum number of bits necessary to code the block itself. Such kind of choice, however, implies the transforming, quantization and coding of three blocks, one for each mode. In order to reduce the hardware complexity, a sub-optimum choice is implemented, it is an "a priori" choice, made before the block transform. It has been verified that a choice based on the minimum energy of the block is very similar to the "a posteriori" choice. Energy is computed from the values of pixels or from those obtained as differences with prediction values, but disregarding the energy linked with the DC component of the block.

Because of the different sampling ratio of luminance or crominance, blocks are grouped in quadblocks composed of 2 contiguous luminance blocks and 2 blocks for the corresponding crominance components C_r and C_b, being the total information of a 16x8 pixel luminance area.

The mode of every four blocks corresponding to a quadblock will be coded as a single word in order to be transmitted, because they correspond to the same area of pixels. As each block can be processed in 3 different ways, $3^4 = 91$ words would be necessary to code all the possibilities within a quadblock, which implies 7 bit words. The difficulty to implement such a large number of coding modes as well as the bit-rate increment produced by these bits, makes necessary to consider a reduced set of processing modes which should however not produce a loss of quality in fixed output bit-rate systems, taking into account that the four blocks in a quadblock are highly correlated.

*) This work has been done within EUREKA-256: "Bit-Rate Reduction system for HDTV Digital Transmission".

2. SIMULATIONS

To optimize the number of coding modes, computer simulations were made of the Encoder-Decoder loop, following different mode selection strategies. Maintaining a constant quality, the strategy which produces a lower output bit-rate can be considered as the optimum one.

Two HDTV sequences produced by RAI were used: CACTUS HD and RENATA HD. These sequences had 1440 pixels/line and 1152 active lines, aspect ratio 16/9.

2.1. Initial conditions.

A program produced by RAI was used as a simulator of the codec. The mode selection strategy was based on 4 modes (independent mode selection of each block in the quadblock), and was taken as reference.

Seven bits per quadblock are required to codify the mode selection. In the used HDTV sequences there are (1440/8 x 1152/8)/2 quad-blocks per frame, and (720/8 x 576/8)/2x 25 quadblock per second, this is, 324,000 quadblock per second. Then, it is necessary adding 324,000 x 7 = 2,268,000 bits per second to codify the mode selection.

2.2. Tests.

The new mode selection strategies were:

- *3 modes.* The modes of each block of luminance was chosen independently and for the two blocks of crominance only one mode was assigned. This mode was obtained on the basis of the minimum energy of the two crominance block addition. 3^3 = 27 different words are needed to codify the different mode selections, that is 5 bits per quadblock and 324,000 x 5 = 1,620,000 bits per second.

- *2 modes.* Only a mode is chosen for the two blocks of luminance, also based in the minimum energy of both blocks. For the two blocks of crominance is chosen another mode. Now we need 3^2 = 9 different words to codify the different mode selections, that is 4 bits per quadblock and 324,000 x 4 = 1,296,000 bits per second.

- *1 mode.* Only one mode is chosen for each quadblock. This mode is chosen on the basis of minimum energy for the two blocks of luminance and the same mode is assigned for the crominance. Now we need 3 different words to codify the different mode selections, that is 2 bits per quadblock and 324,000 x 2 = 648,000 bits per second.

Routines following these strategies were developed and test with 10 frames of the sequences **Renata HD** and **Cactus HD** were made. A fixed Transmission Factor was used, in order to compare the different output bit-rates with the different modes selection strategies. This transmission factor will determine the quantification accuracy, then a fixed quality can be considered in all cases.

2.3. Results

The results are shown in the table I , where we can see the results of the 4 strategies for each Transmission Factor with the sequence **Renata HD**, and in table II, where the results of **Cactus HD** are shown. A percentage of the bit-rate increment over the 4 modes strategy is also written. If this increment is negative, it actually indicates a decrement of the output bit-rate.

Graphs I and II shows the relative output bit-rate (in percentage) for the different mode selection strategies including the necessary bits to code the mode selection for sequences **Cactus HD** (gaph I) and **Renata HD** (graph II).

3. Conclusions

- The number of mode coding bits has a great influence on the magnitude of the output bit-rate.

- For bit-rates higher than 80 Mbit/s, the strategy using only one mode for both crominance components and two modes for luminance performs clearly better than other methods. Contribution codecs work in this bit-rate range.

- For bit-rates lower than 80 Mbit/s, the strategy using only one mode per quadblock is better. Distribution codecs could work in this range.

REFERENCES

- [1] IWP CMTT/2:" Draft new Recommendation AT/CMTT: Transmission of Component-Coded Digital Television Signals for Contribution-Quality Applications at the Third Hierarchical Level of CCITT. October 1989.

TABLE I

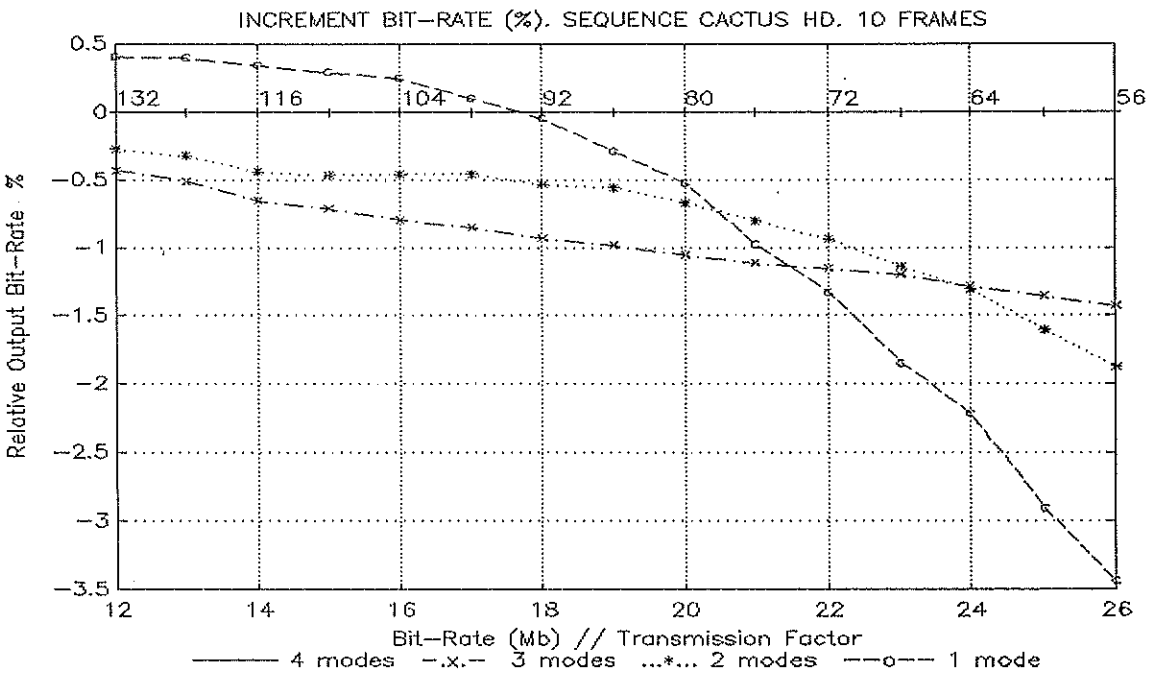
Mode Selection	Bit Rate Total + bits cod	Δ%	Bit Rate Total + bits cod	Δ%	Bit Rate Total + bits cod	Δ%	Bit Rate Total + bits cod	Δ%
Transmi. Factor	15		17		19		21	
4 modes	141.588	0	125.632	0	111.124	0	98.208	0
3 modes	141.204	-.271	125.144	-.388	110.548	-.517	97.572	-.648
2 modes	141.400	-.133	125.344	-.230	110.748	-.337	97.752	-.463
1 mode	141.764	0.125	125.600	-.026	110.900	-.201	97.784	-.430
Transmi. Factor	23		25		27		29	
4 modes	86.700	0	76.564	0	67.704	0	59.896	0
3 modes	86.016	-.791	75.800	-.999	66.924	-1.15	59.100	-1.33
2 modes	86.240	-.531	75.748	-.730	67.116	-.869	59.244	-1.09
1 mode	86.124	-.662	75.748	-1.06	66.640	-1.57	58.592	-2.18

Output bit-rates and relative (%) bit-rate over 4 mode selection, for different Transmission Factors. Sequence **RENATA HD.**

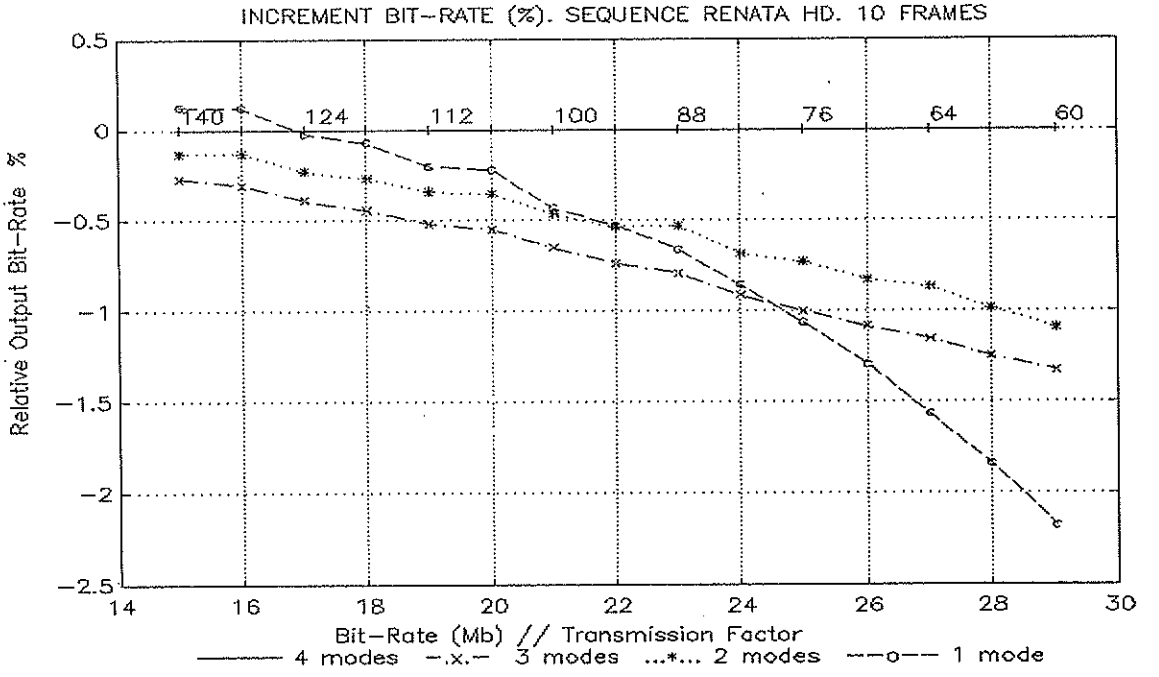
TABLE II

Mode Selection	Bit Rate Total + bits cod	Δ%	Bit Rate Total + bits cod	Δ%	Bit Rate Total + bits cod	Δ%	Bit Rate Total + bits cod	Δ%
Transmi. Factor	12		14		16		18	
4 modes	132.072	0	117.124	0	103.316	0	91.540	0
3 modes	131.504	-.431	116.364	-.650	102.496	-.793	90.688	-.929
2 modes	131.712	-.271	116.612	-.438	102.840	-.462	91.048	-.536
1 mode	132.608	0.404	117.528	0.344	103.572	0.247	91.496	-.047
Transmi. Factor	20		22		24		26	
4 modes	80.436	0	71.516	0	63.532	0	57.160	0
3 modes	79.588	-1.05	70.692	-1.15	62.716	-1.29	56.344	-1.43
2 modes	79.904	-.663	70.848	-.936	62.700	-1.31	56.088	-1.88
1 mode	80.020	-.517	70.564	-1.33	62.124	-2.22	55.196	-3.44

Output bit-rates and relative (%) bit-rate over 4 mode selection, for different Transmission Factors. Sequence **CACTUS HD.**



GRAPH 1



GRAPH 2

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

DCT Domain Modelization of the TV Signal for Quantization*

José I. Ronda, Fernando Jaureguizar, and Narciso García

Grupo de Tratamiento de Imágenes, E.T.S. Ingenieros Telecomunicación
Universidad Politécnica de Madrid, E-28040 Madrid, Spain

The lack of reliable statistical models of the TV signal makes the simulation the only tool for the design and optimization of bit rate compression systems for its digital transmission. This paper studies the possibilities of such statistical models in case of Hybrid DCT TV encoding. The DCT domain energy distribution is studied as well as several approaches to the modelization of the probability density function of each coefficient. As a test for the resulting models, the problem of the optimum Max-Lloyd quantizer design is analyzed.

1. Introduction

The design of bit rate reduction systems for the digital transmission of signals requires a knowledge of the statistical properties of the signal to be encoded. The optimal knowledge consists of a complete statistical characterization of the signal as a stochastic process, thus including its n-th order probability density functions [1]. Under the stationary assumption, a usually satisfying characterization includes only the first order probability density distribution (pdf) and the second order moments of the process (autocorrelation function).

The minimal set of statistical data needed in order to carry out a particular design strongly depends on the encoding technique itself. Transformation schemes often strength the interest on the characterization of the transformed process, namely its first order probability distribution and second order moments. Besides, if the transformation is close to optimal, a further assumption is made that considers incorrelated transform coefficients, thus reducing the knowledge of the autocorrelation function in the transformed domain to that of the energy of the coefficients.

In the case of the TV signal, the need for reliable statistical models of the source stems as the simulation of encoding systems is a highly demanding computational task, both in CPU consume and mass storage capacity. Although simulation can never be completely avoided, an *a priori* statistical characterization can save efforts, mainly in the first stage of the design.

As TV signal generation is highly complex, the only practical approach to the modelization is the empirical estimation of its more relevant parameters. The estimation of the coefficient pdf is achieved through the computation of its histograms; afterwards, they are approximated to classical distributions. the approximation can only be tested *a posteriori*, its performance when applied to the academical problem of the optimum Max-Lloyd quantizer design is proposed as a test.

2. Hybrid DCT Encoding

The efficiency of DCT transform coding techniques for both static and moving image compression is well known [1]. VLSI technology developments in the last years allow for their application on a wide range of bit-rate reduction systems for digital video transmission. Consequently, many of the currently in progress international standardization efforts in this field are considering compression schemes which are based on DCT transform coding [2,3].

The hybrid predictive-DCT scheme which has been used for the modelization is the one followed by CMTT/2 [4,5] for contribution quality TV digital transmission at bit-rates near 30 Mbit/s. This uses prediction along the temporal axis and 2-D DCT transform along the spatial ones, being 8×8 the block size. Three different temporal predictions are considered: *interfield*, *interframe* and *interframe with motion compensation*. After predictions are computed, the corresponding prediction errors are compared to each other and with the block itself (mode *intrafield*), and the one with less energy is chosen for transmission.

The selected block undergoes a transform encoding procedure in order to reduce its spatial redundancy. First it is DCT-transformed; then the resulting coefficients are scaled according both to subjective visibility criteria and to the value of the transmission factor (instantaneous picture quality control parameter), after which they are quantized, coded by means of a VLC, and sent to a buffer for transmission. Meanwhile, the quantized block is reconstructed in the same way the decoder would do and stored in a memory which contains the two previous decoded fields for prediction computing. The need for an output buffer stems as a variable bit rate results from the encoding operation, while a fixed one is accepted by the transmission line. The control of the buffer takes place by means of the transmission factor, which affects the scaling, and is varied according to the buffer occupancy.

This work has been done within Eureka-256: "Bit-Rate Reduction System for HDTV Digital Transmission". It has been partially supported by the Plan Electrónico e Informático Nacional and the Comisión Interministerial de Ciencia y Tecnología of the Spanish Government.

Although the general architecture of the system is well defined, several critical parameters, namely the quantization of the DCT coefficients, the variable length code (VLC) and the buffer control require a careful study, specially considering the extensibility to HDTV and distribution quality TV transmission. For these tasks the modelization of the signal can play a substantial role.

3. Modeling

Since the TV signal can only be considered stationary when observed for a relatively long time, reliable estimation of its parameters can only be obtained by means of the processing of significantly long sequences with different features from one another.

This empirical approach to the modelization has to overcome a practical difficulty: The huge amount of data associated to a sequence of digital TV makes it prohibitive the simulation with sequences longer than a few seconds. For example, the CPU time required for a simulator as the one described in [6] to process a two second sequence is four hours in a 4 MIPS general purpose computer, not considering motion compensated prediction. And the representativeness of the sequence has to be *a posteriori* tested, according to the same criterion: The performance of the resulting system when applied to other sequences.

In this modelization essay three test TV sequences have been processed: *Calendar*, (200 fields), *Renata* (148 fields) and *Renata and Butterfly* (198 fields). All the sequences hold to European format.

3.1. DCT domain energy distribution

Figures 1 and 2 show the observed energy distribution of the DCT coefficients. As a measure of the energy compaction property of the DCT in each case (prediction mode and type) it is worthwhile noting the value of the *gain over PCM*[1], computed as

$$G = \frac{\sum_{k=0}^{7}\sum_{l=0}^{7}\sigma_{kl}^2/64}{[\prod_{k=0}^{7}\prod_{l=0}^{7}\sigma_{kl}^2]^{1/64}}$$

where σ_{kl}^2 is the variance of the coefficient with vertical index k and horizontal index l. Its values for the average of the processed TV sequences is shown in Fig. 3. As it can be seen from this figure, *the gain is substantial in the intrafield mode while very poor in the inter modes.*

Results concerning the energy figures for each sequence are significantly different from one another, which makes consider further simulations necessary, in the search for a reliable long-term energy distribution. Nevertheless, some general qualitative aspects can be established:

- Intrafield mode is characterized by a decreasing energy distribution.
- Inter modes show high energy components in the coefficients with one of the indices equal to 0, being particularly apparent the increasing of energy for $l = 0$ and $k > 1$. This phenomenon, repeated for all the test sequences does not have any evident explanation.
- The DCT shows a much better performance in energy compaction when applied to the intrafield mode than when applied to the inter ones. This evidences a low correlation between elements in the error prediction blocks.

	0	1	2	3	4	5	6	7
0	168086.	1203.	680.	434.	245.	130.	77.	36.
1	976.	337.	241.	162.	93.	59.	39.	16.
2	581.	257.	198.	131.	74.	51.	33.	13.
3	420.	192.	135.	89.	54.	41.	28.	11.
4	391.	175.	106.	66.	42.	33.	22.	9.
5	373.	152.	89.	52.	33.	24.	18.	8.
6	430.	161.	93.	45.	26.	19.	14.	6.
7	488.	133.	68.	32.	20.	16.	12.	5.

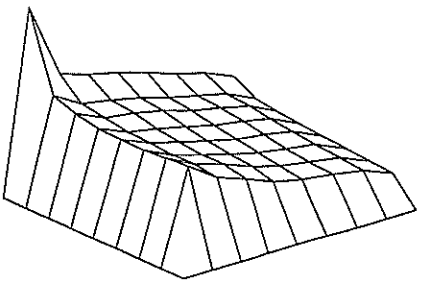


Fig. 1 – DCT domain energy distribution and its logarithmic representation. Luminance, mode intrafield. Average results.

	0	1	2	3	4	5	6	7
0	236.	161.	161.	178.	168.	157.	129.	66.
1	142.	114.	115.	118.	109.	104.	85.	45.
2	157.	117.	111.	108.	97.	93.	77.	41.
3	195.	126.	111.	99.	85.	80.	68.	36.
4	249.	147.	114.	92.	77.	73.	61.	32.
5	287.	160.	113.	86.	72.	66.	55.	28.
6	328.	174.	115.	82.	65.	58.	49.	26.
7	343.	176.	115.	80.	64.	59.	48.	24.

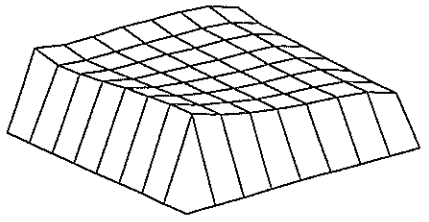


Fig. 2 – DCT domain energy distribution and its logarithmic representation. Luminance, mode interframe (motion compensated). Average results.

Type	Intrafield	Interfield	Interframe
Luminance	15.30	1.19	0.70
Chrominance	18.90	0.43	0.28

Fig. 3 – Gain over PCM (dB) for TV sequences. Average results.

3.2. Probability density function estimation

The best estimation of the fdp of a discrete random variable is its relative frequency histogram. According to it, as the DCT coefficients can be regarded as discrete (if, after their exact computation, rounded with the required precision), the task of estimating their distribution is equivalent to that of the computation of their histogram. So, histograms have been computed, separately for each coefficient order, mode, type and test sequence (see Fig. 4).

In order to present the results in a more compact format, they have been approximated by means of two classical parametrical distributions: the *generalized gamma* and the *generalized gaussian* distributions [7,8]. The expression for the first one is

$$p_G(r,x) = Ax^{r-1}e^{-\alpha x} \quad (0 < r \leq 1) \qquad (1)$$

where r is the parameter which defines the distribution and α and A are adjusted as a function of r so that the subtended area is unitary and the variable presents an a priori stated variance. The corresponding expression for the generalized gaussian is

$$p_N(r,x) = Ae^{-\alpha|x|^r} \quad (0 < r) \qquad (2)$$

where r, A and α play parallel roles to those of their homonymous in the previous formula. Observe that equation (2) for $r = 2$ corresponds to the gaussian distribution and that both equations (1) and (2) equal the laplacian one for $r = 1$

For each considered random variable, the closest gamma and gauss distributions with the same variance have been chosen. The range of parameter searched has been $(0, 1]$ for the gamma pdf and $(0, 2]$ for the gaussian one. Two different approaches have been considered for the best fit adjustment:

- **Best χ^2 test fit:** The value of the χ^2 parameter is computed for each considered distribution and the one which gives a minimum is selected.
- **Minimum square error (MSE) fit:** The distribution which results closest to the empirical one, in the average square error sense, is selected.

Some results are given in Figs. 5-8. It must be admitted that the χ^2 test is clearly *not accomplished* by any variable for any distribution in the considered range of parameters. So, no conclusion can be stated of the form 'coefficient (k, l) *matches* laplacian distribution', since the corresponding statistical hypothesis results rejected with significance level above 99.99%. However, this does not mean that the obtained curves are useless –on the contrary, they can be used for the first order estimation of certain parameters of the corresponding random variables, though the validity of the results has to be necessarily a posteriori verified.

The separate consideration, in the first stage, of each sequence gives some intuitive hints about the representativeness of the resulting data, at least in the negative sense: If the resulting profiles are not assimilable, nothing definitive can be concluded starting from the available data. According to this, conclusions are only given when apply for all the tested sequences:

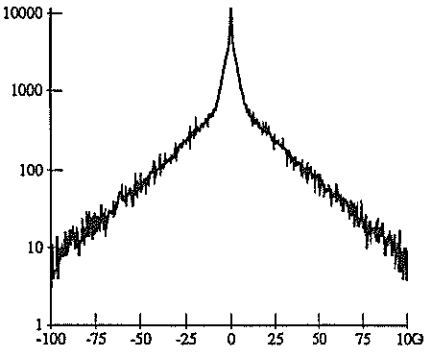


Fig. 4 – Histogram of the DCT coefficient (1,1), Intrafield mode, luminance. Sequence *Renata*.

	0	1	2	3	4	5	6	7
0	0.65	0.21	0.21	0.23	0.25	0.29	0.31	0.36
1	0.20	0.22	0.23	0.24	0.27	0.29	0.32	0.38
2	0.21	0.22	0.24	0.25	0.28	0.30	0.33	0.39
3	0.23	0.24	0.26	0.27	0.29	0.31	0.34	0.41
4	0.22	0.24	0.26	0.28	0.30	0.32	0.35	0.42
5	0.21	0.24	0.26	0.28	0.31	0.33	0.36	0.43
6	0.21	0.23	0.26	0.29	0.32	0.34	0.37	0.44
7	0.21	0.24	0.27	0.30	0.33	0.35	0.38	0.45

Fig. 5 – Best gaussian adjustment for the luminance intrafield coefficients, according to the MSE criterion. Values of the distribution parameter r.

	0	1	2	3	4	5	6	7
0	0.61	0.28	0.27	0.30	0.31	0.37	0.39	0.43
1	0.26	0.27	0.29	0.30	0.33	0.36	0.38	0.44
2	0.27	0.28	0.30	0.31	0.34	0.36	0.39	0.45
3	0.29	0.30	0.32	0.33	0.36	0.38	0.40	0.47
4	0.29	0.29	0.32	0.34	0.37	0.38	0.41	0.48
5	0.27	0.29	0.32	0.34	0.37	0.39	0.42	0.48
6	0.27	0.29	0.31	0.34	0.37	0.40	0.43	0.49
7	0.26	0.29	0.32	0.36	0.39	0.41	0.44	0.50

Fig. 6 – Best gamma adjustment for the luminance intrafield coefficients, according to the MSE criterion. Values of the distribution parameter r.

	0	1	2	3	4	5	6	7
0	0.69	0.72	0.76	0.78	0.85	0.90	0.93	1.00
1	0.71	0.79	0.82	0.87	0.94	0.98	0.99	0.97
2	0.69	0.79	0.84	0.89	0.96	0.99	1.00	0.95
3	0.66	0.76	0.83	0.89	0.97	1.00	1.00	0.95
4	0.64	0.74	0.81	0.88	0.95	0.99	0.98	0.94
5	0.61	0.71	0.77	0.84	0.90	0.93	0.94	0.93
6	0.61	0.69	0.75	0.81	0.86	0.89	0.90	0.90
7	0.60	0.68	0.73	0.78	0.81	0.82	0.85	0.86

Fig. 7 – Best gaussian adjustment for the luminance interframe coefficients, according to the MSE criterion. Values of the distribution parameter r.

	0	1	2	3	4	5	6	7
0	0.84	0.86	0.88	0.90	0.94	0.97	0.99	0.99
1	0.85	0.90	0.92	0.95	0.99	0.99	0.99	0.99
2	0.84	0.90	0.93	0.96	0.99	0.99	0.99	0.99
3	0.82	0.88	0.93	0.97	0.99	0.99	0.99	0.99
4	0.80	0.87	0.91	0.95	0.99	0.99	0.99	0.99
5	0.78	0.85	0.89	0.93	0.96	0.98	0.99	0.98
6	0.78	0.84	0.88	0.91	0.94	0.96	0.96	0.97
7	0.77	0.83	0.86	0.89	0.91	0.92	0.93	0.94

Fig. 8 – Best gamma adjustment for the luminance interframe coefficients, according to the MSE criterion. Values of the distribution parameter r.

	0	1	2	3	4	5	6	7
0	0.1	0.1	0.2	0.2	0.3	0.4	0.3	0.4
1	0.2	0.3	0.3	0.3	0.5	0.4	0.5	0.9
2	0.2	0.3	0.5	0.5	0.7	0.6	0.4	0.4
3	0.4	0.4	0.5	0.7	0.7	0.6	0.6	0.4
4	0.4	0.4	0.3	0.6	0.4	0.5	0.6	0.4
5	0.3	0.4	0.6	0.3	0.3	0.4	0.6	0.6
6	0.3	0.3	0.2	0.4	0.4	0.7	0.6	0.5
7	0.2	0.3	0.3	0.4	0.5	0.7	0.6	0.4

Fig. 9 – Best gamma adjustment for the luminance interframe coefficients, according to the OML criterion. Values of the distribution parameter r.

1 – Inter modes show a behavior close to the laplacian. For most of the coefficients the closest gamma fdp corresponds to parameters between 0.8 and 1.0, and the best gauss approach lies in the range $(0.6, 1)$.

2 – Intrafield mode coefficients are better approximated, considering the gamma fdp, by smaller values of the parameter.

3 – In case of intrafield mode, gamma adjusted fdp outperforms the gaussian one. No general rule can be given about the other prediction modes.

4. The optimal quantizer test

In the approximation of empirical distributions by means of theoretical curves, very different criteria can be given for the selection of the best adjustment. Since the application of the model has to eventually judge its validity, criteria for the adjustment should be application-oriented. As an example, the L-level optimum Max-Lloyd (OML) quantizer design has been selected as the problem in which the modelization should help.

The subsequent problem-oriented measure for the fitness of the adjustment can be stated as follows: *given the theoretical distributions p_1 and p_2, with associated OML quantizers q_1 and q_2, p_1 is a better approximation to the empirical distribution* h *if*

$$D(q_1, h) < D(q_2, h)$$

where $D(q_i, h)$ is the signal to quantization noise relation resulting from the application of quantizer q_i to a random variable characterized by distribution h.

Results show that adjustments selected by means of this measure of fitness (Figs. 9 and 10 show them for 128 level quantizers) *in general do not agree* with those obtained by the the classical and not application-oriented criteria χ^2 and MSE. No general rule can be given about higher performance of any of the considered theoretical distributions above the other.

5. Conclusions

The problem of the statistical characterization of the TV signal in the DCT domain has been studied by means of the evaluation of the energy distribution within the DCT coefficients for each prediction mode, as well as by the estimation of the coefficients fdp. An essay has been carried out in order to modelize DCT coefficients by means of the generalized gaussian and gamma distribution.

	0	1	2	3	4	5	6	7
0	0.4	0.5	0.5	0.6	0.6	0.7	0.7	0.7
1	0.5	0.6	0.6	0.6	0.8	0.7	0.7	0.7
2	0.6	0.6	0.8	0.7	0.8	0.8	0.7	0.7
3	0.6	0.7	0.8	0.8	0.8	0.8	0.7	0.9
4	0.6	0.6	0.6	0.6	0.7	0.8	0.8	0.8
5	0.6	0.6	0.6	0.7	0.5	0.8	0.9	0.8
6	0.6	0.6	0.6	0.6	0.7	0.7	0.9	0.7
7	0.6	0.6	0.6	0.6	0.8	0.7	0.8	0.6

Fig. 10 – Best gaussian adjustment for the luminance interframe coefficients, according to the OML criterion. Values of the distribution parameter r.

It has been apparent that reliable estimations concerning the TV signal imply the processing of many different sequences of enough length, since extrapolation from one arbitrary four second sequence to another are sensibly dangerous. Nevertheless, some relevant phenomena have been observed, such as the low *energy compaction* capability of the DCT in the hybrid encoding and the quasi-laplacian behavior of certain coefficients.

The concept of *application oriented fdp estimation* has been introduced and exemplified in case of optimum Max-Lloyd quantizer design. The best gamma and gaussian approximations for this purpose are given and can be compared with those according to non problem-oriented criteria.

References

[1] N.S. Jayant, and P. Noll. *Digital Coding of Waveforms.* Prentice-Hall, 1984

[2] L. Stenger. *Digital Coding of Television Signals - CCIR activities for Standardization.* Image Communication, Vol. 1, June 1989, pp 29-44.

[3] S. Okubo. *Video Codec Standardization in CCITT Study Group XV.* Image Communication, Vol. 1, June 1989, pp 45-54.

[4] *Draft New Recommendation: Transmission of Component-coded Digital Video Signals for Contribution-quality Applications at Third Hierarchical Level of CCITT, Recommendation G.702,* Document CMTT/303, October 1989.

[5] *Proposed Modifications to Report AD/CMTT: Digital Transmission of Component-Coded Television Signals at 30–34 Mbit/s and 45 Mbit/s.* Document CMTT/321, October 1989.

[6] N. García, F. Jaureguizar, J. I. Ronda, and A. Sanz, *HDTV Parallel Codec Simulator.* Proc. Third Int. Workshop HDTV, Torino, August 1989.

[7] R.C. Reininger, and J.D. Gibson. *Distributions of the Two Dimensional DCT Coefficients for Images.* IEEE Transactions on Communications, Vol. COM-31, June 1983, pp 835-839.

[8] R. Bellora, G. Dimino, and M. Muratori. *Hybrid DCT: Comparison of the Statistics of DCT Coefficients and Processing Modes with and without Motion Compensation.* Proc. of the Third Int. Workshop HDTV, Torino, August 1989.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A HIGH-SPEED ADAPTIVE IMAGE DCT CODER WITH PARALLEL ARCHITECTURE FOR VLSI IMPLEMENTATION

W. LIEBSCH

Heinrich-Hertz-Institut Berlin GmbH, Federal Republic of Germany

A new efficient discrete cosine transform (DCT) for bit rate reduction of video signals is presented. The design combines a parallel two-dimensional DCT architecture with distributed coder functions. By taking advantage of the symmetry in the DCT matrix, the internal clock rate is reduced to a factor of 4. This circuit is applicable in advanced television systems (HDTV) operating at video sampling rate up to 80 MHz and can be realized in CMOS technology as a single VLSI component. This architecture utilizes the advantage of parallel and distributed arithmetic to achieve high-speed performance.

1. INTRODUCTION

The need for transmitting a television video signal through a low bit-rate channel has led to the development of various data compression techniques. These include predictive coding and transform coding. For good quality pictures, predictive coding yields rather low compression rate and is very sensitive to transmission errors. Transform coding based on the discrete cosine transform (DCT) algorithm is one of the most widely used techniques in image data compression and yields better performance, however the computational complexity is much higher. It has been proven to be an almost optimum method, due to the fact that conventional image data has a reasonably high inter-element correlation and image information is mainly concentrated in a few transformed values.

Most published VLSI architectures [1],[2] for two-dimensional transform coding are based on sequential calculation of input sub-blocks, rows and columns to obtain the transform values and require a separate coding block. The procedure to achieve a two-dimensional N-point DCT requires 2*N passes through the signal flow graph, additional memory for matrix transposition and comprehensive control logic. These architectures would necessitate a costly and tricky design. The regularity of the data flow and the circuit structure for VLSI implementation of a discrete cosine transform algorithm has to be taken into account, in order to ease the control, to minimize the number of different stages of the pipeline structure, and to lower storage requirements.

In this paper, a new parallel architecture combining direct calculation of two-dimensional DCT values and an adaptive coding procedure will be introduced. It does not require any scratch-pad memory for matrix transposition and is characterized by a regular structure. It utilizes the advantages of parallel and distributed arithmetic to achieve high-speed performance. A one-chip VLSI implementation for high-speed image data compression is made possible.

2. DCT BY DISTRIBUTED ARITHMETIC

In this design, the calculation of DCT or IDCT values is based on the vector matrix product algorithm and is implemented using a bank of 64 basic elements. Even taking the special pattern in the coefficient matrix into account, the maximum internal clock rate is only one quarter of the pixel sampling rate. A direct calculation of the two-dimensional discrete cosine transform values for an 8x8 block is represented by:

$$F(u,v) = \frac{1}{64} \sum_{j=0}^{7} \sum_{k=0}^{7} x(j,k)\, C(u)\, C(v)\, Ci(j,u)\, Ci(k,v) \qquad (1)$$

$$u=0,1\ldots7 \;;\; v=0,1\ldots7$$

$$Ci(n,m) = \cos\frac{(2n+1)m}{16} \;; \qquad C(k) = \begin{cases} 1/\sqrt{2} & \text{for } k=0 \\ 1 & \text{for } k\neq 0 \end{cases}$$

The matrices X(j,k) and F(u,v) represent the original and the transformed blocks respectively; the product C(u)C(v)Ci(j,u)Ci(k,v) is the discrete cosine transform coefficient matrix. Each of the transformed values is a weighted sum of all the block input data. Taking advantage of the specific pattern in the discrete cosine transform matrix, the number of multiplications and additions can be reduced, according to the degree of decimation. By calculating the coefficient matrix expression above and using distributed arithmetic to compute the polynomial product,

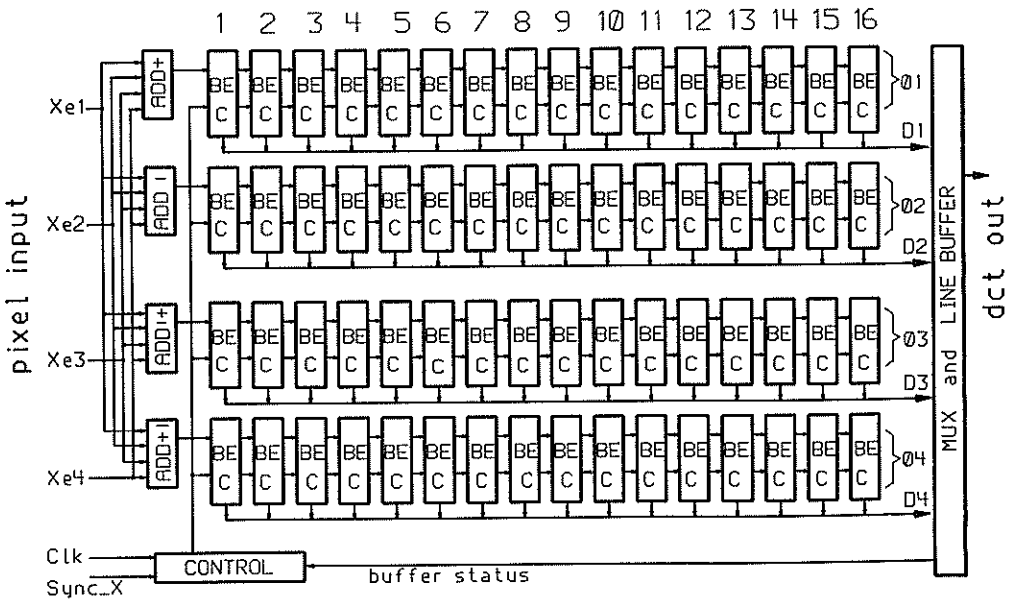


Fig. 1 Block diagram of the two dimensional DCT coder

a separate calculation can be made for each transformed value.

$$F(u) = \sum_{u=0}^{63} \sum_{v=0}^{63} x(v)\ Ck(u,v) \qquad (2)$$

$$Ck(v+8u,k+8u) = \sum_{u=0}^{7} \sum_{v=0}^{7} \sum_{j=0}^{7} \sum_{k=0}^{7} Ci(j,u)\ Ci(k,v)$$

With the decimated coefficient matrix cd(u,v), which is partly shown in Fig. 2 as a graphical representation, the number of arithmetic calculations is reduced by a factor of four. This is utilized fully to lower the internal circuit processing speed for chip realization in CMOS technology and for application in high-speed coder systems.
To maintain regularity of the circuit structure, reduction is made by a factor of four, although a further reduction is possible. The complete calculation of a block with the given size of 8x8 pixels requires 1024 multiplications and additions. 16 cycles are needed for the calculation of each of the 64 discrete cosine transform values. The derived architecture enables a simple arrangement of interconnections of the basic elements and the development of an identical basic element circuit structure. This is important for a full custom VLSI implementation.

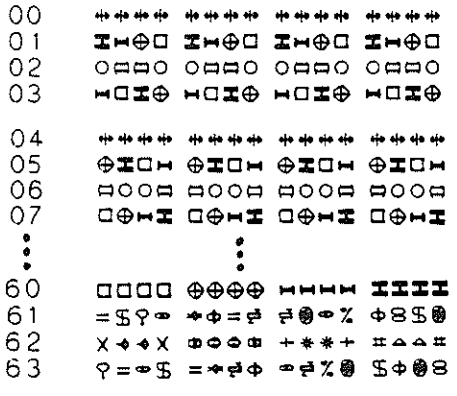


Fig. 2 Graphical representation of the coefficient matrix (partly shown)

3. DCT AND IDCT PROCESSOR

A block diagram of the two-dimensional DCT device based on this approach is shown in Fig. 1. The circuit contains four adder functions and is divided into four data paths. The selection of input pixels and the sequence of DCT output values are shown in Fig. 3. The pixels of an input block are separated into four groups, forming planes G1, G2, G3 and G4. The values are read by tracing the marked line. Four input pixels are calculated simultaneously in adders to reduce the internal processing speed. The IDCT circuit has a similar structure with four adders at the output of the circuit.
Each path of the circuit consists of 16 pipelined basic elements. Using the resulting 16 values, each basic element calculates one DCT value after the pixel precalculation. The basic elements are linked to a common output bus system. Four DCT values are obtained for every four input pixels entered into the DCT processor. Each DCT value is calculated concurrently. A sequential output of the

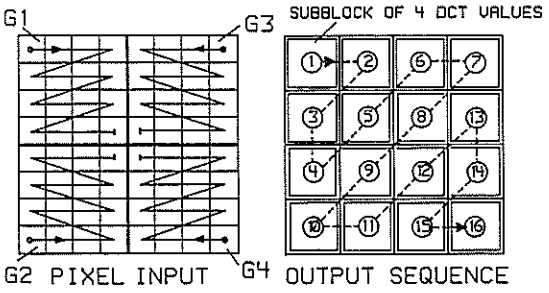


Fig. 3 Pixel input and DCT value output

calculated values is produced by controlling the output buffer in the basic elements. 16 cycles are needed to complete a discrete cosine transform of one block. All basic elements can work on a single output bus, but it is convenient to share the bus to reduce the load for the bus drivers in the basic elements. In Fig. 1 the output bus is divided into four busses.

It is possible to combine DCT and IDCT functions for a one-chip VLSI implementation. Because of the different coefficients for the DCT and IDCT, the basic elements must contain both coefficient sets and additional multiplexer. To combine DCT calculation with an additional coder procedure, it is preferable to design a separate component for DCT coder and IDCT decoder.

4. BASIC ELEMENT

A block diagram of the basic element (BE) is shown in Fig. 4. The structure of a basic element is similar for DCT and IDCT. Each basic element calculates one transform value and consists of a multiplier, an accumulator, an output latch with bus driver and a small control unit. The control unit consists of an associated ROM containing the DCT or IDCT coefficients and some control signals to determine the sequence for outputting the DCT value via the common bus. This sequence can be made adjustable by selecting one of several control signals contained in the ROM. This is useful for an adaptive coding procedure. All basic elements are synchronized by the main control unit. After 16 cycles, the result becomes the appropriate transform value and then is transferred to the output latch. An output will appear at the basic element every 16 cycles. For 15 cycles, the output of the basic elements is disabled and the bus is free for the output of the next basic element values.
All arithmetic functions use a sign-magnitude numbering system. Partial sum overflows during accumulation can be tolerated if the final result is in the dynamic range of a DCT value.

5. ACCURACY OF THE ARCHITECTURE

For circuit realization, the accuracy of DCT values calculated with this parallel structure has been evaluated for a given 8x8 input sub-block in order to determine the minimum number of bits which does not produce any visible degradation after direct and inverse transform.

The overall computational error of this approach results from the rounding operation of the internal values and depends on the selected operand word length. These word length are represented by the following parameters: n2, word length after precalculation, n3, coefficient word length, n4, at multiplier output, n5, at accumulator output, and n6, the final DCT value at the output of the basic element. The number of bits n1 at the input of two-dimensional DCT is normally eight. The optimation procedure is carried out by simulation. The required number of bits for internal operation varies from 8 to 12 bits. For this DCT method, n2=8, n3=8, n4=8, n5=12 and n6=11 bits are sufficient for coding and transmission. To attain a satisfactory level of accuracy, the output of the first basic element (dc value) should contain 12 bits instead of 11. Due to the statistical nature of the input pixels, many of the ac values may be small compared to the dc value and thus would allow for reduced word length. The mean square error (MSE) generated by the limited DCT word length is better than 48 dB related to the maximum pixel value of 255.

6. SCANNING AND CODING

The calculation of DCT values in separate basic elements makes it possible to combine the DCT processor architecture with an adaptive quantizer and coder for bit rate reduction.

Due to the statistical properties of images, the output sequence of DCT values for coding, e.g., zigzag scanning, can easily be fixed by programming the control memory (ROM) in the basic elements without additional hardware. The threshold processing, quantization and coding procedure of the current DCT block values is performed before outputting the DCT values of a basic element. The value of the threshold can be adjusted on a block-to-block basis and depends on the fill-status of the output buffer. The DCT values that are above this threshold are quantized, coded and signed with one bit for transmission. Fig. 4 shows a block diagram of additional coding functions in the basic element.

For the adaptive coding procedure, information is derived from the coding result of the previous basic elements, the current

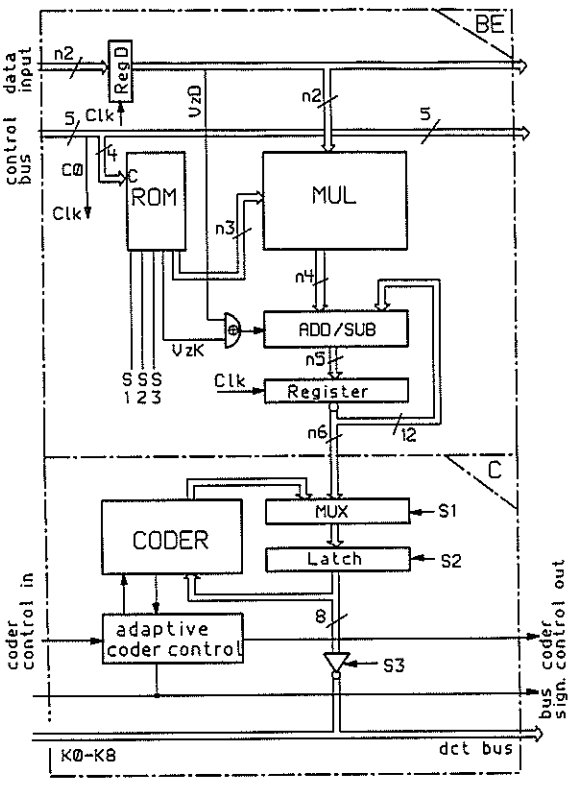


Fig. 4 Basic element (BE) with coder

block activity and the status of the transmission output buffer. A new coding control signal is generated in the active basic element and is fed via a pipeline to the coding module of the next basic element. According to the statistical properties, this procedure begins with the highest order DCT value and the transmission of the block values starts with the first DCT value unequal to zero. The location of a relevant DCT value is marked by one bit. A change of the block scanning is possible, but it must be noted that the overhead to be transmitted in order to decode the information may be a considerable part of the total data.

7. CONCLUSION

The efficient architecture proposed here is applicable for high-speed two-dimensional DCT and IDCT working at video sampling rates up to 80 MHz. This architecture utilizes the advantages of parallel and distributed arithmetic to achieve high-speed performance and can be favourably combined with additional distributed functions for data reduction of video signals.

The resulting circuit structure is highly regular with a minimum of control overhead and, because of the widely identical basic elements, is relatively easy to implement. Using this approach, a two-dimensional discrete cosine transform with a block size of 8x8 pixels can be designed in CMOS technology as a single VLSI component.

REFERENCES

[1] ARNOULD, E., and DUGRE, J.P., Real Time Discrete Cosine Transform an Original Architecture, Proc. ICASSP, 48.6.1 San Diego, 1984.

[2] Jutland, F.,Demassieux,N.,Concordel, G., A Single Chip Video Rate 16x16 Discrete Cosine Transform, ICASSP 86, 15.8.1 Tokyo, 1986.

[3] Chen, W.H., Smith,C.H., Fralick, S.C., A Fast Computational Algorithm for Discrete Cosine Transform, IEEE, Com-25, No 9, 1977.

[4] Lohscheller, H., and Franke, U., Colour Picture Coding-Algorithm Optimization and Technical Realisation, Frequenz 41 Nr.11/12, 1987.

[5] Chen, W.H., Pratt, W.K.,Scene adaptive coder, IEEE Trans., Com-32, No 3, pp. 225-232, 1984.

[6] Madec, G., A comparison between several fast dct algorithms for hardware implementation, PCS 1987, p.177

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

FAST PROGRESSIVE RECONSTRUCTION OF IMAGES USING THE DCT

M. Miran
Honeywell, Inc.
Sperry Commercial
Flight Systems Group
Phoenix, AZ 85036

K. R. Rao
Dept. of Electrical Engineering
The University of Texas at Arlington
Arlington,TX 76019

Fast progressive reconstruction (FPR) of images based on discrete Fourier transform (DFT) and Walsh-Hadamard transform (WHT) has been developed by Takikawa [3]. This technique is now extended to the discrete cosine transform (DCT). The quality of reconstructed images during the intermediate stages based on these transforms is analyzed. This comparison is both subjective and objective. The feasibility of the DCT in this FPR scheme is discussed.

INTRODUCTION

Progressive transmission of images [PIT] is ideally suited for interactive video communication services such as medical imaging, image scanning and retrieval from large data bases, videotex and teletex [1-12]. Transmission of full resolution images over low bit rate networks such as telephone lines ordinarily can take few minutes. In the progressive reconstruction, the image quality is gradually built up from a crude version to the replica of the original. During this build up, the viewer can interactively abort the reconstruction at any stage and scan another image from the library or obtain a near lossless image. The object is two fold; (1) to interactively scan crude versions of the images for initial selection and (2) to reduce the transmission time over low bandwidth channels using data compression techniques. Recognizing the importance of still frame communication, the joint photographic experts group (JPEG) formed from both CCITT and ISO has been developing the international draft standards for the compression and decompression of natural color images [11-12]. A tutorial review of PIT is presented in [10].

Progressive Transmission in Transform Domain

The application of the transform coding concept combines the progressive transmission with an effective data compression technique. The first step is to reconstruct the image from the DC coefficient which represents the average intensity value for each of the sub-blocks. The higher order or AC coefficient transmission allows the image resolution to improve progressively. The main difficulty is to decide which AC coefficient is to be transmitted sequentially from which sub-block.

Lohscheller [9] , Ngan, Leong and Singh [7] and Chitprasert and Rao [6] incorporated human visual system (HVS) sensitivity in the transform coding applied to still images. Dubois and Moncet [8] applied transform coding to progressive transmission of color pictures in NTSC composite format. Elnahas et al [4] extended the transform coding to digital diagnostic images. In 1984, Takikawa [3] developed fast progressive reconstruction (FPR) of images based on discrete Fourier transform (DFT) and Walsh-Hadamard transform (WHT) [14]. In this scheme the image is divided into sub-blocks and various size transforms are applied to groups of transform coefficients of the sub-blocks leading to a fast but somewhat discontinuous build up of an image. Takikawa [3] has suggested that other transforms such as the discrete cosine (DCT) [13] and the discrete sine (DST) [15] may be investigated for the effectiveness in the FPR. The objective of this paper is to investigate the feasibility of FPR based on the DCT and compare its performance with both the DFT and the WHT. This comparison is based on both the visual observation of the reconstructed images at all stages and also on the quantitative parameters such as the signal-to-noise ratio (SNR).

Fast Progressive Reconstruction (FPR) for the DCT

Detailed expressions for the FPR of images based on the DFT and WHT have been developed by Takikawa [3] and hence will not be repeated here. Similar expressions for the DCT will be derived. The 2-D DCT of a 2-D image array x(m,n), m, n = 0, 1, 2, ..., N–1 is defined as [13]

$$X_c(i,k) = \frac{2}{N} C(i)\, C(k) \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} x(m,n)$$

$$\cos\left[\left(\frac{2m+1}{2N}\right)\pi i\right] \cos\left[\left(\frac{2n+1}{2N}\right)\pi k\right] \qquad (1)$$

i, k = 0, 1, 2, 3, ..., N–1, where $C(u) = \frac{1}{\sqrt{2}}$ for u = 0 and C(u) = 1 for u = 1, 2, 3, ..., N–1 and the inverse transform is

$$x(m, n) = \frac{2}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{N-1} C(i)\, C(k)\, X_c(i, k)$$

$$\cos\left[\left(\frac{2m+1}{2N}\right)\pi i\right] \cos\left[\left(\frac{2n+1}{2N}\right)\pi k\right] \qquad (2)$$

$m, n = 0, 1, 2, 3, \ldots, N-1$

The matrix representations of (1) and (2) respectively are $[X_N^c] = [C_N]\,[x_N]\,[C_N]^T$ $[x_N] = [C_N]^T\,[X_N^c]\,[C_N]$ where $[C_N]$ is the N x N DCT matrix and $[\ ]^T$ is the matrix transpose operator. $[x_N]$ and $[X_N^c]$ are the (N x N) matrices in data and DCT domains respectively. DCT matrix can be recursively generated [16] by rearranging the rows and columns of the DCT matrix. To illustrate define $[B_N]$ as $[B_N] = [P_N]\,[C_N]$ where $[P_N]$ is obtained by rearranging the rows of $[I_N]$, an (N x N) unit matrix, from 0, 1, 2, 3 ..., N–1 to 0, 2, 4, ..., N–2, 1, 3, 5, ..., N–1. Then

$$[B_N] = \left[\begin{array}{c|c} [C_{N/2}] & [\tilde{C}_{N/2}] \\ \hline -[\tilde{A}_{N/2}] & [A_{N/2}] \end{array}\right] \qquad (3)$$

where $\left[\tilde{C}_{N/2}\right] = [C_{N/2}]\left[\tilde{I}_{N/2}\right]$, and $\left[\tilde{A}_{N/2}\right] = [A_{N/2}]\left[\tilde{I}_{N/2}\right]$ and $[\tilde{I}_N]$ is an (N x N) secondary diagonal unit matrix. Define $[G_N]$ as $[G_N] = [P_N]\,[X_N^c]\,[P_N]^T$.

Hence $[X_N^c] = [P_N]^T\,[G_N]\,[P_N]$ and

$$[x_N] = [B_N]^T\,[G_N]\,[B_N] \qquad (4)$$

Rearrange $[G_N]$ as

$$[G_N] = \left[\begin{array}{c|c} [G_{N/2}] & [0_{N/2}] \\ \hline [0_{N/2}] & [0_{N/2}] \end{array}\right] + [\hat{G}_N] \qquad (5)$$

where $[G_{N/2}]$ is the upper left quarter of $[G_N]$ and $[\hat{G}_N]$ is the remaining part of $[G_N]$. $[O_N]$ is a (N x N) null matrix. Substituting (5) in (4) gives $[x_N]$ as

$$\left[\begin{array}{c|c} [C_{N/2}]^T[G_{N/2}]\,[C_{N/2}] & [C_{N/2}]^T[G_{N/2}][C_{N/2}][\tilde{I}_{N/2}] \\ \hline [\tilde{I}_{N/2}]\,[C_{N/2}]^T[G_{N/2}][C_{N/2}] & [\tilde{I}_{N/2}]\,[C_{N/2}]^T[G_{N/2}][C_{N/2}][\tilde{I}_{N/2}] \end{array}\right] + [B_N]^T[\hat{G}_N][B_N] \qquad (6)$$

The second term in (6) can be simplified as $[B_N]^T\,[\hat{G}_N]\,[B_N] = [C_N]^T\,[P_N]^T[\hat{G}_N]\,[P_N]\,[C_N] = [C_N]^T\,[\hat{X}_N^c][C_N]$ where $[\hat{G}_N] = [P_N]\,[\hat{X}_N^c]\,[P_N]^T$ and

$$[\hat{X}_N^c] = [P_N]^T[\hat{G}_N]\,[P_N] \qquad (7)$$

From (7) $[\hat{X}_N^c] =$

$$\begin{bmatrix} 0 & X(0,1) & 0 & \cdots & X(0,N-1) \\ X(1,0) & X(1,1) & X(1,2) & \cdots & X(1,N-1) \\ 0 & X(2,1) & 0 & \cdots & X(2,N-1) \\ X(3,0) & X(3,1) & X(3,2) & \cdots & X(3,N-1) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ X(N-3,0) & X(N-3,1) & X(N-3,2) & \cdots & X(N-3,N-1) \\ 0 & X(N-2,1) & 0 & \cdots & X(N-2,N-1) \\ X(N-1,0) & X(N-1,1) & X(N-1,2) & \cdots & X(N-1,N-1) \end{bmatrix} \qquad (8)$$

In (8), $\{[\hat{X}_N^c]\,(2m, 2n)\}$, $m, n = 0, 1, 2, 3, \ldots, (N-2)/2$ are all zeros. Finally $[x_N]$ can be written as

$$[x_N] = \left[\begin{array}{c|c} [x_{N/2}] & [x_{N/2}]\,[\tilde{I}_{N/2}] \\ \hline [\tilde{I}_{N/2}]\,[x_{N/2}] & [\tilde{I}_{N/2}]\,[x_{N/2}]\,[\tilde{I}_{N/2}] \end{array}\right] + [\hat{x}_N] \qquad (9)$$

where $[x_{N/2}] = [C_{N/2}]^T\,[G_{N/2}]\,[C_{N/2}]$ and $[\hat{x}_N] = [C_N]^T\,[\hat{X}_N^c]\,[C_N]$. Equation (9) shows that $[x_N]$ can be broken into $[x_{N/2}]$ plus a remaining part $[\hat{x}_N]$. Applying (9) recursively, we obtain

$$[x_N] = \begin{bmatrix} X(0,0) & X(0,0) & \cdots & X(0,0) \\ X(0,0) & X(0,0) & \cdots & X(0,0) \\ \vdots & \vdots & \cdots & \vdots \\ X(0,0) & X(0,0) & \cdots & X(0,0) \end{bmatrix}$$

$$+ \begin{bmatrix} [\hat{x}_2] & [\hat{x}_2][\tilde{I}_2] & \cdots & [\hat{x}_2] & [\hat{x}_2][\tilde{I}_2] \\ [\tilde{I}_2][\hat{x}_2] & [\tilde{I}_2][\hat{x}_2][\tilde{I}_2] & \cdots & [\tilde{I}_2][\hat{x}_2] & [\tilde{I}_2][\hat{x}_2][\tilde{I}_2] \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ [\hat{x}_2] & [\hat{x}_2][\tilde{I}_2] & \cdots & [\hat{x}_2] & [\hat{x}_2][\tilde{I}_2] \\ [\tilde{I}_2][\hat{x}_2] & [\tilde{I}_2][\hat{x}_2][\tilde{I}_2] & \cdots & [\tilde{I}_2][\hat{x}_2] & [\tilde{I}_2][\hat{x}_2][\tilde{I}_2] \end{bmatrix}$$

$$+ \begin{bmatrix} [\hat{x}_4] & [\hat{x}_4][\tilde{I}_4] & \cdots & [\hat{x}_4] & [\hat{x}_4][\tilde{I}_4] \\ [\tilde{I}_4][\hat{x}_4] & [\tilde{I}_4][\hat{x}_4][\tilde{I}_4] & \cdots & [\tilde{I}_4][\hat{x}_4] & [\tilde{I}_4][\hat{x}_4][\tilde{I}_4] \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ [\hat{x}_4] & [\hat{x}_4][\tilde{I}_4] & \cdots & [\hat{x}_4] & [\hat{x}_4][\tilde{I}_4] \\ [\tilde{I}_4][\hat{x}_4] & [\tilde{I}_4][\hat{x}_4][\tilde{I}_4] & \cdots & [\tilde{I}_4][\hat{x}_4] & [\tilde{I}_4][\hat{x}_4][\tilde{I}_4] \end{bmatrix}$$

$$+ \cdots + \begin{bmatrix} [\hat{x}_{N/2}] & [\hat{x}_{N/2}][\tilde{I}_{N/2}] \\ [\tilde{I}_{N/2}][\hat{x}_{N/2}] & [\tilde{I}_{N/2}][\hat{x}_{N/2}][\tilde{I}_{N/2}] \end{bmatrix} + [\hat{x}_N] \qquad (10)$$

From (10) it is clear that $[x_N]$ can be recursively built up by the combination of a DC term (X (0,0)), a (2 X 2) transform, a (4 X 4) transform, ..., and finally a (N x N) transform.

Reconstruction Process

In the progressive reconstruction, $[X^c_N]$ is decomposed into $(\log_2 N + 1)$ sparse matrices, each of which is inverse transformed by (1 x 1), (2 x 2), (4 x 4), ..., (N x N) algorithms. It should be noted that the progressive reconstruction starts as soon as relevant blocks of information is received. Consider for example the reconstruction process of an (8 x 8) transform. Since N = 8, we decompose $[\hat{X}_8]$ into 4 sparse matrices. So, it takes 4 stages for the complete reconstruction process. **Stage I.** In the first stage, first the data X(0, 0) (the average value of each subblock) received is copied entirely in the 8 x 8 output matrix $[D_8]_I$, which corresponds to the first matrix in (10). **Stage II.** In the second stage, the block of data received is (8). $[\hat{x}_2]$ the 2D-IDCT of (8) is copied in an 8 x 8 matrix so that a matrix $[M_8]_I$ of size 8 x 8 is formed. This is the second matrix in (10). $[M_8]_I$ is added to $[D_8]_I$ in the first stage to produce the reconstructed picture at the second stage i.e., $[D_8]_{II} = [D_8]_I + [M_8]_I$. **Stage III.** In the third stage, the block of data received is $[\hat{X}^c_4]$. $[\hat{x}_4]$ the 2-D IDCT of $[\hat{X}^c_4]$ is copied in an 8 x 8 matrix so that a matrix $[M_8]_{II}$ of size 8 x 8 is formed. This is the third matrix in (10). $[M_8]_{II}$ is added to $[D_8]_{II}$ in the second stage to produce the reconstructed picture at the third stage i.e., $[D_8]_{III} = [D_8]_{II} + [M_8]_{II}$. **Stage IV.** In the fourth stage, the block of data received is $[\hat{X}^c_8]$. $[\hat{x}_8]$ is the 2D-IDCT of $[\hat{X}^c_8]$. $[M_8]_{III}$ is exactly $[\hat{x}_8]$. This $[M_8]_{III}$ is added to $[D_8]_{III}$ in the third stage to produce the reconstructed picture at the fourth stage i.e., $[D_8]_{IV} = [D_8]_{III} + [M_8]_{III}$. For N = 8 this will be the last stage. Hence, at this stage the reconstructed picture will be exactly the original picture, except for the computational errors (finite word length effects). This is the essence of FPR of images by DCT.

SIMULATION RESULTS

The FPR for DCT developed in the previous section is applied to monochrome image Girl (512 512). The pel intensities are represented by 8 bit PCM. FPR is applied to (8 x 8) subblocks of this image. For DFT and WHT, the expressions developed by Takikawa [3] are utilized. Reconstructed images at various stages based on these three transforms are compared both subjectively and objectively. It may be observed that the first stage (DC of each subblock) and the final stage (For 8 x 8 this is the fourth stage.) are identical (subject to computational errors) for all the transforms. For objective comparison normalized SNR (NSNR) is used where NSNR = $-10 \log_{10}$ (NMSE) and

$$\text{NMSE} = \frac{\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}[x(m,n) - x_R(m,n)]^2}{\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}[x(m,n)]^2}$$

Here x(m, n) and $x_R(m, n)$ are the intensities of the original and reconstructed images at row m and column n. Reconstructed images at stage II are shown in Figs. 1-3 for DCT, DFT and WHT respectively. Objective performance of these transforms is described in Tables 1.

SUMMARY AND CONCLUSIONS

The FPR of images developed by Takikawa [3] based on DFT and WHT is now extended to DCT. Detailed expressions of FPR/DCT have been developed, with illustrations for N = 8. Extension to N = 16, 32, ... is straight forward. Both from subjective (Compare the reconstructed pictures at stage II shown in Figs. 1-3) and objective (see Table 1) criteria, WHT is superior to both DCT and DFT. This is because the additions to the reconstructed images at stages II and III for the DCT are based on the terms $[\hat{X}_2]$ and $[\hat{X}_4]$. These are obtained in turn from (8). The 2-D DCT coefficients in both $[\hat{X}_2]$ and $[\hat{X}_4]$ are not the low frequency terms immediately adjacent to the DC component but are farther away. This is in contrast to the WHT where for the intermediate stages of reconstruction, coefficients (representing low sequency) immediately adjacent to the DC component have been utilized. (See Takikawa [3] for the expressions similar to $[\hat{X}_2]$ and $[\hat{X}_4]$ for the WHT). For the DFT also, unfortunately, the same situation as in the case for the DCT arises. To obtain a recursive reconstruction for the DCT, the low frequency coefficients are added at later stages, resulting in undesirable progressive image build up. As is well known, DCT, in general, is superior to other discrete transforms in redundancy reduction. However this advantage has been sacrificed for the sake of recursive reconstruction. As stated earlier, the objective has been to extend the technique of Takikawa to the DCT. It is probably intuitively clear that the discrete sine transform [15] will be equally ineffective, even though similar expressions can be developed.

REFERENCES

[1] W. D. Hoffman and D. E. Troxel, "Making progressive transmission adaptive," IEEE Trans. Commun., vol. COM-34, pp. 806-813, Aug. 1986.

Fig. 1. Reconstructed image of 'Girl', using DCT. Stage II (i.e., $[D_8]_{II}$). Bit rate: 1/2 BPP

Fig. 2. Reconstructed image of 'Girl', using DFT. Stage II (i.e., $[D_8]_{II}$). Bit rate: 1/2 BPP

Fig. 3. Reconstructed image of 'Girl', using WHT. Stage II (i.e., $[D_8]_{II}$). Bit rate: 1/2 BPP

Table 1. Comparison of objective performance (NSNR). For all transforms. Text image: 'Girl'

Stage	Bit rate BPP	NSNR (dB)		
		DCT	DFT	WHT
I	1/8	18.00483	18.00483	18.00483
II	1/2	18.07386	17.89526	21.26269
III	2	18.78011	17.88353	25.13122
IV	8	47.25890	47.17626	47.10894

[2] H. M. Dreizen, "Content-driven progressive transmission of gray-scale images," IEEE Trans. Commun., vol. COM-35, pp. 289-296, Mar. 1987.

[3] K. Takikawa, "Fast progressive reconstruction of a transformed image," IEEE Trans. Inform. Theory, vol. IT-30, pp. 111-117, Jan. 1984.

[4] S.E.Elnahas etal., "Progressive coding and transmission of digital diagnostic pictures," IEEE Trans. Medical Imaging, vol.MI-5, pp.73-83, 6/1986.

[5] L. Wang and M.Goldberg, "Progressive image transmission by transform coefficient residual error quantization," IEEE Trans. Commun., vol. 36, 75-87, Jan. 1988.

[6] B. Chitprasert and K. R. Rao, "Human visual weighted progressive image transmission," IEEE Trans. Commun., vol. COM-38, pp. , 1990.

[7] K. N. Ngan, K.S. Leong and H. Singh, "Adaptive cosine transform coding of images in perceptual domain," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-37, pp. , Nov. 1989.

[8] E. Dubois and J. L. Moncet, "Encoding and progressive transmission of still pictures in NTSC composite form using transform domain methods," IEEE Trans. Commun., vol. COM-34, pp. 310-319, March 1986.

[9] H. Lohscheller, "A subjectively adapted image communication system," IEEE Trans. Commun., vol. COM-32, pp. 1316-1322, Dec. 1984.

[10] K. H.Tzou, "Progressive image transmission: a review and comparison of techniques," Optical Engineering, vol. 26, pp. 581-589, July 1987.

[11] W. B. Pennebaker and J. L. Mitchell, "Standardization of color image data compression. I. sequential coding," Electronic Imaging '89 East, Boston, MA, Oct. 2-5, 1989.

[12] J. L. Mitchell and W. B. Pennebaker, "Standardization of color image data compression II. progressive coding," Electronic Imaging '89 East, Boston, MA, Oct. 2-5, 1989.

[13] K. R. Rao and P. Yip "Discrete cosine transform: algorithms, advantages and applications," Orlando, FL, Academic Press, 1990.

[14] D. F. Elliott and K. R. Rao, Fast Transforms, Algorithms, Analyses and Applications, New York, NY: Academic Press, 1982.

[15] P.Yip and K.R.Rao, "A fast computational algorithm for the discrete sine transform," IEEE Trans. Commun., vol.COM-27, pp.304-307, 2/1980.

[16] H. S. Hou, "A fast recursive algorithm for computing the discrete cosine transform," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-35, pp. 1455-1461, Oct. 1987.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

PICTORIAL TRANSFORM CODING FOR TESSELLATING ARBITRARY SHAPED REGIONS

Yanbin Yu and A.G. Constantinides

Department of Electrical Engineering
Imperial College of Science, Technology and Medicine
London SW7 2BT England

ABSTRACT

An adaptive image coding algorithm is proposed, which has the unique feature of transforming data over arbitrarily shaped regions. The proposed algorithm works in two steps: firstly it segments the image into regions based on an homogeneous criterion; then transform coding is applied to encode the content of the regions. A Shortest Spanning Tree (SST) data structure and segmentation algorithm are employed to produce the partitioning. The major advantage over traditional transform coding and contour/texture methods is that the collective redundancy of image signals and the characteristics of the HVS can be simultaneously utilised.

1. Introduction

In the Cartesian coordinate system, two-dimensional and orthogonal transformations can be easily defined on rectangular shaped regions. Transform coding of an image is therefore usually defined and conducted within blocks. In this paper, an investigation is made into extending the variable block size transform algorithms [1] to a more accurate pattern - producing an arbitrarily shaped region transform.

Traditional coding methods, such as DPCM and transform coding, mainly rely on the exploitation of statistical redundancy. By contrast, texture-contour methods [2,3,4], make use of characteristics of the HVS from the very beginning. While such contour-texture methods have achieved impressive coding results, however, they still require some refinement. The glaring drawback of most existing schemes is the lack of an efficient means of representation of image data within regions. Polynomials usually serve as approximation functionals over given regions, and they are, the authors believe, not as efficient as a transform method. Subjectively, false-contour effects will occur due to the lack of details in the reconstructions. The following summarises basic requirements that an ideal coding algorithm should have:

- The chessboard and the stripe-based patterns of modelling, on which the fixed blocksize and variable blocksize transform coding are based, do not tally precisely with real world image signals. Such coding algorithms bear an inflexible and implicit assumption: that characteristics of the image signal within rectangles are homogeneous. That is, if a given block consists of two or more different signals, say, an edge crossing the block, coding performance will inevitably suffer as a result.
- Extensive studies [5,6] have also demonstrated that the DCT behaves very closely to the optimal transform in most circumstances. This is particularly true in dealing with highly correlated image data.
- Edges, which define the shape of objects, are vital to the final perceptual result, and much more effort should be paid to preserve them as faithfully as possible.

Obviously, the above requirements cannot all be met at the same time by employing a block based scheme. It is thus quite natural to develop new transform coding schemes based on arbitrarily shaped regions.

If the problem is viewed from the angle of the transform, image coding could be thought of as having experienced three different development phases or stages. In the first stage, coding algorithms divide images into blocks of uniform size, classifying the blocks in accordance with a variety of criteria perceived as useful. Most existing transform coding schemes were investigated in this first developmental stage.

The variable size transform may be considered as typical of the second stage. Instead of fixing the block size and varying bit-allocation, block size is made variable to effect adaptation. Since the algorithms invented in these two stages are all constrained by rectangular blocks, they suffer drawback of destroying the inner structure of image data.

The third stage goes under the name of *region-oriented methods*. Coding algorithms developed in this phase should be able to select dynamically the location, shape and size of regions for transformation. A perfect adaptation to the inner structure of image data and coding requirement can be achieved. The other merit is that edges are well preserved, talking full advantage of the HVS.

2. The Basic Ideas

Bearing the above considerations in mind, we are now in a position to study and develop a coding algorithm combining transform and contour-texture methods. The proposed algorithm works in two steps: First it segments the image under criteria called *parameter homogeneity*. The image is separated into two pieces of information, being boundaries and region content. Thereafter, region content is transformed through the DCT, and boundaries are encoded by chain-coding methods. Finally, transform coefficients are quantised and encoded with distortion, and boundaries are handled by an error-free strategy in order to explicitly preserve edges. At the receiver, an inverse transform is carried out and the image is rebuilt from these two pieces of information.

The essential difference between the proposed AST (Arbitrary Shape Transform) and other contour-texture schemes is the method of representing image data over given regions. In AST, image data are represented by transform coefficients, whereas in the traditional contour-texture methods polynomial coefficients are used. The aim of this scheme is to maintain the relatively high compression ratio and preserve edges at the same time. There are two essential operations in the scheme: image segmentation and transform on arbitrarily shaped regions.

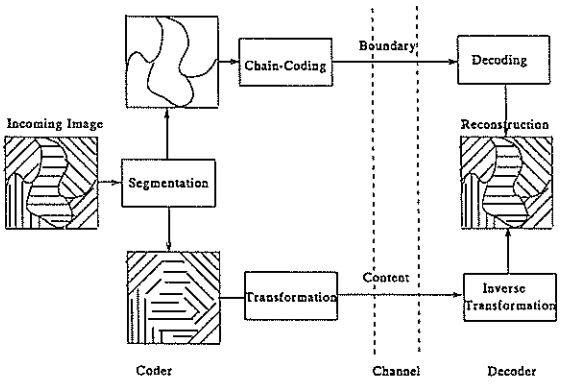


Figure 1. Arbitrary Shape Transform (AST) Coding

3. The Representation of Data in an Arbitrarily Shaped Region

With respect to image compression, the aim is to use the minimum number of functionals to approximate the data in a given region that minimises the error measure. The number of functionals should be kept small to make compression possible. The error measure is normally MSE (Mean Square Error) or SSE (Sum Square Error), and it should be stressed that such measures do not take into account the characteristics of the HVS. However, in view of the inherent philosophy of the proposed algorithm, this error measure is still applicable.

3.1. Polynomial Representation

Polynomial approximation has a simple analytical format and can be easily defined on an arbitrarily shaped region, even in two or more unconnected regions. Most existing research [7,8,4] in the field of image coding employ 2-D polynomials using first few terms to approximate the data. A 2-D polynomial has the following general form:

$$g(x,y)=a_0+\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_i x^i+b_j y^j+c_i x^i y^{n-i}\right)$$

where n is the maximum order of the polynomial.

However, an optimal choice of the maximum order n is still an open question, depending on the characteristics of the given region. It is common practice to choose 3 as the maximum order. The other relevant problem is the quantisation of polynomial coefficients. An upper limit of 8 bits seems reasonable, given that the original data is represented by 8 bits. To date, however, the theoretical analysis of the statistical characteristics of coefficients of polynomials is inadequate. As a consequence, we do not know whether more bits should be allocated to encoding each coefficient or whether a larger number of coefficients be employed with a smaller number of bits being assigned to each. The optimal trade-off between these two seems *ad hoc* and very difficult to determine.

3.2. Transformation

Using a polynomial to represent data shares the same formulation as transform approximation. The general problem of approximating a signal $g(x,y)$ by a transform can be expressed as follows:

$$g(x,y)=\sum_{i=0}^{n_1}\sum_{j=0}^{n_2} a_{ij}f_{ij}(x,y)$$

where $f_{ij}(x,y)$ are the basis vectors of the transform, $n_1+n_2=n$ is the total number of basis vectors and if the DCT is employed, then:

$$f_{00}(x,y)=1$$

$$f_{ij}(x,y)=cos\left(\frac{\pi\cdot i\cdot(x+\frac{1}{2})}{N_x}\right)cos\left(\frac{\pi\cdot j\cdot(y+\frac{1}{2})}{N_y}\right)$$

Beside the merit of a sub-optimal basis, there is also a set of well developed theories and techniques for handling the DCT coefficients. It is desirable to extend transform methods to operate on arbitrarily shaped regions, retaining their higher efficiency in terms of representation of image data.

A straightforward way of applying the DCT to arbitrarily shaped regions is to define an enclosing rectangular region around the one of interest. Hereafter, the transform is applied to this rectangular block. This procedure, called *Blockisation*, is to make the transformation possible on arbitrarily shaped regions. To minimise the number of coefficients, the rectangle should be large enough to just enclose the given region. The problem emerging immediately is how to deal with the pixels which are outside the given region but inside the rectangle. One technique, namely Padding Data, is to be investigated in the following.

3.3. Padding Data

The final goal is to encode the original data efficiently, so the dummy pixels should be chosen in a way that minimises the error measure over the given region. These dummy pixels, however, should not be of the same intensity values as pixels in the original image. This is because if we are to follow the philosophy of the scheme, the characteristics of data within a region will differ from those in an adjacent one.

1) Linear Prediction:

Based on the homogeneity principle, data within a given region can be extended to fill the whole block. Linear prediction, or DPCM, is a solution at hand. Nevertheless, in reality, owing to the arbitrary region shape, it is rather difficult to derive a general prediction equation. A simplified method, named the *Neighbourhood Growing* algorithm, provides a better solution. This is a two-pass approach, as follows:

(1) Starting off with the upper-left corner of the block, the algorithm scans the pixels in a row by row manner. In the second pass, the region is scanned in reverse order.

(2) For each pixel $p(x,y)$

IF it is a pixel residing inside the region or processed pixel residing outside the region, GO to next pixel;

ELSE check the neighbour pixels $p(x+i,y+j)$; $i,j=-1,0,1$;

IF any of these pixels is either inside the region or a processed pixel: calculate the mean of these pixel(s), replace the current pixel with this value, and mark it as a processed pixel.

ELSE GO to the next pixel.

If the given region is connected and just enclosed by the block, the algorithm described above is guaranteed to process every unknown pixel. The final effect of this approach is to spread the value of boundary pixels over the whole block, causing very small transitions. In addition, the computation involved is cheap.

2) Deconvolution Method

It is very important to view the problem from the angle of

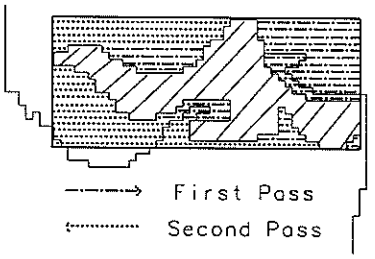


Figure 2. Neighbourhood Growing Algorithm

linear system theory. Image data in an arbitrarily shaped region can be considered as the product of multiplication between a homogeneous signal and a region window. In the spectral domain, the resulting spectrum is the convolution of the spectra of the signal and the window. The task of padding data can be regarded as restoring those *original* pixels corrupted by the windowing effect. Therefore, this is virtually a problem of deconvolution. Since *the constrained iterative method* [9,10] can make use of the prior knowledge in both spatial and spectral domains, it is suited for eliminating this windowing effect. The method can be depicted as in Figure 3.

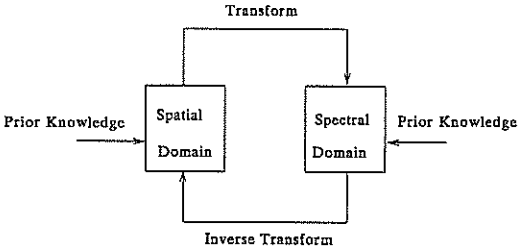


Figure 3. Constrained Iteration Method

The transform involved in the method is usually the DFT (Discrete Fourier Transform) since the phase information can be explicitly utilised. However, here the purpose is to encode image data, so that the DCT is the designated transform. The DCT rather than the DFT should be used in the constrained iteration method. Consequently, in the following analysis, terms such as spectrum will refer to coefficients of the DCT rather than that of the Fourier transform. The prior knowledge with respect to image coding is as follows:

- In the spatial domain, pixels within a given region should have their original values, while those outside the given region should have limited amplitude. Say, pixel values cannot be below zero or above 255 in an 8 bits system.
- The transition between pixels from inside to outside a given region should be as smooth as possible, since any abrupt change in amplitude will cause unnecessary high frequency components.
- In the spectral domain, coefficients normally are *triangularised*. This means that higher order coefficients have the smaller statistical variance and the upper-left corner of the coefficient matrix contains the major part of the image energy. The variance of coefficients usually declines in a zig-zag scan [5] of the coefficient matrix.

The constrained iteration method described above, however, cannot converge to the desired state after a few iterations. In practice, the method should be applied in conjunction with a linear prediction method, such as neighbourhood growing algorithm, to accelerate convergence.

4. Segmentation Techniques

The Recursive Shortest Spanning Tree (RSST) [8], or Region Adjacent Graph (RAG) method [2] equivalently, can make use of global information in the merging operations and the number of segmented regions is also easy to control. This method has a hierarchical structure and is guaranteed to produce connected regions. In brief, the algorithm iteratively joins adjacent regions based on some predefined criterion. The incoming image is mapped onto a graph, one pixel corresponding to one vertex. Costs for every possible merge are then calculated, sorted and stored in a table. Merging is always conducted in such a way that the distortion or cost caused by merging is minimised. It is therefore a stepwise optimal data driven approach, and there is no restriction imposed on the shape of the final segmentation result. Moreover, it can also offer the possibility of embedding flexible merging criteria in context with application.

The choice of cost function depends on the desired purpose of segmentation. A common cost measure is SSE (Sum Square Error):

$$S.S.E. = \sum_{x,y \in region} (g(x,y) - h(x,y))^2;$$

where $g(x,y)$ is the approximation function and $h(x,y)$ is the original data over the given region. In the case of transform coding, the following approximation functional is readily derived:

$$g(x,y) = T^{-1}QT\Big[h(x,y)\Big]$$

where T is the AST (Arbitrary Shape Transform), and T^{-1} is the inverse AST and Q is the coefficient quantisation operation, respectively.

5. Simulations and Results

To assess the effectiveness of the proposed model, computer simulations have been conducted. Techniques for dealing with different sizes of transformation region are also investigated. In this section the details of AST coding algorithms and simulation results are presented.

One piece of information required to enable image reconstruction is the boundary description, which also represents edges. After segmentation, boundaries are simplified as four-way connected lines. The thin-line coding technique developed by Biggar and Constantinides [11] is employed. This method, called Adjacent Direction Run Code, is basically a *differential chain code* [12, 13] with runlength coding. Intersections of edges are represented by a predefined codeword. In the process of encoding, the position of line division is placed on a stack for later tracing, so some bits can be saved by not explicitly addressing the beginning of every boundary.

The full-frame AST coding scheme, however, has two drawbacks. One is that the number of segmentation regions is limited by resource, i.e. the number of bits available, for encoding the boundaries. As a result, the segmentation operation cannot produce enough regions to effectively track the local image statistics. It can be observed that in an active area of an image, where high frequency components have to be retained, there are no obvious edges. In those areas, regions of small size are preferred, but the segmentation operation fails to furnish such ideal partitioning. Also, some false contouring occurs due to the algorithm having to segment very large regions in order to apply AST, costing a lot of bits to address. The second drawback is due to the global processing structure, which on one hand utilises the

whole image to increase efficiency, but on the other requires a large amount of computation memory for manipulating image data causing delay. In the following discussion, an alternative approach is proposed to overcome the drawbacks mentioned above.

The approach is virtually a stripe version of the proposed AST coding algorithm. The incoming image is firstly separated into horizontal, or vertical data stripes, depending on the scanning method. For each stripe, the AST algorithm described above is then applied. The benefits are twofold. Imposing stripe structure will produce block-like segmented regions. This means that part of the boundaries coincide with stripe boundaries, saving bits. Memory requirement and inherent time delay will also be substantially reduced. However, all these benefits are achieved at the penalty of losing generality and some straight line structure may appear.

Figure 4 shows the reconstruction of the *Amalia* image and its corresponding segmentation. Compared with results of a full frame approach, more than 1.5 dB improvement is obtained in terms of SNR. When compared with results obtained by the Scheme based on blocks, the blocking effect has been eliminated, although only limited SNR improvement can be observed.

(a) DCT Coded: 21.39dB (b) AST Coded: 22.70dB

(c) RSST Coded: 21.13dB (d) AST Segmentation

Figure 4. The Coding Results by the Alternative AST Approach

6. Conclusions

In this paper, an entirely new paradigm of transform coding is proposed and computer simulations are presented. The unique characteristic of the proposed algorithm is the breaking of the rectangular block constraint which exists in most traditional transform coding schemes.

Traditional schemes, nevertheless, ignore the importance of partitioning, and they rely on effecting the adaptation within the given blocks. It is quite common that blocks lies on the change place of the statistics, therefore, much efforts have to be paid.

The intrinsic mechanism of VST [1] and AST is virtually the same. By varying boundaries and positions of regions, we can reduce the overall number of different patterns. Consequent transformation, therefore, becomes more efficient.

References:

[1] Y.B. Yu and A.G. Constantinides, "Variable Block Size and Position Transform Coding", *Fourth European Signal Processing Conference*, Grenoble, France (Sept. 1988).

[2] M. Kunt, A. Ikonomopoulos, and M. Kocher, "Second-Generation Image-Coding Techniques", *Proc. IEEE*, Vol. 73, pp. 549-574 (Apr. 1985).

[3] M.J. Biggar, "Source Coding of Segmented Digital Image and Video Signals", *Ph.D Thesis, Imperial College, University of London*, London (1987).

[4] V.J. Stanger and A.E. Symons, "The Application of Image Analysis Techniques to Low Bit Rate Coding of Colour Video-Conferencing Sequences", *ALVEY Project Report* (Sept. 1987).

[5] R.J. Clarke, *Transform Coding of Images*, Academic Press (1985).

[6] A.N. Netraval and J.O. Limb, "Picture Coding : A Review", *Proc. IEEE*, Vol. 68, pp. 366-406 (Mar. 1980).

[7] M.Eden, M. Unser, and R. Leonardi, "Polynomial Representation of Picture", *Signal Processing*, Vol. 10, pp. 385-393 (1986).

[8] M.J. Biggar, O.J. Morris, and A.G. Constantinides, "Segmented-Image Coding: a Performance Comparison with the Discrete Cosine Transform", *IEE Proceedings Part F*, pp. 121-132 (1988).

[9] U. Franke, R. Mester, and T. Aach, "Constrained Iterative Restoration Techniques: A Powerful Tool in Region Oriented Texture Coding", *Fourth European Signal Processing Conference*, North-Holland, Grenoble, p. 1145 (Sept. 1988).

[10] R.W. Schafer, R.M. Mersereau, and M.A. Richards, "Constrained Iterative Restoration Algorithms", *Proc. of IEEE*, Vol. 69, pp. 432-450 (April 1981).

[11] M.J. Biggar and A.G. Constantinides, "Thin Line Coding Techniques", *Proc. International Conference on Digital Signal Processing*, Florence, pp. 471-475 (Sept. 1987).

[12] H. Freeman, "Computer Processing of Line-Drawing Images", *Computer Survey*, Vol. 6, pp. 57-97 (July 1974).

[13] T. Kaneko and M. Okudaira, "Encoding of Arbitrary Curves Based on the Chain Code Representation", *IEEE Trans. Comm.*, Vol. COM-33, pp. 697-706 (July 1985).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

IMPROVED PERMUTATION CODES AND THEIR APPLICATION TO DISCRETE COSINE TRANSFORM IMAGE CODING

Takahiro SAITO+ , Takashi KOMATSU+, and Hiroshi HARASHIMA ++

+ Department of Electrical Engineering, Kanagawa University
3-27-1 Rokkakubashi, Yokohama, 221, JAPAN
++ Engineering Research Institute, The University of Tokyo
2-11-16 Yayoi, Tokyo, 113, JAPAN

Permutation codes, previously developed by Berger et.al.,which codes do not require any multiplication for encoding and can quantize a very high dimensional input efficiently, however cannot show satisfactory performance for high rates or low dimensions. To cope with this problem, extending the concept of permutation codes, we developed new improved permutation codes, and incorporated improved permutation codes into discrete cosine transform image coding. The simulation results demonstrated that improved permutation codes can quantize transform coefficients of a high frequency efficiently.

1.INTRODUCTION

Conventional discrete cosine transform(DCT) image coders quantize the transform coefficients using a scalor quantizer. The work herein introduces permutation codes[2], previously proposed by Berger et.al., or improved permutation codes, developed herein extending the concept of permutation codes, into DCT image coding instead of a scalor quantizer, and improves its coding performance.

We formerly demonstrated that the normalized AC transform coefficients of a sampled image are well modeled as having a spherically symmetric distribution[1]. For the probability distribution, the quantizer whose representative vectors, codewords, are arranged on the surface of concentric hyperspheres will show excellent performance with its low encoding complexity. Permutation codes[2],[3] and a gain/shape vector quantizer[4] have this property. In gain/shape vector quantization, it takes a large number of multiply-adds to quantize a given input vector, and hence the number of dimension is limited to a low number. On the other hand, in permutation codes, it does not take any multiply-add to quantize a given input vector, and a very high dimensional input vector can be quantized directly. Permutation codes, however, cannot show satisfactory performance for high rates or low dimensions, and require unfeasible operation precision for encoding a reproduction index in high dimensional cases.

To cope with these problems, we develop new improved permutation codes, and form an algorithm to encode a reproduction index,an algorithm which involves only integer operations with limited operation precision. We omit a full detail of the algorithm for encoding a reproduction index from this paper on account of limited space. Furthermore improved permutation codes are incorporated into DCT image coding, and computer simulations are conducted on monochrome images.

2.EXTENSION OF PERMUTATION CODES

Two types of permutation codes, Variant I and Variant II permutation codes, were developed[2]. The work herein deals with and improves only Variant II permutation codes, but of course the extension is applicable to Variant I permutation codes.

2.1 Permutation Codes[2]

The codewords $\vec{y}_1,\vec{y}_2,\ldots,\vec{y}_M$ of Variant II permutation codes are chosen in the following manner. The first codeword $\vec{y}_1$ is a n-dimensional vector of the form

$$\vec{y}_1=(\overleftarrow{\mu_1,\ldots,\mu_1}^{n_1\rightarrow},\overleftarrow{\mu_2,\ldots,\mu_2}^{n_2\rightarrow},\ldots,\overleftarrow{\mu_k,\ldots,\mu_k}^{n_k\rightarrow}) \quad (1)$$

$$,\mu_1>\mu_2>\cdots\cdots>\mu_{k-1}>\mu_k=0 \quad (2)$$

$$\sum_{i=1}^{k} n_i=n \quad (3)$$

The value of μ_k is fixed at zero on the assumption that the source distribution is symmetric about the origin. The other codewords are formed by assigning algebraic signs to the components of $\vec{y}_1$ in all possible ways, and then permuting these signed components in all possible ways. The number of codewords M is

$$M=2^{(n-n_k)}\cdot n!/\prod_{i=1}^{k}[(n_i)!] \quad (4)$$

The bit rate I is

$$I=n^{-1}\lceil\log_2 M\rceil \text{bit/sample} \quad (5)$$

where $\lceil\log_2 M\rceil$ denotes the smallest integer equal to or greater than $\log_2 M$. For a given input vector x, minimum distortion mapping is accomplished by the simple algorithm described below.

1)Replace the n_1 components of $\vec{x}$ largest in absolute value by either $+\mu_1$ or $-\mu_1$, the sign chosen to agree with that of the components it replaces.
2)Replace the n_2 components of $\vec{x}$ next largest in absolute value by either $+\mu_2$ or $-\mu_2$, the sign chosen to agree with that of the components it replaces.

:

k)Replace the n_k components of $\vec{x}$ smallest in absolute value by 0.
k+1)Use the codeword that results from these replacements to represent $\vec{x}$.

For Variant II permutation codes, the sorted index arrangement and the sign pattern must be encoded. We have developed an algorithm to encode the sorted index arrangement, an algorithm which involves only integer operations with limited operation precision, by introducing two techniques named stepwise coding and path-splitting into the Schalkwijk's coding algorithm[5]. In the coding algorithm, the limitation of operation precision induces an increase in a bit rate compared with the ideal bit rate given by Eq.(5). As operation precision is limited more severely, this coding loss increases gradually. Operation precision is limited to 32 bits herein, which severe limitation induces little coding loss.

Define the random variable η_j to be the jth largest of the absolute values of the components of $\vec{x}$. The best choice of the parameters μ_i ($i \neq k$) for given n_i's are

$$\mu_i = \frac{1}{n_i} \sum_{j=s_{i-1}+1}^{s_i} E[\eta_j] \qquad (6)$$

$$, s_i = n_1 + n_2 + \cdots + n_i$$

$$s_0 = 0$$

The order statistics $E[\eta_j]$ is used for the design of Variant II permutation codes. We herein estimate the order statistics using a long training sequence of data. Berger et.al. developed an iterative technique that searches for the values of k, n_i, and μ_i that minimize the MSE values for a given bit rate and dimension. We herein used this iterative technique.

2.2 Improved Permutation Codes

Since all the codewords of permutation codes are arranged on the surface of the same hypersphere, permutation codes cannot cope with the norm distribution of vectors emitted by an information source, which fact leads to unsatisfactory performance for high bit rates or low dimensions. We develop improved permutation codes aimed at enhancing the suitability for the norm distribution.

In improved permutation codes, for a given input vector $\vec{x}$ the average ξ_i ($i \neq k$)

$$\xi_i = \frac{1}{n_i} \sum_{j=s_{i-1}+1}^{s_i} \eta_j \qquad (i < k),\ \xi_k = 0 \qquad (7)$$

is computed and the scalor-quantized value ξ_i of the ξ_i is substituted for the μ_i of Variant II permutation codes given by Eq.(6). The above equation is derived by omitting the expectation operation $E[\cdot]$ from Eq.(6). The ξ_i's also satisfy

$$\xi_1 > \xi_2 > \cdots > \xi_{k-1} > \xi_k = 0 \qquad (8)$$

, and utilizing this property the reproduction level ξ_i of the ξ_i must be encoded. This extension enables permutation codes to cope with the norm distribution of vectors generated by an information source.

Figure 1 illustrates the coding process in improved permutation codes. In Fig.1, quantization and reconstruction are performed in the order, ①, ②, ···, ⑦, and three different scalor quantizers of 4 representative levels are used to encode the ξ_i. The reproduction level of the ξ_1, "2.0", is encoded with a binary code of 2 bits, but we can reduce a code length to less than 2 bits for the ξ_2 and ξ_3. The representative levels of "4.5" and "2.5" of the scalor quantizer for the ξ_2 will never be chosen as a reproduction level, because the reproduction level of the ξ_1 is "2.0" and the ξ_i's satisfy Eq.(8). Therefore the ξ_2 is scalor-quantized with the two remaining representative levels of "1.8" and "1.2", and the reproduction level of "1.8" is encoded with a binary code of 1 bit. In the same manner, a code length for the ξ_3 is reduced to 1 bit.

Improved permutation codes embodies the idea that

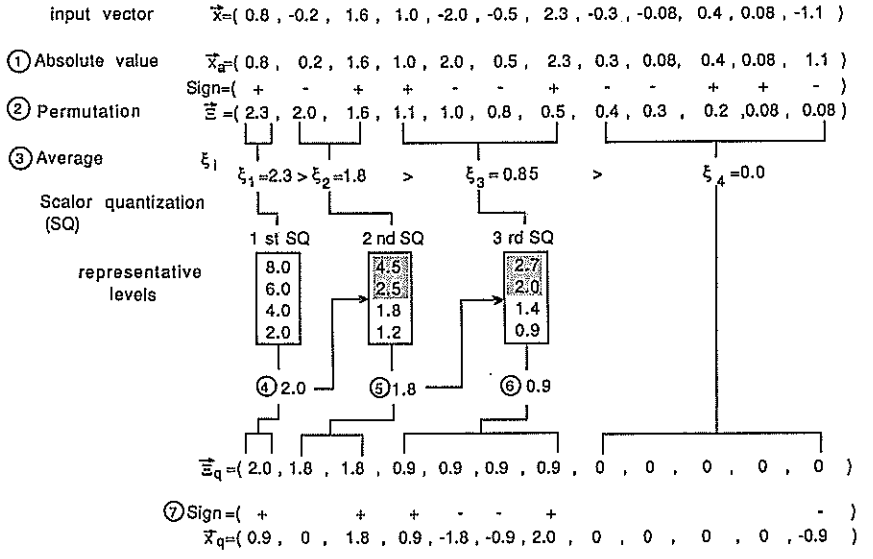


Figure 1 Quantization process in improved permutation codes.

different kinds of permutation codes are mixed together and used to encode a given input vector and that for a given input vector the optimum permutation codes yielding the minimum distortion are chosen. This extension of permutation codes is expected to enhance its suitablility for various types of information sources. For compression of speech waveforms Lu et.al. have developed another variation of permutation codes based on the above-mentioned idea, named composite permutation codes[3], which show much the same performance as improved permutation codes. In composite permutation codes, however, it takes a number of multiply-adds to choose the optimum permutation codes according to the nearest-neighbour rule for a given input vector, as in vector quantization. On the other hand, in improved permutation codes it does not take any multiply-add to quantize a given input vector as in permutation codes. Furthermore improved permutation codes are designed more easily than composite permutation codes. In the design of composite permutation codes, their first codewords are designed by the LBG algorithm[7] as in the design of a vector quantizer. On the other hand, in the design of improved permutation codes, we employ the values of k and n_i that are determined for permutation codes by the Berger-Jelinek-Wolf algorithm[2], and have only to design the scalor quantizers for the ξ_i's by the Lloyd's algorithm[6].

2.3 Quantization Performance

Figure 2 shows rate versus distortion performance of a 4-dimensional gain/shape vector quantizer(G/S VQ), 100-dimensional permutation codes (PC), and 100-dimensional improved permutation codes (PC•I) for a 4-dimensional Pearson Type VII distribution whose probability density is given by

$$P_{(n)}(\vec{x}) = \frac{2^{\nu}\cdot(\nu-1)^{\nu}\cdot\Gamma(\nu+\frac{n}{2})}{\Pi^{\frac{n}{2}}\cdot\Gamma(\nu)\cdot[2(\nu-1)+|\vec{x}|^2]^{\nu+\frac{n}{2}}} \qquad (9)$$

$$, \ \nu > 1$$

where the values of n and υ are fixed at 4 and 2.5

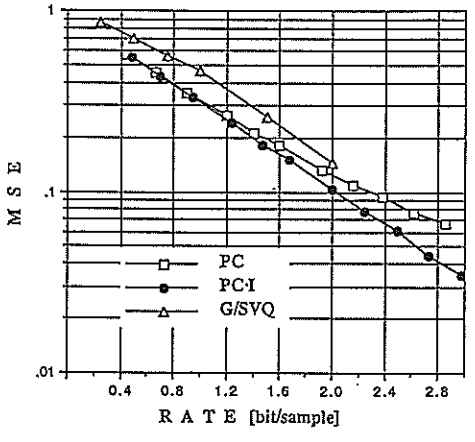


Figure 2 Comparision of mean squared error quantization performance for a Pearson Typr VI distribution.

respectively. The Pearson Type VII distribution is a good spherically symmetric probability model of the distribution of the normalized AC transform coefficients[1]. Table 1 shows the number of representative levels of the scalor quantizer used in PC•I, and these numbers of Table 1 yield nearly optimum performance in most cases. An input vector fed to PC or PC•I is formed by concatenating 25 successive 4-dimensional vectors emitted by the information source having the 4-dimensional Pearson Type VII distribution.

Table1 Number of representative levels of scalor quantizer used in improved permutation codes.

k	Number of representative levels					
	L_1	L_2	L_3	L_4	L_5	L_6
2	8	1	—	—	—	—
3	4	4	1	—	—	—
4	4	4	4	1	—	—
5	4	4	4	4	1	—
6	4	4	4	4	4	1

In Fig.2, PC outperforms G/S VQ especially at the bit rate under 1 bit/sample, but its superiority decreases with higher bit rates. PC•I shows better performance than PC, but its superiority decreases with lower bit rates. When the number of dimension is 100 or less, PC•I outperforms PC especially at the bit rate over 1 bit/sample.

3.APPLICATION TO DCT IMAGE CODING

In conventional DCT coders, a large number of the high frequency transform coefficients are usually discarded, because the number of bits assigned to each of those transform coefficients falls below 1 bit. Herein those high frequency transform coefficients are quantized using permutation codes or improved permutation codes, whereas the other low frequency transform coefficients are scalor-quantized.

In the DCT coder using permutation codes(PC) or improved permutation codes(PC•I), named DCT-PC and DCT-PC•I, a given input image is divided into 16x16 pel subblocks, 2-dimensional DCT is performed on each subblock, and the bit assignment matrix is determined. In every transform subblock, the coder normalizes and quantizes the transform coefficients whose allocated number of bits exceeds 2 bits using a uniform scalor quantizer designed for a standard Gaussian distribution of zero mean and unit variance, and encodes their quantization outputs with Huffman-codes designed for the standard Gaussian distribution. In every transform subblock, the coder additionally selects the transform coefficients to which 2 bits or less are assigned in the order of decreasing variance to pick up 128 transform coefficients, thus forming a 128-dimensional vector. The 128-dimensional vector is encoded using permutation codes or improved permutation codes which are designed on a training sequence of data taken from many natural images outside test images. Improved permutation codes employ the number of representative levels shown in Table 1. The coding system additionally encodes the overhead information : the bit assignment matrix, the normalization

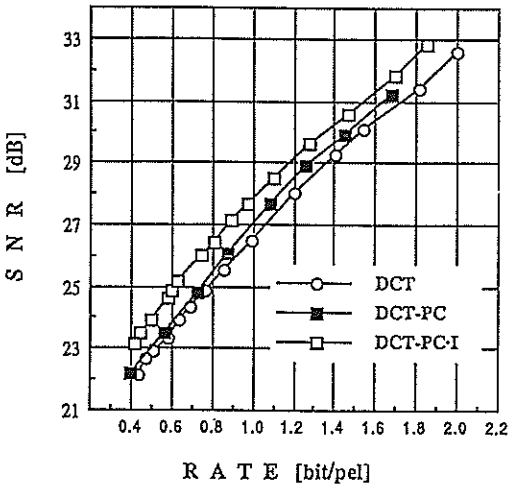


Figure 3 Rate versus SNR performance of nonadaptive DCT coders for the test image "Tulip".

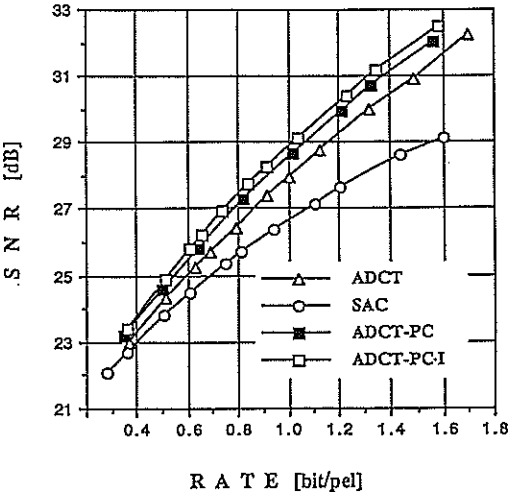


Figure 4 Rate versus SNR performance of adaptive DCT coders for the test image "Tulip".

factors, and the identification labels by which the decoder is informed which transform coefficients are gathered as a component of the 128-dimensional vector.

Figure 3 and Fig.4 show the simulation results of coding performance for the ITE (The Institute of Television Engineers of Japan) test image "Tulip", which is composed of 512x480 pels with each intensity value uniformly quantized to 256 levels. SNR is defined by

$$SNR = 20 \log_{10} (255/\sqrt{M.S.E.}) \qquad (10)$$

In Fig.3 and Fig.4 several DCT coders are compared :
(1)"DCT" -- Conventional DCT coder where all the normalized transform coefficients are encoded by using a uniform scalor quantizer and Huffman-codes both of which are designed for the standard Gaussian distribution.
(2)"DCT-PC" and "DCT-PC·I"
(3)"ADCT","ADCT-PC", and "ADCT-PC·I"---- Adaptive versions of DCT, DCT-PC, and DCT-PC·I where the adaptive technique based on an activity index[8] is employed.
(4)"SAC" --Scene adaptive DCT coder previously proposed by Chen and Pratt[9]. SAC employs Huffman-codes designed on the same training sequence of data that is used for the design of PC and PC·I.

In Fig.3, DCT-PC·I gives the best performance and outperforms DCT by about 1.5 [dB]. In Fig.4, ADCT-PC I outperforms ADCT and SAC especially at the bit rate over 0.8 bit/pel.

4.CONCLUSIONS

Extending the concept of permutation codes, we developed improved permutation codes. The simulation results showed that improved permutation codes are more useful than prototypal permutation codes, previously developed by Berger et.al., as a means for quantizing vectors generated by a time-discrete information source. Furthermore we introduced improved permutation codes into discrete cosine transform (DCT) image coding. The simulation results demonstrated that improved permutation codes can quantize transform coefficients of a high frequency more efficiently than prototypal permutation codes and a scalor quantizer with entropy coding.

REFERENCE

[1] T.Saito et. al. ,"Gain/shape vector quantizer for multidimensional spherically symmetric random source", Trans. IECE Japan,vol.J68-B,pp.904-911,Aug.1985.
[2]T.Berger, F.Jelinek, and J.K.Wolf,"Permutation codes for sources", IEEE Trans. Inform. Theory, vol.IT-18, pp.160-169,Jan.1972.
[3] L.Lu, G.Cohen, and P.Godlewski, "Composite permutation coding of speech waveforms",Proc. IEEE Int. Conf. Acoust.,Speech & Signal Process., pp.2359-2362, April 1986.
[4]M.J.Sabin and R.M.Gray, "Product code vector quantizers for waveform and voice coding", IEEE Trans.,Acoust.,Speech & Signal Process., vol.ASSP-32, pp.474-488,June 1984.
[5]J.P.M.Schalkwijk, "An algorithm for source coding" ,IEEE Trans. Inform. Theory, vol. IT-18, pp.395-399, May 1972.
[6]S.P.Lloyd,"Least squares quantization in PCM ",IEEE Trans. Inform. Theory, vol.IT-28,pp.129-137,Mar. 1982.
[7]Y.Linde,A.Buzo and R.M.Gray,"An algorithm for vector quantizer design",IEEE Trans. Commun., vol. COM-28,pp.84-95,Jan. 1980.
[8]W.H.Chen and C.H.Smith,"Adaptive coding of monochrome and color images", IEEE Trans. Commun., vol.COM-25,pp.1285-1292,Nov.1977.
[9]W.H.Chen and W.K.Pratt,"Scene adaptive coder",IEEE Trans. Commun.,vol.COM-32,pp.225-232,Mar.1984.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

NEW HYBRID SPLINE-LINEAR INTERPOLATION FOR THE FAST CT AND MR IMAGING

Samuel MATEJ

Institute of Measurement and Measuring Engineering, Slovak Academy of Sciences
Dúbravská cesta 9, 842 19 Bratislava, Czechoslovakia

The modern tomographic applications like imaging of 3-D objects or dynamic processes require reconstruction techniques of extra high speed. One of the most promising reconstruction methods is the Direct Fourier Method (DFM). Provided an appropriate reconstruction parameters (discretization parameters, filter type, 2-D IFT dimension, interpolation type) are used, the DFM offers results of the same quality as the classical Convolution BackProjection Method yet in considerable shorter time. The most important reconstruction parameter is the accuracy of spectrum data interpolation. The interpolation proposed enables the fast image reconstruction of high quality. This fact is demonstrated on results of our simulation research.

1. INTRODUCTION

The tomographic imaging methods represent one of the most topical and rapidly developing areas of image science. Especially, the modern 3-D and dynamic methods both in CT (X-ray) and MR tomography seem to be topical for scientific research, because the study of dynamic processes or moving organs is limited in the contemporary commercial systems. Namely, in the CT tomography, the commonly used reconstruction method - Convolution Back-Projection Method (CBPM) - does not fit the new fast 3-D and dynamic tomographic methods, because of high computational costs, even when a specialized hardware is available. The Direct Fourier Method (DFM) seems to be a convenient candidate for such applications. The classical MR methods collecting the Fourier spectrum data on the rectangular raster lines are slow. However, the new methods have been proposed, which collect data from the entire spectrum plane in one or in several measurements (e.g. placed on spiral trajectories). The fast methods both in CT and MR tomography have a common feature, namely that the measured data correspond to the Fourier spectrum data that are not given on the rectangular raster. Consequently, in both cases the reconstruction process requires an interpolation from the given Fourier spectral raster to the rectangular one.

2. IMAGE RECONSTRUCTION

As well known, in the CT-tomography, the measured data - projections - can be described by Radon transformation [1]:

$$p(s;\theta) = \mathcal{R}\,[f(x,y)] = \int_{-\infty}^{\infty} f(s\cdot\cos\theta - q\cdot\sin\theta,\ s\cdot\sin\theta + q\cdot\cos\theta)dq\ ,$$

where s is the distance from the space origin and θ represents angle of the projection line. The Direct Fourier Method is based on the formula:

$$f(x,y) = [\ \mathcal{F}_2^{-1}\mathcal{F}_y\ ]\ p(s;\theta) = \mathcal{F}_2^{-1}P(S;\theta)\ ,$$

where $\mathcal{F}$ denotes the Fourier transformation operator. According to the discrete version of the projection slice theorem [1], the 2-D Fourier spectrum of the function f(x,y) is obtained on a polar point raster. Since the 2-D inverse Fourier transformation (2-D IFT) does not posses a fast algorithm for the polar raster, an interpolation of the Fourier spectrum data to a rectangular raster has to be performed.

Analogously, an interpolation has to be performed in the fast MR methods, where the measured data are given on a Fourier spectral raster which is not rectangular. An effective implementation of the interpolation algorithm in the case of spiral raster is based on the following principles. The FID signal corresponding to the values of the 2-D Fourier spectrum on a spiral has to be sampled by the same number of samples in each revolution of the spiral. Then the collected data are regularly distributed on the angular lines. In comparison with the polar raster each angular line has an offset given by the spiral parameters. Taking this offset into account the interpolation computation for the spiral point raster can be carried out analogously as for

the polar point raster.

From the viewpoint of reconstruction quality, sufficiently precise interpolation represents the crucial point of DFM reconstruction [3,4, 5]. Moreover, in [1,5,6] was showed, that for particular type of interpolation, reconstruction errors can be substantially suppressed providing the number of sampling points in the Fourier spectrum domain is radially increased 2^{β}-times, where $\beta=0,1,2,\ldots$ The choice of an appropriate value of the parameter β depends on the precision of the interpolation used.

The discrete nature of an actual reconstruction problem leads to reconstruction errors, namely the aliasing errors, and the spatial oscillations caused by a truncation of the Fourier spectrum (Gibbs phenomenon). These artifacts may be reduced by filtering the function using a filter whose values smoothly taper to zero at the cutoff frequency. This is reached by an expense of slightly smoothed image. As an compromise, the COS filter has been found [5,7,8] to be the most convenient.

Furthermore, the reconstruction errors caused by the truncation of the image spectrum and the nonexact interpolation are encountered also in a surrounding of the reconstructed image. These are aliased in the space domain. Such effect can be suppressed by enlarging the image domain (i.e. object surrounding). The 25% enlargement has been found [5,7,8] to be sufficient.

3. INTERPOLATION

The 2-D interpolations have been analyzed from the viewpoint of two controversial demands - interpolation accuracy, and speed. We have considered the following interpolation types in our research: nearest neighbor, linear, bilinear, spline [6] interpolation and near-exact interpolation proposed by Stark [4]. Among the simple types of interpolation, the bilinear interpolation is sufficiently fast, but its accuracy is low, unless the raster density is radially increased approximately 8-times [6,7]. On the other hand, more precise interpolations like spline interpolation are less convenient if the computational demands are to be considered. Since the Fourier spectrum of tomographic images varies much faster in the radial direction than in angular one, the overall interpolation accuracy is much more sensitive to interpolation accuracy in the radial direction. Based on this result, the hybrid two-stage interpolation has been designed, which utilizes the spline de Boor's cubic interpolation [9] in the radial direction and the linear interpolation in the angular one.

Interpolation of the value in the point A (see Fig.1) is performed using the following formula (equivalently for the point B):

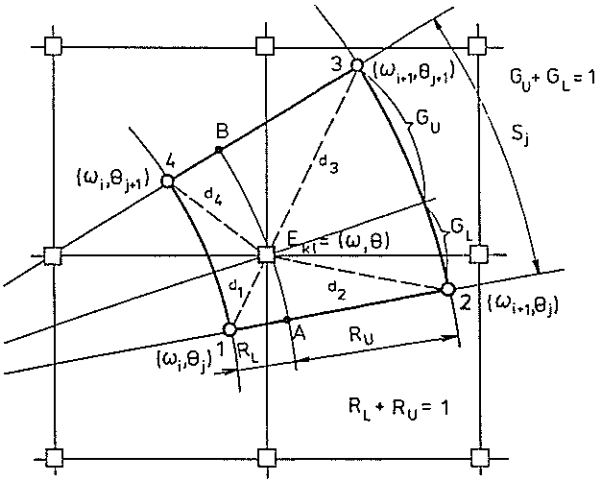


Figure 1

$$F(A) = u_{i,j} + p_{i,j}R_L +$$

$$+ \left\{ 3(u_{i+1,j} - u_{i,j}) - p_{i+1,j} + 2p_{i,j} \right\} R_L^2 +$$

$$+ \left\{ -2(u_{i+1,j} - u_{i,j}) + p_{i+1,j} + p_{i,j} \right\} R_L^3 ,$$

where $u_{i,j} = F(\omega_i,\theta_j)$, and
$p_{i,j}$ denotes the partial derivation in the radial direction;

Partial derivations are computed for each angular sector at once on the basis of [9]. Then the R-raster value is calculated as:

$$F(\omega,\theta) = G_U \cdot F(A) + G_L \cdot F(B) .$$

An appropriate raster scale has been adopted into these formulae - see radial (R) and angular (G) distances in Fig.1.

The proposed hybrid spline-linear interpolation enables to increase the computational speed considerably in comparison with the spline interpolation. At the same time the reconstruction quality remains preserved.

4. RESULTS

Both, the MR measurement of data on the spiral trajectories, and the polar raster in the Fourier spectrum domain have been simulated for the mathematically defined phantoms (the well known Shepp and Logan [10] phantom). In the simulation research, different data discretiza-

tion and reconstruction parameters have been used. We have simulated the noise of MR measurement too. It has been found that reconstruction of the images containing objects of low local contrast is especially sensitive to the choice of the reconstruction parameters (2-D IFT dimension, filter type, interpolation accuracy). The simulation results using the DFM with distinct interpolations applied have been compared with the results obtained by the CBPM.

The reconstruction quality has been evaluated on the basis of mathematical measures [1] and visual quality [8]. Fig.2 demonstrates (102x102 phantom) that the same visual quality can be obtained via the CBPM (top left image) and the DFM using bilinear interpolation (DFM/BI) and the parameter β=3 (top right image), hybrid spline-linear interpolation (DFM/SPLI) and the parameter β=1 (bottom left image) and spline interpolation (DFM/SP) and the parameter β=1 (bottom right image). The rest reconstruction parameters (optimal [7,8]) are as follows: COS type filtering, 25% 2-D IFT enlargement, 150 projection angles. The use of interpolation proposed by Stark [4] (DFM/ST) does not provide sufficient reconstruction quality in the case of the "low" contrast image reconstruction.

The simulation results are demonstrated also for real tomographic image reconstructed via CBPM (Fig.3 - top left) and DFM/SPLI (top right) with the same parameters as above. The difference images (absolute values - left, normalized values - right) are placed in the bottom of Fig.3.

Figure 2

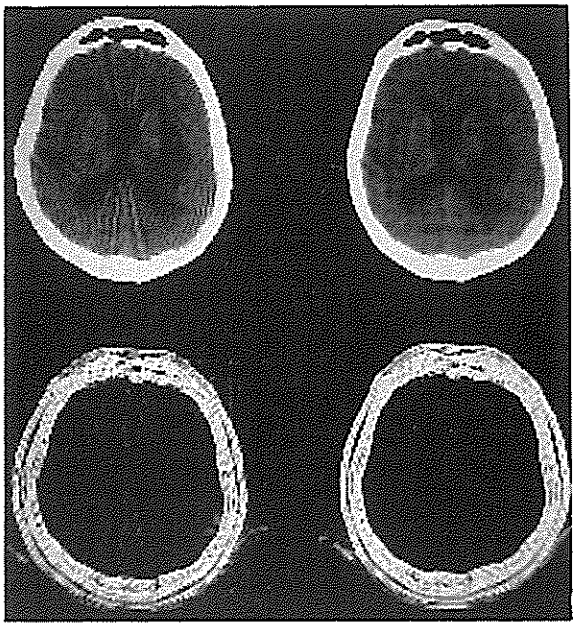


Figure 3

The reconstruction speed has been evaluated on the basis of arithmetic operations expressed as a number of the equivalent multiplications. Such numbers (in millions) for the CT and MR tomography reconstruction (using the same reconstruction parameters as in the previous case) are depicted in Fig.4 and Fig.5. It is evident that the DFM using the hybrid

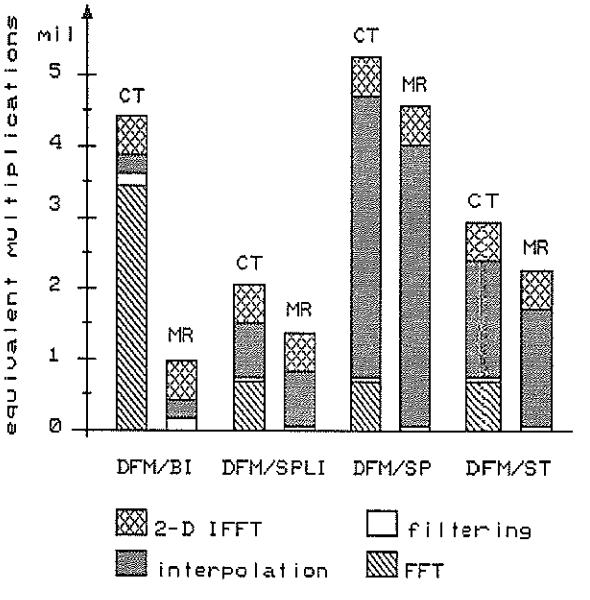


Figure 4

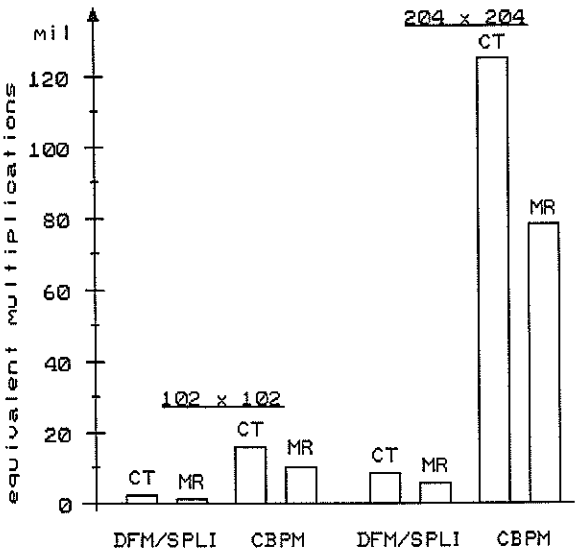


Figure 5

spline-linear interpolation is the most advantageous. For reconstruction of the 102x102 images, the DFM/SPLI method needs approximately 7.8-times less computational time than the CBPM. Further, this rate increases nearly proportionally to the image dimension (the CBPM computational time increases slightly less than 8-times while the DFM/SPLI computational time increases only 4.4 times for the twofold image dimension enlargement). These results have been confirmed by the actual computation time measurements too.

As for memory demands, we established that the differences between the methods investigated are not great.

5. CONCLUSIONS

Using the proposed hybrid spline-linear interpolation yields the reconstruction quality comparable with that obtained in the case of the 2-D spline interpolation, while the reconstruction speed is considerably increased. The interpolation procedure proposed makes possible an implementation of the very fast DFM having the comparable reconstruction quality as the convolution backprojection method has. We suppose that the hybrid spline-linear interpolation represents a convenient candidate for the fast MR methods and for the modern 3-D and dynamic applications in CT tomography. It can be applied also in such cases of general signal/data processing that require the interpolation from a polar-like raster or other regular raster like concentric square raster, fan beam projections raster.

REFERENCES

[1] Herman,G.T., *Image Reconstruction from Projections - The Fundamentals of Computerized Tomography*. Academic Press, New York, 1980.

[2] Ahn,C.B., Kim,J.H. and Cho,Z.H., "High-speed spiral echo planar NMR imaging - I.," *IEEE Trans. Med. Imaging*, vol. MI-5, 1986, pp.2-7.

[3] Mersereau,R.M., "Direct Fourier transform techniques in 3-D image reconstruction," *Comput.Biol.Med.*, vol.6, 1976, pp.247-258.

[4] Stark,H., Woods,J.W., Paul,I. and Hingorani,R., "An investigation of computerized tomography by direct Fourier inversion and optimum interpolation," *IEEE Trans. on Biomed. Eng.*, vol.BME-28, No.7, 1981, pp.496-505.

[5] Matej,S., *A Contribution to the Application of Projective Reconstruction Methods in Tomographic Measurement Systems*. Ph.D. Dissertation. Slovak Academy of Sciences, Bratislava, 1987.

[6] Niki,N., Mizutani,R.T., Takahashi,Y. and Inouye,T., "A high-speed computerized tomography image reconstruction using direct two-dimensional Fourier transform method," *Systems Computers Controls*, vol.14, No.3, 1983, pp.56-65.

[7] Matej,S., "Fast transform methods of image reconstruction from projections and their parallel implementation on SIMD-type computer," in *Computer Analysis of Images and Patterns*. Eds.: L.P.Yaroslavskii, A.Rosenfeld, W.Wilhelmi. Mathematical Research series, Vol.40, Akademie-Verlag, Berlin, 1987, pp.40-47.

[8] Bajla,I., Matej,S. and Bognárová,M., "Computer simulation of the Fourier method of image reconstruction from projections in tomography," in *Advances in Biomedical Measurement*. Eds.: E.R.Carson, P.Kneppo, I.Krekule. Plenum Press, New York, 1988, pp.269-280.

[9] de Boor,C., "Bicubic interpolation," *J. Mat. Physics*, vol.41, 1962, pp.212-218.

[10] Shepp,L.A. and Logan,B.F., "The Fourier reconstruction of a head section," *IEEE Trans. Nucl. Sci.*, vol.NS-21, No.3, 1974, pp.21-43.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

IMAGE RECONSTRUCTION FROM LINE-INTEGRAL DATA: A REGULARIZATION APPROACH

Emanuele Salerno, Anna Tonazzini

Istituto di Elaborazione della Informazione - Consiglio Nazionale delle Ricerche
Via S.Maria, 46, I-56126 Pisa, Italy

The problem of reconstructing images on the basis of very sparse and noisy line-integral data is addressed. The strategy adopted has been that of the standard Tikhonov regularization theory which allows a unique and stable solution to be selected for an ill-posed, ill-conditioned inverse problem. The image is first modeled as a finite Fourier series and then a set of coefficients which have low norm are searched from all those consistent with the data. The basic result obtained was that regularization can improve the quality of reconstructed images with respect to the traditional least squares method for particularly small data sets and low signal-to-noise ratios.

1. INTRODUCTION

The problem of reconstructing images from line integral data is very common and important in image processing applications. As a typical example, we can mention X-ray tomography. In some cases, only a very few data may be available, making the problem extremely underdetermined. This is, for instance, the case when reconstructing temperature fields in furnaces from acoustic time-of-flight measurements (sonic pyrometry) [2]. The reconstruction of images is an inverse ill-posed problem [6], due to the non-uniqueness and instability of the solution. In order to select a solution which is unique and robust against noise (regularization), the class of feasible solutions can be restricted by imposing constraints that exploit additional information on the solution and/or the noise. Standard regularization consists in reformulating the problem as a well-posed, well-conditioned, constrained optimization problem [1,5,6]. In practice, a *cost functional*, or *stabilizer*, representing some measure of *regularity* in the solution, is optimized on the set of images consistent with data. The consistency of the solution will be specified in the form of constraint maps. Under weak conditions on the cost functional and the constraint maps, a unique solution can be found by solving a related unconstrained convex problem in which the cost functional and the constraint maps appear as linearly combined via suitable *regularization parameters.*
To reduce the originally infinite dimensionality of the problem, we first adopted a Fourier parametrization for the solution, obtaining a discrete problem, the unknowns of which are the Fourier coefficients of the model. Under the assumption of uniformly distributed noise with known variance, we considered a constraint set containing the functions which match the data within the variance of the noise. From this infinite set, we selected a unique solution to the problem by optimizing the quadratic cost functional which measures the energy of the solution vector. The performance of the technique was then analyzed as a function of the number of data, the regularization parameter, and the signal-to-noise ratio. In cases where the data set is particularly small, the regularization produces similar effects to that produced by increasing the number of data. This result is interesting because, in practical applications, additional data are often very difficult, or even impossible, to collect.

2. FORMULATION OF THE PROBLEM

The image to be reconstructed is first modeled as a continuous 2D function f(x,y). Let us assume that the available data are related to the unknown function via the following line-integrals:

$$g_k = \int_{l_k} f(x,y)\, dl + n_k, \qquad k = 1, \ldots, N_d \qquad (1)$$

where, for any k, l_k is a straight path contained in the support of f(x,y) (see Figure 1), n_k is the *system noise*, assumed to be independent of f and l_k. This is obviously an inverse problem, easily recognizable as being ill-posed in the sense of Hadamard [6], in that the solution cannot be unique.

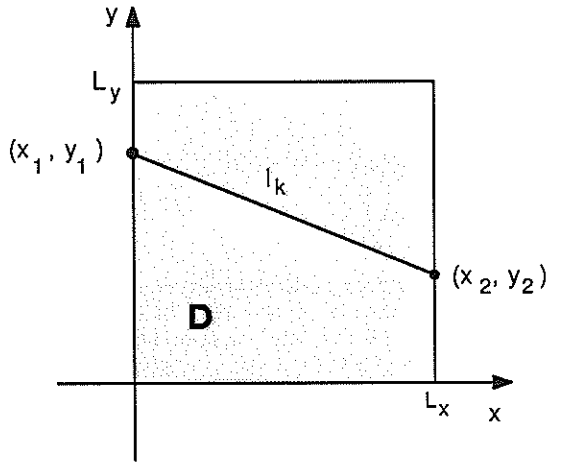


Figure 1

Support region for f(x, y) with the k-th integration path indicated

A way to reduce the set of possible solutions can be the introduction of some a priori knowledge we may have on the problem. For instance, we may know that the true solution can be suitably described by a parametric model, such as a 2D finite Fourier series. Let us suppose, as shown in Figure 1, that the support region of f(x,y) is rectangular, with dimensions L_x and L_y. If we think of f(x,y) on D as part of a function which is symmetric with respect to axes x and y and, furthermore, periodic with periods $2L_x$ and $2L_y$ in directions x and y, respectively, then we can expand it as a double Fourier series with only cosinusoidal terms. Defining the normalized coordinates:

$$u = \frac{x}{L_x} \qquad (2\,a)$$

$$v = \frac{y}{L_y} \qquad (2\,b)$$

we obtain:

$$f(x,y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} A_{ij} \cos(i\pi u)\cos(j\pi v) \qquad (3).$$

If (u_{1k}, v_{1k}) and (u_{2k}, v_{2k}) are the normalized coordinates of the ends of the k-th path and L_k its length then the integrals (1) become [2]:

$$g_k = A_{00}L_k + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} A_{ij}\left[\frac{L_k}{2}\frac{\sin(i\pi u_{2k} + j\pi v_{2k}) - \sin(i\pi u_{1k} + j\pi v_{1k})}{i\pi(u_{2k} - u_{1k}) + j\pi(v_{2k} - v_{1k})}\right.$$

$$\left. + \frac{L_k}{2}\frac{\sin(i\pi u_{2k} - j\pi v_{2k}) - \sin(i\pi u_{1k} - j\pi v_{1k})}{i\pi(u_{2k} - u_{1k}) - j\pi(v_{2k} - v_{1k})}\right] + n_k$$

$$k=1,\ldots, N_d \qquad (4),$$

where the term with i=j=0 is excluded from the summation. On the basis of the expected behaviour of f(x,y), we can limit the maximum orders for i and j, obtaining a finite parametrization for f, with N_p parameters. If we order the double-indexed coefficients A_{ij} so as to form a single indexed vector **a** ($a_m = A_{i_m j_m}$, $m = 1, \ldots, N_p$), Eq. (4) will assume the form of a linear system:

$$\mathbf{g} = \mathbf{H}\mathbf{a} + \mathbf{n} \qquad (5),$$

where the coefficients of matrix H are given by the bracketed terms in (4) and by the lengths L_k. It can be noted that the form of H, and its properties, depend on the integration paths and on the parametrization chosen. If $N_d \geq N_p$, and H^TH is non-singular, then, neglecting the noise, the least squares solution of the system exists and is unique. The possibility of neglecting the noise depends not only on the signal-to-noise ratio (SNR), but also on the condition number of the matrix H^TH. In the present case, for any fixed parametrization, we observed that the condition number of H^TH is highly dependent on the number and position of the integration paths. In particular, any symmetry should be avoided, if at all possible, in order to prevent nearly equal rows and/or columns in matrix H, and consequently a high condition number or even singularity in H^TH. The condition number decreases as the number of integration paths increases; furthermore the integration paths should cover the support region D as uniformly as possible, to avoid any *bias* in the data. If additional a priori knowledge on the behaviour of f(x,y) and/or on statistics of the noise is available, an alternative way to reduce noise sensitivity, keeping the same number of data, is offered by the regularization theory [1,4,5], which makes it possible to solve (5) even if $N_d < N_p$. Following this approach, the problem can be reformulated as a well-posed, well-conditioned constrained optimization problem. In practice, a *cost functional*, representing some measure of *regularity* for f, is

optimized on the set of feasible solutions. In this paper, we assume that the system noise is uniformly distributed around the measurement values with known variance σ^2. This information can be expressed as a constraint relation on the vector **a** to be estimated. If C(**a**) is the cost functional, chosen on the basis of the desired regularity properties for f, the problem may be formulated as:

$$\text{minimize} \quad C(\mathbf{a}) \tag{6 a}$$

$$\text{subject to} \quad \|\mathbf{g} - H\mathbf{a}\|^2 \leq N_d \sigma^2 \tag{6 b}.$$

The constraint (6 b) defines the set of feasible solutions. If C(**a**) is a convex functional then problem (6) has a unique solution which can be computed as the solution of an equivalent unconstrained optimization problem:

$$\min_{\mathbf{a}} \|\mathbf{g} - H\mathbf{a}\|^2 + \lambda\, C(\mathbf{a}) \tag{7}$$

for a suitable non-negative value of the *regularization parameter* λ. In our case, we did not calculate the exact value for λ, but we considered λ as a weight that determines a compromise between the degree of regularity of the solution and its fit to the data. Furthermore, we considered C(**a**) as being a quadratic convex functional of the form:

$$C(\mathbf{a}) = \|B\mathbf{a}\|^2 \tag{8}.$$

Given this assumption, the solution of problem (7) is the following:

$$\hat{\mathbf{a}} = \left(H^T H + \lambda\, B^T B\right)^{-1} H^T \mathbf{g} \tag{9},$$

where the superscript T means the transpose matrix. The results reported here are only relative to the choice, for B, of the identity matrix I; this corresponds to a minimum energy criterion for the solution $\hat{\mathbf{a}}$.

3. PERFORMANCE ANALYSIS

In order to analyze the performance of this reconstruction technique, we fixed a particular test function, exactly parametrizable via a Fourier series with a finite number of coefficients. After the exact calculation of the line integrals along prescribed paths, we added a certain amount of error to generate synthetic data (Eq. 4) to be used for the numerical simulations.
With fixed N_p, $\{i_m, j_m\}$, the performance of the technique was studied as a function of the number of data N_d and of the signal-to-noise ratio SNR, and compared with that of the simple least squares solution of the system (5), corresponding to the choice $\lambda = 0$ in (9). By this approach, we noted that the use of a $\lambda > 0$, with constant N_d, improves the reconstruction in terms of error reduction. If we define as *relative error* the following function:

$$\varepsilon(x,y) = \left| \frac{f(x,y) - \hat{f}(x,y)}{f(x,y)} \right| \times 100 \tag{10}$$

for constant N_p, N_d and SNR we can plot the maximum of ε and its mean value over the support of f, as functions of λ. As examples, we show two diagrams (Figures 2 and 3) obtained with SNR = 28 dB, $N_d = N_p = 15$ (Figure 2), $N_d = 24$ and $N_p = 15$ (Figure 3). The large difference between the errors for $\lambda = 0$ and the errors for small positive values of λ should be noted. The simulations performed for different values of N_d and SNR have shown qualitatively similar behaviours with respect to λ. However, we observed that the benefit of using regularization is more evident for low SNR's, and, for increased N_d, the influence of λ on the errors is continually decreased. The latter was a somewhat predictable result, because data redundancy acts as some sort of regularization.
The choice of the regularizing functional should be made on the basis of prior knowledge on the image to be reconstructed. The performance of different cost functionals, either quadratic or not, such as entropy, is under investigation.

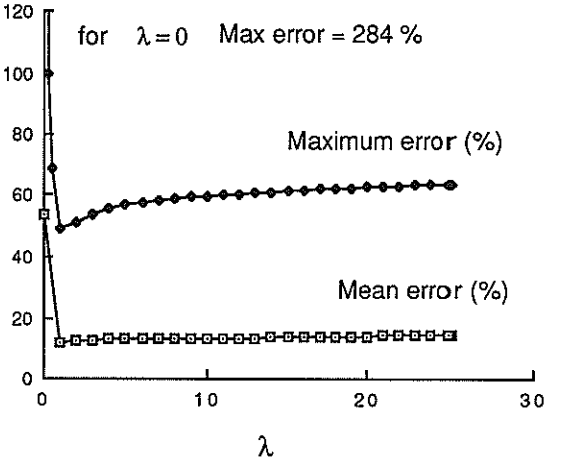


Figure 2

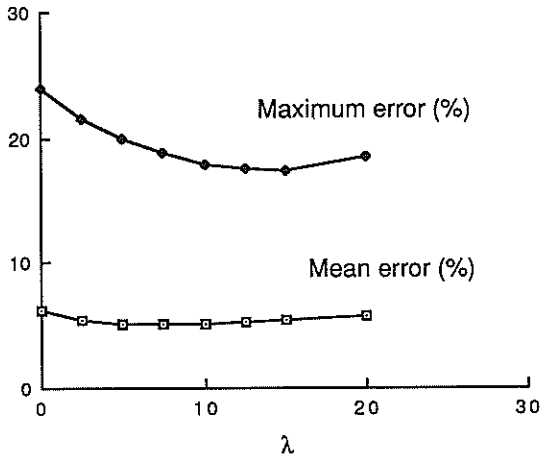


Figure 3

5. CONCLUSIONS

In this paper, we have discussed the reconstruction of images when the data available are sparse and noisy line-integral values. This problem is an ill-posed one, because of the infinite dimensionality of the solution space and the finite dimensionality of the data space. The first step towards its solution was a Fourier parametrization of the function to be reconstructed. However, the problem may still remain singular, and is certainly ill-conditioned. The performance of a particular technique, based on standard regularization theory, has been studied and compared with that of the classical least squares technique. The relative errors of the reconstructed images have been taken as evaluation indexes of the performance. We chose as stabilizer a quadratic functional given by the square norm of the solution vector. The first observation is that the method is not general as it can be applied only when the images to be reconstructed admit a finite parametrization. The method is effective in cases with high noise level and short data sets. The use of regularization can be seen, in some cases, as an alternative to increasing data redundancy for least squares solutions.
Future research on this topic should analyze the performances of different quadratic and non-quadratic cost functionals, and the possibility of applying different parametrizations on the unknown function.

REFERENCES

[1] Bertero, M., Poggio, T.A. and Torre, V., Proc.IEEE **64**, 8 (1988) 869.

[2] Green, S.F., J.Acoust. Soc. Am. **77**, 2 (1985) 759.

[3] Luenberger, D.G., Optimization by vector space methods (Wiley, New York, 1969).

[4] Luenberger, D.G., Linear and Nonlinear Programming (2-nd Ed. Addison-Wesley, 1984).

[5] Poggio, T.A., Torre, V. and Koch, C., Nature **317**, 26 (1985) 314.

[6] Tikhonov, A.N. and Arsenin, V.Y., Solution of ill-posed problems (Winston-Wiley, Washington, 1977).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

MULTICRITERION CROSS-ENTROPY OPTIMIZATION MODEL AND ALGORITHM FOR IMAGE RECONSTRUCTION FROM PROJECTIONS

Wang Yuanmei and Lu Weixue

Institute of Biomedical Engineering, Zhejiang University
Hangzhou, 310027, People's Republic of China

In this paper, a multicriterion cross-entropy optimization approach to image reconstruction from projections is described. We apply a unexplored multicriterion cross-entropy optimization algorithm based on weighted sum scalarization method to solve this problem. Computer simulation is given.

1. INTRODUCTION

The mathematical problem of the image reconstruction is to estimate the picture function from its line integrals. Yuanmei Wang and Weixue Lu have proposed that multicriterion entropy optimization approach to image reconstruction and have obtained many results.

In this paper we attempted to develop a new approach to the reconstruction problem using multicriterion cross-entropy optimization theory. Computer simulation results are given.

2. MULTICRITERION CROSS-ENTROPY MINIMIZATION MODEL ALGORITHM

The principle of minimum cross-entropy provides a general method of inference about an unknown picture density when there exist a prior estimate of this picture and new information about it in the form of constraints. The principle states that,of all the densities that satisfy the constraints, one should choose the posterior with the least cross-entropy. The cross-entropy minimization in the special case of discrete spaces and uniform priors is equivalent to entropy maximization.

It is useful to view a multicriterion cross-entropy optimization as a simultanuous minimization of cross-entropy function of an image, the sum of the local smoothness and peakness functions of an image, and the squared error function between the original projection data set and the projection data due to the reconstructed image. The vector optimization problem of the multicriterion cross-entropy image reconstruction from projections may be discussed as follow:

$$\text{D1.}\quad \min f(x)=(f_1(x),f_2(x),f_3(x)) \qquad (1)$$

where

$$X\triangleq\{x \mid x\in R^n \quad x_j>=0,\ j=1,2..,n \quad \sum_{j=1}^{n} x_j=1\}$$

The cross-entropy function $f_1(x)$ of an image is $\sum_{j=1}^{n} x_j \ln x_j/z_j$. Here $\{z_j\}_{j=1}^{n}$ is a given prior density. The sum of the local smoothness function and peakness function of the image $f_2(x)$ is $\frac{1}{2}\ x^T Sx+\frac{1}{2}x^T x$. Here matrix S is a nonuniformity matrix [3], and α is a positive number choosen experimently to yield the best reconstruction. The squared error function between the original projection set and the projection data due to the reconstruction $f_3(x)$ is $\frac{1}{2}\beta(y-Ax)^T(y-Ax)$. Here A is m n projection matrix, y is a m-dimensional projection data vector, and β is constant to be determined experimently.

Among the many possible ways of obtaining a scalar problem from a vector optimization problem, a common method for finding noninferior solution of a VOP is to convert into a scalar problem using the weighing formulation of the form D2:

$$\text{D2.}\quad \min \sum_{i=1}^{3} w_i f_i(x)=w^T\cdot f(x) \qquad (2)$$

where the row weight vector

$$w\in W=\{w \mid w\in R^3,\ w_j>0 \text{ for each } j=1,2,3, \text{ and } \sum_{j=1}^{3} w_j=1\} \qquad (3)$$

Since the results of solving an optimization model using (2) can vary significantly as the weighting coefficient change, and since very little is usually known about how to choose these coefficients, a necessary approach is to solve the same problem for many different values of w_i. Still, confronted with these solutions, one must then choose among them, presumably on the basis of his intuition. If we want w_i to reflect closely the importance of objectives, all function

should be expressed in unit of approximately the same numerical values. We can also transform (2) to the form:

$$(x)=\sum_{i=1}^{3} w_i f_i(x)\cdot c_i \qquad (4)$$

where c_i must be choosen by the decision maker.

The best results are usually obtained if $c_i = 1/f_i^0$, here $f_i^0 \neq 0$, i=1,2,3.

It follows from the Kuhn-Tucker condition that there exist nonnegative η such that the gradient of the Lagrangian for problem D2.

$$L=w^T f(x) e^T + \eta\left(\sum_{j=1}^{n} x_j - 1\right) \qquad (5)$$

is zero, which implies

$$-\sum_{k=1}^{3} w_3 c_3 \frac{\partial f_k(x)}{\partial x_j} + \eta = 0 \quad , j=1,2,..,n \qquad (6)$$

i.e.

$$\ln x_j = \ln z_j - z_j + w_1^{-1} c_1^{-1} w_3 c_3 \beta\left(\sum_j a_{ji}\left(y_i - \sum_j a_{ij} x_j\right)\right) - w_1^{-1} c_1^{-1} w_2 c_2 \left(\alpha \sum_j s_{ji} x_i + x_j\right) + w_1^{-1} c_1^{-1} \eta \qquad (7)$$

Let

$$v_j(x) = w_1^{-1} c_1^{-1} w_3 c_3 \beta\left(\sum_j a_{ji}\left(y_i - \sum_j a_{ij} x_j\right)\right) - w_1^{-1} c_1^{-1} w_2 c_2 \left(\alpha \sum_j s_{ji} x_i + x_j\right) - z_j \qquad (8)$$

$$\zeta = w_1^{-1} c_1^{-1} \eta$$

Then Eq.(7) may be written

$$\sum_j x_j = 1 \quad , j=1,2,..,n \qquad (9)$$

Hence

$$x_j = \frac{z_j \exp\{v_j(x)+\zeta\}}{\sum z_j \exp\{v_j(x)+\zeta\}} \qquad (10)$$

where ζ has been selected as that

Notice that this is always nonnegative so that the inequality constraint $x_j \geq 0$ is satisfied.

We see that we have solved the jth nonlinear equation explicity for the jth unknown and put everything else on the right side. The Jacobi method is to guess at x, say x^0, and do the iteration.

$$x_j^{(k+1)} = \psi_j\left(x_j^{(k)}\right), \quad j=1,2,..,n \qquad (11)$$

where

$$\psi_j(x) = \frac{z_j \exp\{v_j(x)\}}{\sum_j z_j \exp\{v_j(x)\}} \qquad (12a)$$

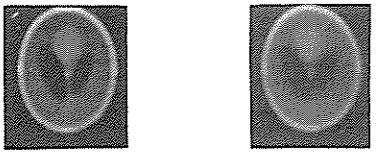
$$\psi(x) = (\psi_1(x), \ldots, \psi_n(x))^T \qquad (12b)$$

3. COMPUTER SIMULATION RESULTS

The effectivebess of the multicriterion cross-entropy minimization and algorithm was shown by examples of reconstruction for Shepp-Logan head model phantom. A comparision to single objective optimization methods were given. We have used a mormalized mean root squared error measure for evaluation of the different algorithms. The simulation results are reported in Table 1 and Fig. 1. For each algorithm the number of iteration is shown.

Table 1

Algorithm	NMS error	Number of iteration
Kashyap and Mittal	0.2532	6
Minerbo–MENT	0.4312	8
Multicriterion cross–entropy	0.09231	4



(a) (b)

Fig. 1. Head Phantom and Multicriterion Cross-Entropy Reconstruction with 114 x 114 Digitization.

We can see from simulation results that multicriterion cross-entropy minimization method performs better than the other single criterion methods.

REFERENCES

[1] Wang Yuanmei and Lu Weixue, Multicriterion Image Reconstruction and Implementation, Computer Vision, Graphics and Image Processing, Vol.46,(1989), pp.131-135.

[2] Wang Yuanmei and Lu Weixue, Multicriterion Maximum Entropy Approach to Image Reconstruction, Signal Processing, Theory and Applications (IV), edited by J.L.Lacoume, Elsevier Science Publisher B.V. (1988).

[3] R.L.Kashyap and M.C.Mittal, Picture Reconstruction from Projections, IEEE Trans. on Computer, Vol.C-24, No.9,(1975), pp.914-925.

[4] G.Minerbo, MENT: A Maximum Entropy Algorithm for Reconstructing a Source from Projection Data, Computer Graphics and Image Processing, Vol.10,(1979),pp.48-68.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

SYMBOLIC AND NUMERIC DATA FUSION FOR THE THREE-DIMENSIONAL RECONSTRUCTION OF VASCULAR NETWORKS

M. GARREAU, J.L. COATRIEUX, R. COLLOREC, C. CHARDENON

Unité INSERM 335 - Laboratoire Traitement du Signal et de l'Image
Université de Rennes I - Campus de Beaulieu - 35042 RENNES CEDEX - FRANCE

A framework is proposed to reconstruct in 3-D the coronary artery network from two X-ray views. The main feature consists in merging algorithmic procedures (aimed at the vessel detection, the matching and the reconstruction) together with a qualitative symbolic model of the object. The processing steps involved in the process are reviewed. The model generation and its description are then depicted. Preliminary results obtained from simulation and phantoms data are finally provided.

1. INTRODUCTION

Two main imaging modalities provide information on the conformation of vascular network. The most recent means come from Magnetic Resonance Imaging (MRI) which allows to image the vessels from the blood movement. The recent advances show that accurate descriptions can be expected in the future. The more traditional way to analyse arterial and veinous trees comes from Digital Subtraction Angiography (DSA). High resolution devices are now available and lead to significant improvements in structural and functional state assessment. Scan sections (MRI) and projection views (DSA) suffer from the same drawback : they do not provide a full access to the organs in a three dimensional (3-D) space. In the last two decades, a number of methods have been proposed to reconstitute 3-D data sets from parallel slices. They range from surface (polygon and voxel based) and volume rendering [1][2]. However a true 3-D reconstruction is not straightforward when dealing with a set of 2-D X-ray projections of vascular network because of the limited number of views : contrast medium propagation and organ movement reduce the temporal interval when the image can be recorded. These reasons have increased the interest of matching methods based on Computer Vision principles [3][4] although computed tomography approaches can be used [5][6][7].

The method here described departs from the previous ones by the fusion of data related to the image information and of a high level model of the object. The reconstruction and labelling process is directly stated in 3-D when the previous attempts, making use of knowledge based systems, focused on 2-D labelling (see the exemplary work of [8]).

2. METHODOLOGICAL FRAMEWORK

The overall scheme is shown Figure 1. It includes :

a) Spatio temporal image sequence acquisition

This stage deals with the correction of distortions, spatial calibration and image preprocessing. In order to have accurate geometrical matching and reconstruction, the distortions produced by the imaging device must

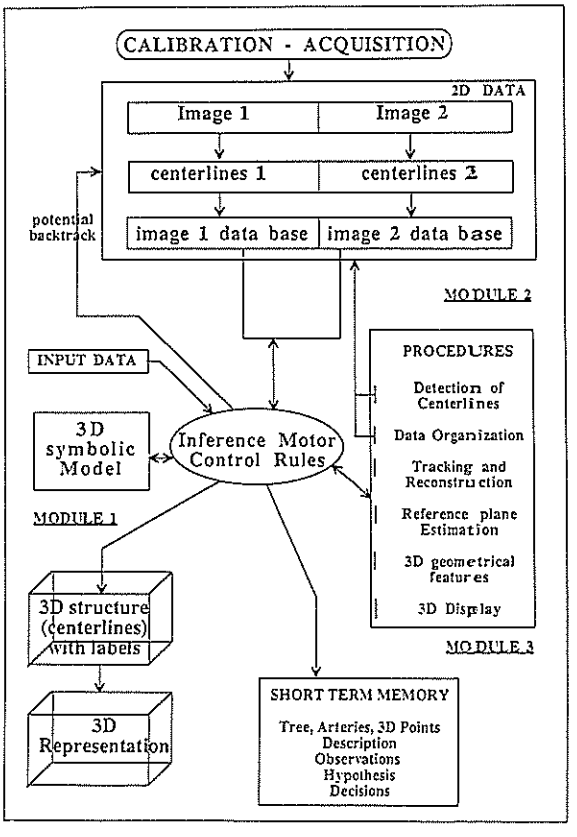


Figure 1

be minimized. The most important one comes from the image intensifier curvature. It can be corrected by imaging a rectangular grid and an a posteriori elastic transformation to recover the initial grid shape. The spatial calibration is performed using a set of markers of known locations inside a cube. The geometrical relations between 3-D points and their 2-D projections are then estimated :

$$t\begin{pmatrix} u_1 \\ v_1 \\ 1 \end{pmatrix} = M_1 \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} ; \qquad t\begin{pmatrix} u_2 \\ v_2 \\ 1 \end{pmatrix} = M_2 \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$

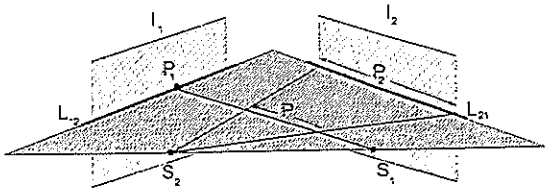
where :
x,y,z : 3-D point coordinates,
M_1 and M_2 : the calibration matrices relative to the two projection planes,
u_1, v_1 and u_2, v_2 : the corresponding 2-D coordinates of this 3-D point.

The image preprocessing corresponds to the subtraction of images recorded before and after contrast medium injection which enables to enhance the vessel network from the background (other interfering tissues). When the organ is submitted to movement, this stage is crucial. Additional preprocessing can be carried out to reduce the low frequency contents of the background.

b) Detection of the arterial tree
Despite the number of works devoted to the segmentation of vessels, ranging from syntactic method, mathematical morphology up to knowledge based approaches (see [9][10] for a review), precise, robust and complete delineation of branches remain an open problem. The method here applied is based on the search of seed points located on the centerlines, the tracking of the vessels and the control of the boundary detection all along the centerlines [11][12].

c) 3-D reconstruction (Figure 2)
This task can be decomposed into two subtasks : (1) find two corresponding points in each image; (2) reconstruct the 3-D point which has these two projections. The first one is equivalent to a matching problem and can be solved by means of epipolar constraint. When several candidates are detected, the right solution can be provided by using local similarities (see [4] for a detailed review). When these two points have been paired, the reconstruction is specified by the intersection of the two epipolar lines formed by the points and the source focuses (some approximation is usually necessary).

d) 3-D representation
Once the three dimensional centerlines of the arteries have been obtained, the local estimations of cross sectional lumen in shape and size can be achieved using the edges already detected in the two views. Well known techniques to render 3-D data sets can be used : surface approximation (concatenation of cylinders and polygonal description) as well as ray tracing or octree encoding for volume (voxel based) display (refer to [13][1] for a full survey)
The previous paragraphs identified the procedures and the resulting 2-D and 3-D data sets involved in the reconstruction method. The next section describes the symbolic model and the resolution process.



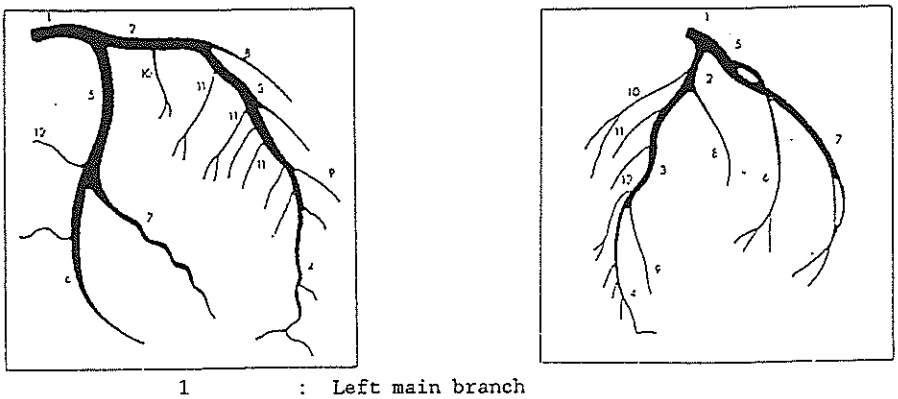
I_1, I_2 : projection planes
S_1, S_2 : focuses of sources
P : 3-D point
P_1, P_2 : 2-D projections of P
L_{12}, L_{21} : epipolar lines in I_1, I_2

Figure 2

3. RECONSTRUCTION AND LABELLING

The 3-D coronary network (Figure 3) has been modelled in a qualitative way.

Structural relations (originated and originating arteries) and geometrical properties on relative positions and orientations are defined by means of a tree-like description whose elements are the vessels. Geometrical informations are referenced in relation to typical planes containing the two main arteries (Circonflex and Anterior Inter-Ventricular) which, when estimated from the patient specific data base, will serve as the common coordinate system. These two planes delimitate the subspaces within all the branches can be derived. The type of distribution (balanced, right and left dominant) has been introduced to deal with individual variants.The knowledge base has been elaborated in part from the design of a 3-D graphical generator. This tool enables the expert to draw the vascular tree in a wire like mode. So the knowledge engineer can identify the underlying relations and the order of the expert 's reasoning. Segments of vessels and branches can be built , updated, removed and connected to other branches. The validation is performed by projection of the 3-D construction onto conventional incidences. Inconsistencies can be retracted by coming back to the 3-D space[14].



1 : Left main branch
2 - 3 - 4 : Anterior inter-ventricular artery
5 - 6 : Circonflex artery
7 : Left main lateral artery
8 - 9 : diagonal arteries
10 - 11 : septal arteries
12 : left atrial artery

Figure 3

The reconstruction-labelling process can now be stated as follows :

a) initialization : the analysis of the two views provides an ordered list of candidate segments (a segment is a set of connected centerline points delimited by decision nodes, see below) for the tree root. The ranking can be defined from the mean grey level along a segment or from the width associated to it. The best candidate is chosen to initiate the reconstruction.

b) reconstruction : the reconstruction of each branch is performed through 3-D segments using epipolar constraints. A 3-D segment is a set of 3-D points built from the pairing of two 2-D segments extracted in the image planes.

c) decision node processing : the decision nodes identify branching points, crossing points and extremities of vessels. When the reconstruction tracking arrives to a decision node in one (or two) image(s), a set of hypothesis is generated in the following order :

- identification of the nature of the point (data driven : first hypothesis generated by 2-D configurations and model driven : hypothesis ordering of possible interpretation)
- pairing of successor segments (if the decision node is not depicted as end of branch)
- label allocation of the two 3-D segments to reconstruct from the node according to the previous matching.

d) iteration and reference plane estimation : the first estimation of the reference planes is biased because of the very limited number of 3-D points belonging to AIV and CX arteries. They must be refined iteratively as soon as new 3-D data have been derived. This update can take place at each decision node detected on their pathway. A least square approximation is run to re-estimate the planes based on the set of available points.

The steps b, c and d are then reiterated except when a failure occurs.

e) backtracking : Inconsistent matching between data and model is carried out at the decision node level. Failure can occur in the node naming or in the subsequent pairing and labelling. It induces the reexamination of the previous deductions by backtracking. This process consists in selecting the remaining alternatives associated to the anterior decision node (with the inverse order labelling-pairing -node naming). In case of persistent inconsistency, the traversal of the space search continues and if necessary up to the tree root (evoking the next candidate of the initial list).

4. RESULTS

This method has been implemented on a HP9000 computer in D-Prolog and procedural programs written in Fortran. The validation has been carried out on simulated data (error free), line like phantoms (to be free of the detection errors) and on tubular like phantoms made of flexible tubes filled with usual contrast medium. These two last phantoms have been designed according to the heart anatomy and arranged on an elliptical shaped support. The results obtained from the tubular phantoms are reported here.

The images (resolution 512*512 and dynamic range 8 bits) were obtained from a Digitron-Siemens through the digitalization of the analog output. The reconstruction-labelling process

has been applied to the views "left Transversal" and "left Anterior Oblique" with axial angles 60°, 20°. The resulting 3-D structure has been backprojected onto a third incidence (left Anterior Oblique (LAO) with axial angle 60°). Precise geometrical description of the tubular phantom was not available and the accuracy of the reconstruction has been estimated on this third image (LAO 60°). Centerlines deviations ranging from 1 to 3 pixels have been observed. The small angulation of the two incidences as well as the uncertainties should explain these local errors. They are not related to the knowledge part but linked to all the procedural routines (calibration, detection, reconstruction). Before volume rendering, a smoothing, by means of B-splines, has been applied to the 3-D centerlines. Instead of applying classical scheme to produce synthesis images, a specific representation has been adopted : a sphere centered and rolling on the centerlines, whose diameters are estimated from the 2-D contours detection, provides the cylinder like structure of the vessels.
The surface normal computation, needed for shading, is in this case straightforward. Clearly the other techniques reported in [2] can replace this rough representation.

5. CONCLUSION

An approach has been proposed to reconstruct the coronary arteries from biplane technique. The main difference with previous works is that the labelling operates simultaneously to the reconstruction and it is directly stated in 3-D. The critical steps involved in the scheme have been emphasized : the precision in reconstruction is highly dependent on the distortion, the calibration, the angulation between views and the vessel detection : the symbolic and numeric data fusion relies on the availability of a structural and geometric model (taking into account the interindividual variations). Moreover the matching of the model and the image data is only possible through a common coordinate system. The main additional requirement is that the two image planes have to be recorded at the same instant to avoid any spatial shift of the moving structure.

REFERENCES

[1] Coatrieux, J-L., Three-dimensional medical imaging, Special Issue, IEEE Engineering Medecine Biology Society Magazine, Nov. 1990, to appear.

[2] Barillot, C. and Gibaud, B. and Lis, O. and Luo, L. and Bouliou, A. and Le Certen, G. and Collorec, R. and Coatrieux, J-L., CRC Critical review in Biological Engineering, 15, 4, 1988, pp. 269-307.

[3] Venaille, C. and Mischler, D. and Coatrieux, J-L. and Catros, J-Y. , Reconstruction tridimensionnelle de réseaux vasculaires en angiographie, 7ème congrès AFCET/INRIA, Tome 3, pp. 1533-1547, Paris, Nov. 1989.

[4] Faugeras, O.D., Quelques pas vers la vision artificielle en trois dimensions, TSI, 7, 6, 1988, pp. 548-590.

[5] Grangeat, P., Analyse d'un système d'imagerie 3-D par reconstruction à partir de radiographies X en géométrie conique, Thèse de doctorat ENST, 1988.

[6] Trousset, Y. and Saint-Felix, D. and Rougée, A. and Chardenon, C., Multiscale Cone-Beam X-Ray reconstruction, SPIE Medical Imaging IV, Newport Beach, Feb. 1990.

[7] Hamon, C. and Roux, C. and Coatrieux, J-L. and Collorec, R., An analytical method for 3-D tomographic reconstruction from a small number of projections using a simple interpolation scheme, Proc. Conf. IEEE ASSP, Albuquerque, 1990.

[8] Smets, C. and Verbeeck, G. and Suetens, P. and Oosterlinck, A., A knowledge-based system for the Three-dimensional reconstruction of the cerebral blood vessels from a pair of stereoscopic angiograms, in Pattern Recognition and Artificial Intelligence, Gelsema Es and KANALL-N Eds, Elsevier, North Holland, 1988, pp. 425-435.

[9] Toumoulin, C. and Collorec, R. and Coatrieux, J-L., Vascular network segmentation in subtraction angiograms : a comparative study, Medical Informatics, 1990, to appear.

[10] Garreau, M. and Coatrieux, J-L. and Collorec, R. and Chardenon, C., A knowledge based approach for 3-D reconstruction and labeling of vascular network from biplane angiographic projections, IEEE Med. Imaging, to appear.

[11] Garreau, M., Signal, Image et Intelligence Artificielle. Application à la décomposition du signal électromyographique et à la reconstruction et l'étiquetage 3-D de structures vasculaires, Thèse, Université de Rennes I, Oct. 1988.

[12] Collorec, R. and Coatrieux, J.L., Vectorial tracking and directed contour finder for vascular network in digital subtraction angiography, Pattern Recognition Letters, 8, pp. 353-358, 1988.

[13] Coatrieux, J-L and Barillot, C., A survey of 3-D display techniques to render medical data, NATO Advanced Research Workshop on 3D Imaging in Medicine, Travemünde, Springer Verlag Ed., June 1990, to appear.

[14] Coatrieux, J-L. and Garreau, M. and Collorec, R. and Carrault, G., Signal, Image et Intelligence Artificielle en Médecine, Proc. Entretiens de Lyon, Springer-Verlag pp. 264-279, Ed., 1988.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

An Accuracy Model for Binary Pattern Reconstruction from Projections

Yongjian Bao[1]
Deutsches Herzzentrum Berlin
Augustenburger Platz 1, 1000 Berlin 65, FRG

Abstract

The central point of the model-based methods for binary image reconstruction from two projections is to acquire a reference model. A method is presented in this paper to derive an accuracy model from an image with several grey levels. The method has been used in a multi-resolutional reconstruction technique to estimate the accuracy models with high resolution. The preliminary simulation experiments have shown that a considerable improvement to the reconstruction results is achieved with this method.

I Introduction

Recently, the techniques of tomographic image reconstruction from incomplete projection data have received rapidly growing attention, since these render an extension of tomographic imaging in more wide application fields possible [**Reeds and Shepp 1987**]. Under some circumstances, only a few views of projections are available due to, e.g., limited imaging time, poor imaging device, limitation of geometrical position, etc. On the other hand, however, a reconstruction task may be simply to estimate the tomographic shape rather than reconstructing a high quality image such as the commercial CT images. Therefore, the techniques of reconstruction from a few views are required. These techniques have been proposed and investigated from the conventional transform-based method [**Reeds and Shepp 1987**], the stochastic estimation method [**Prince and Willsky 1989**], [**Bresler and Macovski 1989**] and the algebraic method [**Herman 1974**], [**Slump and Gerbrands 1982**], [**Reiber *at al.* 1983**], respectively to different applications.

In modern cardiological laboratories biplane X-ray system has been employed for the cardiovascular investigations, by which two X-ray angiograms in the different directions can be simultaneously acquired. An estimation of the cross-sectional shape of the coronary arteries from the two angiographic projections may be very useful for the vascular stenostic diagnosis and hemodynamic study [**Reiber *at al.* 1983**], [**Kenet *at al.* 1987**]. Because of injection of an X-ray sensitive contrast agent and an appropriate background removing, an arterial cross section can be assumed as a binary pattern, which is represented by a binary matrix of pixel value 1 being the pattern and pixel value 0 the background. As described above, there only two projections are available, which are adjusted to be orthogonal to each other. Under these assumptions, the projection equations (line integrals) are reduced to sum the number of pixels of value 1 in row and column directions of the binary matrixes. Now, the reconstruction problem can be stated to estimate a binary matrix **G**

$$g(i,j) = \begin{cases} 1, \text{ if the pixel (i,j) lies inside of the pattern,} \\ 0, \text{ otherwise,} \end{cases} \quad (1)$$

from its row and column sums:

$$\sum_{j=1}^{N} g(i,j) = p_y(i) \quad (2)$$

$$\sum_{i=1}^{M} g(i,j) = p_x(j) \quad (3)$$

$$1 \le i \le M,\ 1 \le j \le N$$

where the vectors $\mathbf{P}_x$ and $\mathbf{P}_y$ are two projections, respectively.

The problem (1)-(3) is underdetermined because the number of unknown variables (M·N) exceeds the number of equations (M+N-1). Because we have only two projections, the transform-based method is impossible to be applied. To estimate an arbitrary cross-sectional shape, the stochastic method is also not adequate due to the difficulty of modelling the shape with parameters. A algebraic method based directly on the equations (2) and (3) is flexible to incorporate a priori knowledge about the binary patterns into the reconstruction process and therefore is employed by many researchers to deal with the problem. The early researchers have tried to directly process the problem (1) - (3) and failed to obtain any reliable results under only two projections [**Herman 1974**]. In order to compute a unique solution to the problem, a priori knowledge has to be incorporated.

II The Model-Based Method

Actually, there are many solutions satisfying the equations (1) - (3). In this sense the problem does not fulfil the Hardmad's conditions [**Demoment 1989**] and therefore is ill-posed. In order to solve the problem, a regularization is required. If we have a priori binary pattern **M** as a reference pattern, a simple way of performing a regularization is to minimize a distance between the reconstructed pattern **G** and the reference pattern **M**. The regularized problem can be now formulated in the following:

1: Author's correspondence address: Yongjian Bao, Sekr. FR 3-3, CG/FB 20, Technische Universität Berlin, Franklinstr. 28-29, 1000 Berlin 10, FRG.

minimize { $J(\mathbf{G}, \mathbf{M})$ }, (4)

subject to the constraints in (1) - (3).

Generally speaking, $J(\cdot,\cdot)$ can be viewed as a penalty function defined between two patterns. Slump *et al.* (1982) proposed an approach to maximize the shape resemblance between the reconstruction and the reference model. First, a cost matrix **C** is derived from a priori model **M**, in order to define the penalty value c(i,j) for each matrix position (i,j), $1 \le i \le M$ and $1 \le j \le N$, of being a pattern pixel, i.e., taking the value 1. Then the distance between **G** and **M** is defined to be the total penalty value of the reconstructed pattern **G**:

$$J(\mathbf{G},\mathbf{M}) = \sum_{i=1}^{M} \sum_{j=1}^{N} g(i,j) \cdot c(i,j). \quad (5)$$

The equations (1) - (5) thus described a regularized binary pattern reconstruction from two orthogonal projections. The integer programming techniques have been efficiently employed to solve the problem **[Slump and Gerbrands 1982]**, **[Reiber *et al.* 1983]**, **[Bao 1989]**. Actually, the reconstructed shape is essentially determined by the regularization (4) and (5) because of the large ambiguity in the equations (1) - (3). In this paper we will present a new method of computing the cost matrixes.

III Computation of Cost Matrix

An original method of computing cost matrix **C** from a given binary model **M** was presented in **[Slump and Gerbrands 1982]**, which can be described as a three steps procedure. First, all c(i,j) is set to negative if the pixel (i,j) lies inside of the model pattern, otherwise to zero:

$$g(i,j) = \begin{cases} -1, \text{ if m(i,j) lies inside of the model pattern,} \\ 0, \text{ otherwise.} \end{cases} \quad (6)$$

Second, set k = 0 and start an iteration: For all c(i,j)=k, compute a new penalty value for it:

$$c(i,j) = 8 + k - \sum_{p,q \in \mathcal{N}_8} \varepsilon(k - c(p,q)), \quad (7)$$

where $\mathcal{N}_8$ is a 8-neighbourhood of pixel (i,j) and $\varepsilon(\cdot)$ denotes the indicator function:

$$\varepsilon(i,j) = \begin{cases} 1, \text{ if } t > 0, \\ 0, \text{ if } t \le 0. \end{cases} \quad (8)$$

Then, k is replaced by k+8. If there is no any element c(i,j) = k, all elements of **C** have received a final penalty value, otherwise for those elements c(i,j) = k, a new penalty value is further computed with Eq.(7).

Lastly, set all negative c(i,j) to zero:

$$c(i,j) = 0, \text{ if } c(i,j) = -1. \quad (9)$$

The method turns out to be a "core-expending" procedure which is conceptually depicted in Figure 1.

The method was designed for computing cost matrix from the binary model patterns **[Slump and Gerbrands 1982]**. In a multiresolutional binary image reconstruction technique, we want to estimate a cost matrix from an intermediate reconstruction result at a lower spatial resolution than an original model, but with several grey levels **[Bao *et al.* 1990]**. If a thresholding is directly applied to these images, only a very rough cost matrix can be computed with Eq.(6)- (9), because of poor resolutions. The model-based method simply described in section III is an optimizing process under conditions. With this point in mind, we want to compute a cost image both of preserving the detail features of the model and being smooth, otherwise the optimizing may give rise to a undesired result.

VI An Accuracy Model

As discussed in the last section, the accuracy of the cost matrixes are essentially determined by the accuracy of the model's contour. From this point of view, we first interpolate a grey level model into high resolution and then threshold it. A cost matrix is thus computed from the thresholded binary model at a high resolution simply with Eq.(6) - (9). Lastly, we map the cost matrix again back to a lower resolution to control the optimizing.

In our use, a 2x2 non-overlapped pyramid was selected to a multiresolutional binary image reconstruction **[BAO *et al.* 1990]**. Both the interpolation and the back-mapping are carried out in the pyramid, i.e., via the pyramid operations **EXPAND** and **REDUCE**. The operation REDUCE has been described elsewhere **[BAO *et al.* 1990]**, here only the EXPAND is discussed. To interpolate an image into high resolution, a sinc function of the following form

$$h(x,y) = h_s(x) \cdot h_s(y) = \{\sin(\pi x/\Delta x)/(\pi x/\Delta x)\} \cdot \{\sin(\pi y/\Delta y)/(\pi y/\Delta y)\} \quad (10)$$

is employed as the interpolating kernel due to its elementary response function in frequency domain, where Δx and Δy are sampling distances. Because of the geometric 2x2 non-overlapped structure, an EXPAND should interpolate an image into a double resolution in a fashion depicted in Figure 2. Evidently, all new image pixels will be interpolated because the pixels of an image at a high level lie always in-between the image pixels at the level below it. To interpolate these image pixels, both the kernels $h_s(x)$ and $h_s(y)$ must be asymmetrically sampled for the pixels positioned in even and odd grid points, respectively. Let w(p,q) denote the digital sampled 4x4 kernel and a second subscription denote this position correspondence, i.e., $w_{s,0}$ for the even grid positions and $w_{s,1}$ for the odd grid positions, the operation EXPAND can represented by the convolution:

$$g_l(i,j) = \sum_{p=0}^{3} \sum_{q=0}^{3} w_{s,(i \bmod 2)}(p) \cdot w_{s,(j \bmod 2)}(q) \cdot g_{(l+1)}(i'/2 - 2 + p, j'/2 - 2 + q), \quad (11)$$

$i' = i + i \bmod 2$ and $j' = j + j \bmod 2$,

where l and $l+1$ indicate the pyramidal levels and $0 \le l < L$ (L is the top of pyramid). Applying EXPAND continuously, we can expend an intermediate reconstruction result at a higher pyramidal level into to an arbitrary lower level, i.e., at a higher resolution. In a desired resolution, the expended image is thresholded into a binary model. The method described in section III can be directly applied to the model for computing a cost matrix. This cost matrix is reduced to a desired resolution (pyramidal level) to continue the pyramid reconstruction process. Certainly, the required resolution of the cost matrix must not be the same of the intermediate reconstructed image, from which the cost matrix is computed. Practically, we want to estimate a cost matrix from the intermediate result to improve the actually used cost matrix at the level below it. We would discuss the pyramid reconstruction technique in further details elsewhere **[Bao 1990]**

V Results

In Figure 3, we showed our result by comparing the use of cost matrix estimation and improvement with a unchanged cost matrix in binary image reconstruction, where a three-levels pyramid was applied. The display window is divided into three column, we call them column 2, 1, and 0 corresponding to pyramidal level 2, 1, and 0. The pyramidal level 0 is the bottom of pyramid and corresponds to the original resolution of the given projection data and a priori given cost matrix.

The reconstruction without improving cost matrix is shown in Figure 3(a). Three images at the top of each column are the pyramid reconstruction results at level 2, 1, and 0, respectively. Note the result in level 0 is already binary and yields the final reconstructed result. In the middle of each column, the cost images actually used at these levels are displayed with a inverse look-up-table transform. In this example, they are simply the pre-computed reduced copies of the original cost image, which is shown at the bottom of column 0, and not changed during the reconstruction. In Figure 3(b) we showed a result of the same projection data and the same priori cost matrix but with an estimation and improvement of cost matrix discussed in section IV. As shown, two additional cost images are displayed at the bottom of column 2 and 1. The cost image in column 2 is estimated from the intermediate reconstruction result at pyramidal level 2 by using the method discussed in the last section. This cost matrix is used to improve the pre-computed reduced copy of the priori cost image. This results in a new cost matrix at level 1, which is actually employed to control the reconstruction at this level. The new cost matrix is displayed in the middle of column 1. Analogously, the cost image at the bottom of column 1 is estimated from the intermediate result at level 1 and used to improve the cost image at level 0. A final reconstructed result is displayed at the top of column 0. It is evident that this result is considerably improved than the result in Figure 3(a).

In the case of a binary model, the method of section VI can be still applied in some extent. An example is given in Figure 4. Figure 4(a) is the same result of Figure 3(a). To simplify the illustration, only the final result at the pyramid bottom is given. In figure 4(b), the priori given binary model is first expended into a double spatial resolution by means of EXPAND in Eq.(11) and then thresholded by a thresholding value 0.6. A cost image is computed by applying Eq.(6) - (9) to the expended model and reduced back to the original resolution. A such processed cost image is used to control the optimizng in Eq.(4). The result and the cost image are displayed in Figure 4(b), respectively. As can be observed, this result is indeed better than the result of Figure 4(a).

IV Conclusions

An accuracy model estimation method has been presented. It was designed in our multiresolutional algorithm for binary image reconstruction. From a grey level image, an accuracy binary model may be estimated, which is in turn used to compute an accuracy cost matrix by simply applying Eq.(6) - (9). Through estimating and improving the cost matrixes in a pyramid reconstruction process, a considerable improvement to the reconstruction results has been achieved.

The method can also be applied to a binary model. In this case, the effect is actually smoothing a cost matrix. Because the model-based reconstruction method is based on a conditioned optimizing, a smooth cost matrix is helpful to make the optimizing process continuous in a sense of the geometric shape of the reconstructed patterns.

IIV Acknowledgments

This work is supported by the Deutsche Forschungsgemeinschaft. The author thanks H.U. Lemke, Fachbereich Informatik, Technische Universität Berlin for his supports. I would also like to thank S. H. Stiehl, Fachbereich Informatik, Universität Hamburg for his constructive discussions and suggestions.

IIIV References

[Bao *et al.* 1990]
Y. Bao, H. Oswald, and E. Fleck, The Pyramid Structure and Its Applications to Binary Image Reconstruction from Two Projections, in **Proc. IEEE CASSP-90**, April 3. - 6. 1990, Albuquerque, U.S.A.

[Bao 1990]
Y. Bao, On Multiresolutional Reconstruction of Binary Image from Two Projections, Technical Report of FB 20, TU Berlin, being prepared.

[Bresler and Macovski 1987]
Y. Bresler and A. Macovski, Three-Dimensional Reconstruction from Projections with Incomplete and Noisy Data by Object Estimation, **IEEE Trans. Acoustic, Speech and Signal Processing**. Vol.ASSP-35(8), 1987, 1139 - 1152.

[Demoment 1989]
G. Demoment, Image Reconstruction and Restoration: Overview of Common Estimation Structures and Problems, **IEEE Trans. Acoustic, Speech and Signal Processing**, Vol.ASSP-37(12), 1989, 2024 - 2036.

[Herman 1974]
G.T. Herman, Reconstruction of Binary Pattern from a Few Projections, **Proc. International Computing Symposium 1973**, Ed. by A. Günther et al., Horth-Holland, Amsterdam, 1974, 371 - 379.

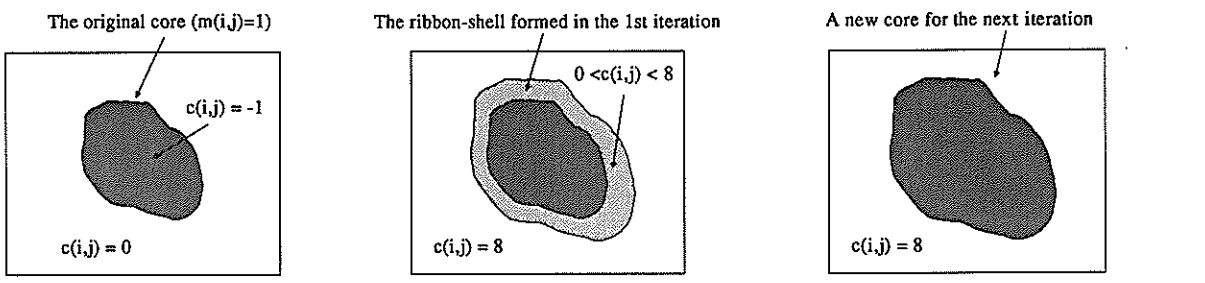
[Kenet *et al.* 1987]
R.O. Kenet, E.M.Herrold, J.P. Hill, J.Waltman, A. Diamond, P.Fenster, J. Barba, M. Suardiaz, J.S. Borer, Reconstruction of Coronary Cross-Sections from Two Orthogonal Digital Angiograms, **IEEE Computers in Cardiology 1986**, IEEE Press, 1987, 273 - 276.

[Prince and Willsky 1989]
J.L. Prince and A.S. Willsky, A Hierarchical Algorithm for Limited-Angle Reconstruction, **Proc. IEEE ICASSP-89**, Vol.3, 1989, 1468 - 1471.

[Reeds and Shepp 1987]
J.A. Reeds and L.A. Shepp, Limited Angle Reconstruction in Tomography via Squashing, **IEEE Trans. Medical Imaging**, Vol.MI-6(2), 1987, 89 - 97.

[Reiber *et al.* 1983]
J.H.C. Reiber, J.J. Gerbrands, G.J.Troost, C.J.Kooijman, and C.H. Slump, 3-D Reconstruction of Coronary Arterial Segments from Two Projections, in **Digital Imaging in Cradiovascular Radiology**, Ed. by P.H. Heintzen and R. Brennecke, Georg Thieme Verlag, 1983, 151 - 163.

[Slump and Gerbrand 1982]
C.H. Slump and J.J. Gerbrand, A Network Approach to Reconstruction of the Left Ventricle from Two Projections, **Computer Graphics & Image Processing**, Vol.18, No.1, 1982, 18 - 36.



(a) A binary model pattern is copied into the cost matrix as an original core

(b) In the 1st iteration, the pixels 0 < c(i,j) < 8 constitute a ribbon-shell surrounding the core.

(c) Adding the shell to the original core a new core is formed for the next iteration.

Figure 1. Computation of Cost Matrix from a Binary Model Pattern

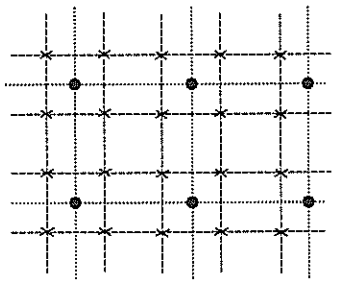
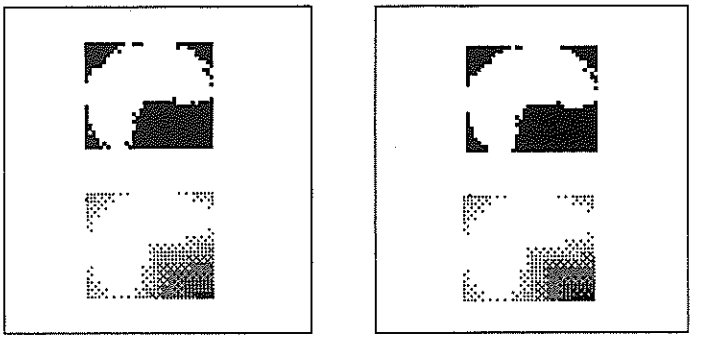


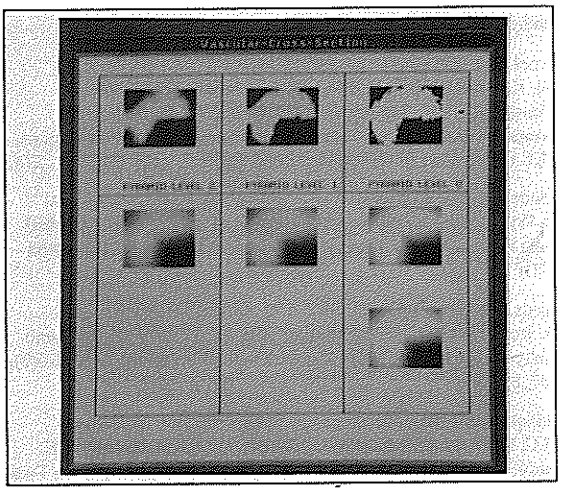
Figure 2. The Interpolation Fashion



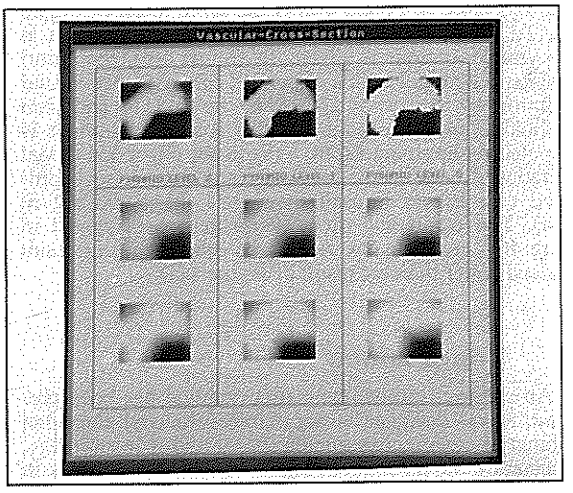
(a) A reconstruction result from a priori given model

(b) A reconstruction result by expending a priori model to a high resolution

Figure 4. An Example of Application of the Accuracy Model Method to Binary Model



(a) A pyramid reconstruction result from a priori given model



(b) A pyramid reconstruction result by estimating and improving the cost matrixes during the pyramid process

Figure 3. Reconstruction Results from the Improved Cost Matrixes vs. a Priori Model

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

THREE DIMENSIONAL RECONSTRUCTION OF BIOLOGICAL STRUCTURES IN A SUPERCOMPUTING ENVIRONMENT

A. GUIDAZZOLI*[1] , G. FABIANI* , C. FRUSCHELLI** , C. ALESSANDRINI**

* CINECA, Via Magnanelli, 6/3- 40033 Casalecchio di Reno BO, Italy
** Istitute of Histology and General Embriology, University of Siena, Italy

This paper aims to reveal the advantages of high quality computer reconstruction versus physical models, for biological structures in medical research. The collaborative efforts of biologists and computer scientists have resulted in a method to develop an automatic reconstruction of biological structures obtained from digitized serial sections in the supercomputing environment of CINECA.
High quality computer reconstruction proves more useful than physical models because they can be rotated rapidly, manipulated to bring substructures in and out of view, and furthermore can be subjected to quantitative analysis. Computer graphics and image processing, in a supercomputing enviroment, have made new advancements in this area possible. The reconstructed structure, obtained from 281 sections, consists of nearly 30000 polygons. Once visualized in real time, the features of the system also provide an automatic recording of the animation of the structure on a U-MATIC SP tape.

1. INTRODUCTION

The nature of the biological structures is clearly tridimensional, therefore the information about their spatial order is essential in medical research. In this area several computer graphics techniques, developed in an interactive environment, workstation-supercomputer, can be successfully used [1] while with physical models, the by-hand drawing of the structures becomes both tedious and often leads to a less effective representation.
We report the results of a research aimed by a double goal: the first one is the automatic reconstruction of a specific structure (a set of lymphatic capillaries) starting from 281 digitized parallel cross-sections obtained from an optical microscope ; the second one is the production of an output able to perform not only an effective visualization but also to storage the results.
This three-dimensional model can be real time visualized, rotated rapidly, manipulated to bring substructures in and out of view. It is also possible to isolate substructures, make zooms, change viewpoints and perform quantitative analysis.
A great computational power is mandatory because of the large number of digitized images to analyze. We had to isolate, in fact, in each section, the structures of interest from the others as blood vessels, etc. by image processing techniques.
The "frame by frame" recording of different animations of the reconstructed structure seems to be a good solution as final output. This recording is performed by a VTR controller connected to the graphical workstation. The videotape can be considered as a new portable record device to retrieve the materials for research and educational purpose.

2. MATERIALS AND METHODS

The specimen of a child-conjunctival biopsy has been processed in the Instute of Histology and General Embriology. A spatial 3D reconstruction is obtained by the following four main steps, [2] :

1. acquisition and digitalization of 281 microscope serial cross-sections from photographs.
2. image processing : two-dimensional contours are automatically extracted from individual slices, providing a first detection of the structure.
3. modelling: the contours in adjacent slices are connected to form triangle strips in order to obtain a 3D model of this particular set of lymphatic capillaries.
4. rendering and animation: a movie of the animated structure is automatically produced.

The research has been developed at the Laboratory for Scientific Visualization of CINECA [3,4], ViScLab, which is composed by different workstations connected by an ETHERNET local network via TCP/IP; furthermore, the graphic workstations

1) fellowship by ITALCAD s.p.a MILANO, Italy

are connected by Hyperchannel (50 Mbits/sec) to CRAY to allow an interactive access to supercomputers.

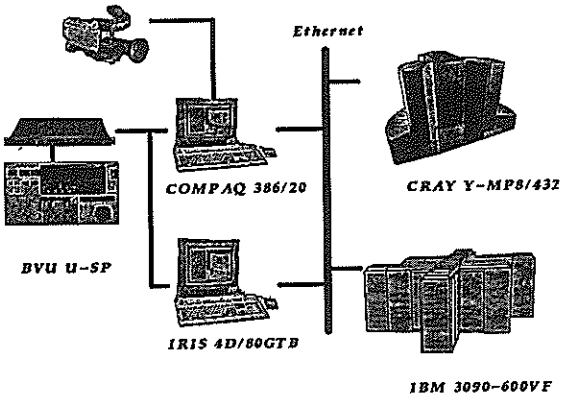


Figure 1. The supercomputing environment

The upper scheme (fig. 1) shows the system used for this application.
It consists of:

- CRAY Y-MP8/432 running under UNICOS with 24 Megawords;
- IRIS workstation 4D/80GTB;
- Sony DXC-3000K camera (resolution 500x582);
- VISTA (AT&T) board installed on COMPAQ 386/20;
- VTR controller LYON-LAMB;
- videotape recorder VTR Sony BVU950P PAL.

2.1 Data Acquisition

From a theoretical point of view, the three-dimensional computer-aided reconstructions are built up on a virtual sampling of the structure in a 3-D coordinate grid. The spatial frequencies along the axes should be choosen according to the sampling Shannon Theorem [5]. In order to adhere to these specifications the specimen of the child-conjunctival biopsy has been prepared according to the TEM method. On the bioptic specimen, embedded in Epon-Araldite, it has been constructed a prism with rectangular base and orthogonal sides, using an ultramicrotrome LKB NOVA. The 281 serial sections, with a rigorous thickness of 1 micron, have been put on a slide, stained with toluidine blue and photographed with light microscope ZEISS Axiomat at 32x . This procedure has also provided three reference marks on each section in order to make an easier realignement of the different section pictures.

The acquisition of the black and white photographs has been performed by a SONY camera and the subsequent digitalization by a VISTA board with a resolution of 740x578 pixels, 8 bit planes. The local realignement between subsequent sections has been made by the reference marks using an overlay video technique.

2.2 The structure detection

The procedure, implemented on CRAY Y-MP8/432, processes all the 281 digized sections (fig. 2a) in order to detect the regions belonging to the set of lymphatic capillaries. The algorithm is composed of three main stages. The first one set the thresholding value automatically according to the histogram statistics of each digitized section ,the second one produces the binary image which point out the regions of interest (fig. 2b). Finally a boundary tracking algorithm produces the files in which the contour data are stored. At this stage the biologists can interactively select the regions of interest for the edge detection, (fig. 2c).

Figure 2. Synthesis of processing steps

- a) digitized cross-section 206
- b) binary image after image processing
- c) binary image: edge detection

2.3 Modeling

The 3-D model has been created interconnecting adjacent contours of each section with triangles that approximate the surface from which the contours are originated. For this goal it has been used MOVIE-BYU [6] in an interactive enviroment IRIS-CRAY, which is useful to solve branching case. The wireframe model (fig.3) of the entire set of capillaries consists of nearly 30000 polygons.

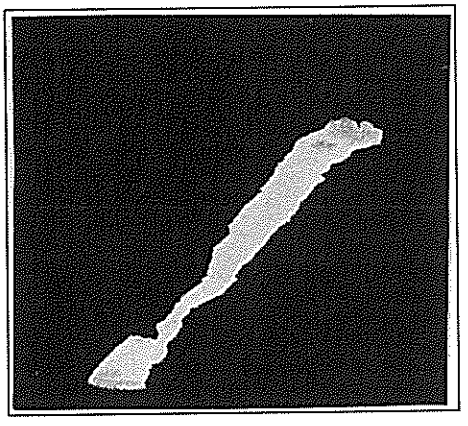


Figure 3. Wireframe model of one capillary

2.4 Rendering and animation

The visualization environment, developed by CINECA, directly connected to the CRAY Y-MP8/432, allows the users an interactive animation of models with any kind of complexity. This task is easily obtained by using the MPGS package which offers true distributed processing: the CRAY system processes the computational intensive data and the workstation performs local graphics manipulations like high speed polygon rendering, [7], (fig. 4). The MPGS driver for the VTR controller has been modified to adapt it to the different features of the European video standard PAL. This video standard has several specifications which are better than the NTSC ones as the scanning line number (625 PAL , 525 NTSC). Nevertheless the high resolution display of the graphic workstation (1280x1024) cannot be directly integrated to the present video standard because of the limited bandwith [8], so it is necessary to create a true video encoded signal.
This kind of graphic output and mainly the " frame by frame " recording of complex models can point out new details, moreover the animation emphasizes the different configuration in space of adjacent structures. The figure 5, for example, shows two different portions of lymphatic capillaries.

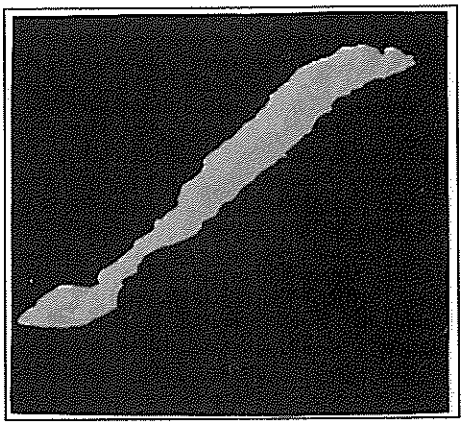


Figure 4. Shaded model of one capillary

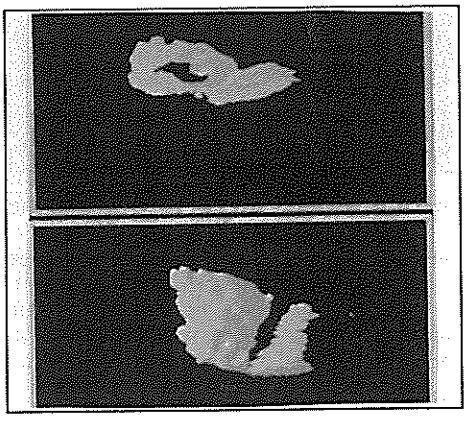


Figure 5. Details of the reconstructed structure

3. ANIMATION

Before the production of the final movie, the different parts of the structures have been visualized together and separately in a preview session.
The aim of this operation was to define the most interesting animations of the reconstructed structure from the biologists point of view using a shading which could fit the PAL video standard.

Finally a movie has been produced on a UMATIC/SP tape following a definite list of rotations and zooms (fig. 6).

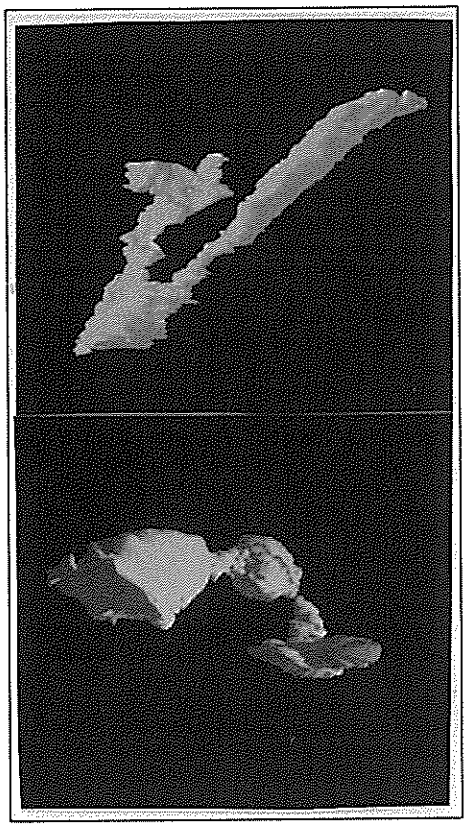


Figure 6. Two frames of the whole structure

It is worth noting that an animation of a complete rotation around the Z axes takes 720 frames (nearly 29 seconds) with the recording of 2 equal frames for 1 degree of rotation. It has required 30 minutes of recording because of the preroll BVU time. The movie is composed of 14 different frames per second (in theory 25 different frames per second are required), however the movements of the structure seem to be enough fluid and the final result on movie is pleasant to see.
Finally for this kind of shaded models the loss of information due to the video resolution does not lead to serious consequences as it is possible to notice in the frame of figure 7 which is referred to a zoom of one lymhatic capillary completely reconstructed along the 281 sections.

4. CONCLUSIONS AND FUTURE TRENDS

The flexibility and adaptability of these procedures and the economic convenience of visualization with respect to more traditional techniques, will permit straightforward applications of computer reconstruction in different research fields.
New improvements are certainly possible by means of volume rendering techniques which allow not only to reconstruct 3D structures but also to simultaneously visualize parts of the model as opaque or transparent. Nevertheless ,when a movie format is choosen as final output, the best results will occur only with the HTDV system whose high resolution will integrate for the best video standard , data and graphic systems.

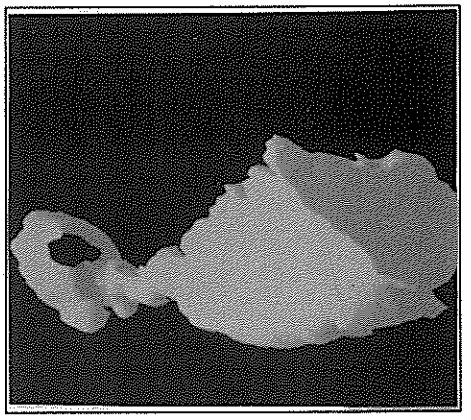


Figure 7. Zoom on one reconstructed capillary

The authors wish to thank Prof. Ing. Remo Rossi, director of CINECA, for his encouragement and the interest for the present work.

REFERENCES

[1] A.Leith, M.Marko, and D.Parsons, "Computer Graphics for Cellular Reconstruction ," IEEE Computer Graphics & Applications, Vol.9, No.5, 1989,pp.16-23.
[2] M. Vannier et al., "3-D Reconstruction Imaging", Special Symp. on Maturing Tech., 1988, pp 68-73.
[3] Fabiani G., Lanzarini M., Rossi R.: "Grafica e Supercomputer: il Laboratorio Grafico del CINECA", PIXEL, n.9, pp.45-52, 1988.
[4] Fabiani G, Lanzarini M., Moltedo L.: "Scientific Visualization In Supercomputing Environments" Proceedings of the workshop IFIP WG5, Jul.1989.
[5] D.P.Huijsmans, W.H.Lamers, J.A.Los, and J.Strackee, "Toward Computerized Morphometric Facilities: A Review of 58 Software Packages ... " The Anatomical Record, 1986,216,pp.449-470.
[6] H. Christianson and T.W. Sederberg,:"Conversion of Complex Contour Line Definitions into Polygonal Element Mosaics", Computer Graphics (SIGGRAPH '78 Proc.) 12, 1978, pp 187-192.
[7] K. Akeley, T. Jermoluk : "High-Performance Polygon Rendering", Computer Graphics (SIGGRAPH '88 Proc.) 22, 1988, pp 239-246.
[8] T.De Fanti, and D.Sandin: "The Usable Intersection of PC Graphics and NTSC Video Recording," IEEE Computer Graphics & Applications, Vol.7, No.10, Oct.1987,pp.50-58.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Grain noise modelling in ultrasonic non-destructive testing

Luis Vergara-Domínguez and Jose M. Páez-Borrallo
Dpto. Señales, Sistemas y Radiocomunicaciones, ETSI Telecomunicación-UPM,
Cdad. Universitaria, 28040 Madrid, Spain

Many materials present an internal grain microstructure. When these materials are subjected to ultrasonic non-destructive testing, the grains behave like scattering centers producing non desired backscattered noise that can difficult de detection of true deffects. This paper is devoted to the modelling of the probability density and the space-time correlation functions of the grain noise complex envelope. The analytical expressions are verified by means of real data measured in austenitic stainless steal specimen.

INTRODUCTION

Detection of defects in materials can be made by means of collecting and processing a given number of signals corresponding to the echoes backscattered by the materials when they are excited by an ultrasonic pulse.

A good defect detection implies that the echoes due to the defect should be clearly discriminated from other undesired echoes and from transducer or any other type of noise. One typical source of undesired echoes is the internal grain microstructure which is significantly important in many types of materials. Each grain behaves like a scattering center, producing an echo that, isolated or (more usually) superimposed with other echoes coming from other grains, can hide the echoes produced by a possible defect. Also, what is less important, these undesired echoes can be erroneously associated to defects (false alarms).

A good knowledge of the statistical characteristics of the grain noise might be very usefull to know potential differences that can be exploited an what their limitations are. The models could also be applied to the generation of large sets of synthetic ultrasonic records to be used in the design and testing of processing algorithms.

2.ENVELOPE PROBABILITY DENSITY FUNCTION

Let $z(\mathbf{x},t)$ denote the backscattered analytic signal corresponding to a time t (measured with respect to the instant when we sent the last pulse) and a transducer position x. Let us assume thet there are no defects, so only grain noise components are contributing to the backscattered signal. Then we can express

$$z(\mathbf{x},t) = \sum_{i=1}^{N(\mathbf{x})} A_i(\mathbf{x}) f(t-\tau_i(\mathbf{x})) \qquad (1)$$

where $N(\mathbf{x},t)$ is the total number of scatterers contributing in $\mathbf{x}$-t (those ones which are embraced by the receiver radiation diagram when the transducer is in position $\mathbf{x}$), $f(t)$ is the complex envelope of the ultrasonic RF pulse, that is

$$f(t) = p(t)\exp(\omega_0 t) \qquad (2)$$

This work has been supported by TECNATOM SA, Km 19 C. N I, San Sebastián de los Reyes, 28709 Madrid, Spain.

being $p(t)$ the RF pulse envelope and ω_0 the carrier pulsation, $A_i(\mathbf{x})$ is the (real) backscattering factor introduced by each grain; this factor will depend on the grain cross-section seen by the transducer from each position. Finally $\tau_i(\mathbf{x})$ is the time delay with which it is received the echo pulse corresponding to each grain; it will depend on the distance between the grain and the transducer.

Let us define the form factor

$$a_i(\mathbf{x},t) = A_i(\mathbf{x})p(t-\tau_i(\mathbf{x})) \qquad (3)$$

then we can express

$$z(\mathbf{x},t) = \exp(j\omega_0 t)\sum_{i=1}^{N(\mathbf{x})} a_i(\mathbf{x},t)\exp(j\phi_i(\mathbf{x})) \qquad (4)$$

where $\phi_i(\mathbf{x})$=-$\omega_0\tau_i(\mathbf{x})$, and the corresponding complex envelope will be

$$e_c(\mathbf{x},t) = \sum_{i=1}^{N(\mathbf{x})} a_i(\mathbf{x},t)\exp(j\phi_i(\mathbf{x})) \qquad (5)$$

This last expression is formaly equivalent to the one obtained in [1] where the aim was sea clutter modelling in radar applications; thus, we can apply to our case analytical results there derived.

First, we will concentrate on the problem of modelling the PDF of $e_c(\mathbf{x},t)$. In the next, to alleviate the notation, we will not express the dependence on $\mathbf{x}$-t. As it is done in [1], we will assume that the form factors a_i are independent and identically distributed random variables (i.i.d.rr.vv) having a K-type PDF of the form

$$p(a) = \frac{2b}{\Gamma(1+\nu)}\left(\frac{ba}{2}\right)^{1+\nu} K_\nu(ba)\ ,\quad \nu > -1 \qquad (6)$$

and the phases ϕ_i are i.i.d.rr.vv.uniformely distributed between 0 and 2π. Also, the form factors a_i are independent of the phases ϕ_i.

The choice of (6) as the the form factors PDF is not arbitrary in two senses. First, by means of changing the parameter ν (the parameter b is just a scale factor), we can consider a wide family of PDF's, ranging from a log-

normal distribution for ν values near to -1, to a Rayleigh distribution for ν values greater than 9. Hence, as ν approaches -1, it increases the probability of finding scatterers contributing with a significantly larger form factor than the rest. Secondly, this choice allows an analytical derivation of the envelope PDF.

Under the above hypothesis it can be shown that the envelope $e=|e_c|$ is also K-distributed (this is demonstrated in [1]) in the form

$$p(e)=\frac{2b}{\Gamma(M)}\left(\frac{be}{2}\right)^{M}K_{M-1}(be),\quad M>0 \qquad (7)$$

where K_{M-1} is the $(M-1)$-th order modified Bessel function of the third kind and the parameter M is given by

$$M=(1+\nu)N \qquad (8)$$

Note that M in (7) play the part of $1+\nu$ in (6). Let us dedicate some words to explain the physical meaning of M.

For a given ν value (remember that $\nu>-1$), M increases with N, the actual number of scatterers. If N is large enough to make the product $N(1+\nu)$ greater than 10, the envelope PDF will be Rayleigh; that is, the in-phase and in-quadrature components will be gaussian. This is the conclusion we arrive to, when the central limit theorem is applied to the superposition of the individual grain echoes expressed in (5). A Rayleigh distribution has been assumed for the grain noise modelling by different authors [2]-[4], but the grain sizes considered were considerably small in comparison with the center wavelength (in other words N was allways a very large value). We can then say that the applicability of the central limit theorem depends on M instead of depending directly on N.

3. EVALUATING THE PARAMETER M

We start from (8). First, assuming a uniform grain volume density ρ we can express

$$N=\rho V \qquad (9)$$

where V is the volume embraced by the transducer radiation diagram. If we consider that the contribution due to the diagram secondary lobes is irrelevant and that the main lobe (beam) is cylindrical (this last hypothesis wil be further removed) the volume V can be written in the form

$$V=\beta\sigma cS \qquad (10)$$

where β is the number of pulses that fit into a line along the beam depth, c is the propagation velocity, σ is the pulse duration (i.e., $\beta\sigma c$ corresponds to the cylinder heigth) and S is the area of the cylinder cross-section.

On the other hand it can be shown that considering, as it is usual, a gaussian pulse, it is

$$1+\nu=\frac{2\frac{\pi}{\beta^2}erf^2\left(\beta\sqrt{2}\right)E^2\left[A^2\right]}{\frac{\sqrt{\pi}}{\beta\sqrt{2}}erf(2\beta)E\left[A^4\right]-2\frac{\pi}{\beta^2}erf^2\left(\beta\sqrt{2}\right)E^2\left[A^2\right]} \qquad (11)$$

Multiplying (9) by (11) and making $\beta\rightarrow\infty$ (this is a reasonable hypothesis in a practical case) we finally arrive to

$$M=\sqrt{2\pi}\rho\sigma cS\frac{E^2\left[A^2\right]}{E\left[A^4\right]} \qquad (12)$$

The above expression indicates that the effective number of scatterers is directly related to:ρ, the average grain volume density; the σc product; and the cross-section area S. All this later is a consequence of the direct relation between M and the actual number of scatterers N.

The parameter M is also directly related to the quotient $E^2[A^2]/E[A^4]$. If all the grains would have the same cross-section, this quotient will be equal to 1. Nevertheless, if there are some grains with a much greater cross-section than the rest, this quotient will go away from 1. Actually, if we assume that the cross-section A is also K-distributed with a parameter m, we have that

$$\frac{E^2\left[A^2\right]}{E\left[A^4\right]}=\frac{m}{2(1+m)},\qquad m>0 \qquad (13)$$

When $m\rightarrow 0$ the quotient tends to 0, when $m\rightarrow\infty$ the quotient tends to 0.5, which implies that we will allways have a reduction of the effective number with respect to the actual number of scatterers.

Although (12) has been derived under the hypothesis of a cylinder beam, the results can be easily extended to any type of beam if we "locally" approximate the beam by a cylinder. We just have to substitue S in (12) by a local area $S(t)$ For example, in an ideal conical beam it is easy to show that

$$S(t)=\pi\left[\frac{ct}{2}\tan(\Delta/2)\right]^2 \qquad (14)$$

where Δ is the angular beamwidth.

Finally, coming back to the complete notation, which include the dependence on x-t, we have in general

$$M(\mathbf{x},t)=\sqrt{2\pi}\rho(\mathbf{x})\sigma c(\mathbf{x})S(t)\frac{E^2\left[A^2(\mathbf{x})\right]}{E\left[A^4(\mathbf{x})\right]} \qquad (15)$$

4. SPACE-TIME CORRELATION FUNCTION

We define the complex envelope space-time correlation function as

$$R(\mathbf{x},\mathbf{x}',t,t')=E\left[e_c(\mathbf{x},t)e_c^*(\mathbf{x}',t')\right]=$$

$$=E\left[\sum_{l=1}^{N(\mathbf{x})}\sum_{m=1}^{N(\mathbf{x}')}a_l(\mathbf{x},t)a_m(\mathbf{x}',t')\exp(j(\phi_l(\mathbf{x})-\phi_m(\mathbf{x}')))\right] \qquad (16)$$

Taking into account the statistical independence between the phases ϕ_l and the form factors a_l we can write

$$R(\mathbf{x},\mathbf{x}',t,t') = \sum_{l=1}^{N(\mathbf{x})}\sum_{m=1}^{N(\mathbf{x}')} E\left[a_l(\mathbf{x},t)a_m(\mathbf{x}',t')\right]E\left[\exp(j(\phi_l(\mathbf{x})-\phi_m(\mathbf{x}')))\right] \quad (17)$$

Each term in the above double summatory is the contribution to the space-time correlation, of pairs of scatterers correlations. The first scatterer is embraced by the transducer radiation diagram in position x, and the second one in position x'. We may devide the set of scatterer pairs, into two different categories. The first categorie corresponds to the scatterer pairs where both scatterers are the same (the scatterer must necessarily be inside the overlapping region between the transducer radiation diagrams in positions x and x'). The second categorie corresponds to the pair os scatterers where both are different. Assuming statistical independence between two different scatterers, we have that the second categorie will not produce any contribution to the space-time correlation.Then we can express the space-time correlation as

$$R(\mathbf{x},\mathbf{x}',t,t') = \sum_{l=1}^{\alpha(\mathbf{x},\mathbf{x}')N(\mathbf{x})} E\left[a_l(\mathbf{x},t)a_l(\mathbf{x}',t')\right]E\left[\exp(j(\phi_l(\mathbf{x})-\phi_l(\mathbf{x}')))\right] \quad (18)$$

where $\alpha(\mathbf{x},\mathbf{x}')$ is the percentage of scatterers contributing when the transducer is in position x, that also contribute when the transducer is in position x'. On the other hand

$$E\left[a_l(\mathbf{x},t)a_l(\mathbf{x}',t')\right] = E\left[A_l(\mathbf{x})p(t-\tau_l(\mathbf{x}))A_l(\mathbf{x}')p(t'-\tau_l(\mathbf{x}'))\right] = E\left[A_l(\mathbf{x})A_l(\mathbf{x}')\right]R_p(t-t'-\tau_l(\mathbf{x})+\tau_l(\mathbf{x}')) \quad (19)$$

where $R_p(t)$ is the pulse envelope autocorrelation function. Note that x and x' are near vectors, because there must be some overlapping between the beams in both positions to make $R(\mathbf{x},\mathbf{x}',t,t')$ different from zero. So we can consider in practice $A_l(\mathbf{x})\approx A_l(\mathbf{x}')$, $\phi_l(\mathbf{x})\approx\phi_l(\mathbf{x}')$, and hence

$$R(\mathbf{x},\mathbf{x}',t,t') = N(\mathbf{x})E\left[A^2(\mathbf{x})\right]\alpha(\mathbf{x},\mathbf{x}')R_p(t-t') \quad (20)$$

where we have assumed that $E[A_l^2(\mathbf{x})] = E[A^2(\mathbf{x})] \ \ \forall l$.

Finally the space-time correlation coefficient will be

$$R(\mathbf{x},\mathbf{x}',t,t')/R(\mathbf{x},\mathbf{x},t,t) = \alpha(\mathbf{x},\mathbf{x}')\rho_p(t-t') \quad (21)$$

where $\rho_p(t-t') = R(t-t')/R_p(0)$.

Expression (21) indicates that the complex envelope space-time correlation is a separable function. The space correlation is the beam overlapping factor $\alpha(\mathbf{x},\mathbf{x}')$; the time correlation is the pulse envelope autocorrelation function.

Some final comments about the factor $\alpha(\mathbf{x},\mathbf{x}')$. Of course, this factor will depend on the particular beam geometry. Nevertheless we can stablish an upper limit for it. If w_{max} is the maximum spatial beamwidth, it must be

$$\alpha(\mathbf{x},\mathbf{x}') \le 1-\frac{|\mathbf{x}-\mathbf{x}'|}{w_{max}} \qquad |\mathbf{x}-\mathbf{x}'| \le w_{max}$$

$$\alpha(\mathbf{x},\mathbf{x}') = 0 \qquad |\mathbf{x}-\mathbf{x}'| \ge w_{max} \quad (22)$$

It is a little bit more tedious to show that the squared envelope space-time correlation coefficient is given by

$$R_{e^2}(\mathbf{x},\mathbf{x}',t,t')/R_{e^2}(\mathbf{x},\mathbf{x},t,t) = \alpha(\mathbf{x},\mathbf{x}')\rho_{p^2}(t-t')+m \quad (23)$$

where $\rho_{p^2}(t)$ is the correlation coefficient of the squared pulse envelope. It is easy to show that for a gaussian pulse

$$\rho_{p^2}(t) = p(t)/p(0) \quad (24)$$

that is, the time correlation coefficient is just the normalized pulse envelope. The constant value m is due to the nonzero mean-value of an envelope record. We will use (23) and (24) in the verification experiments of the next section.

5.EXPERIMENTAL RESULTS

In this section we have tried to verify the analytical results previously presented. The data have been collected in austenitic stainless steal specimen, with an average grain size equal to 1,5 *mm*. In all the experiments the transducer center frequency and bandwidth was around 1 *MHz*.

First we have tried to test if the envelope was *K*-distributed. To do it we have used a specimen composed by two austenitic stainless steal pieces joined by a welding. We have generated three sets of data corresponding respectively to measures taken in the first piece, the second piece and the welding. No flaws were present in the specimen. Each data set was generated locating the transducer in a number of positions in the corresponding part of the specimen. In each location we sent an ultrasonic pulse and we registered the first 48 samples of the echo signal. The sampling frequency was 3 *MHz* and the specimen depth was such that the 48 samples practically correspond to a complete depth round throught it. The different transducer location were randomly chosen, far enough to avoid beam overlapping. With the above adquisition scheme we assure that the different samples are practically uncorrelated. This is a requirement for an efficient PDF estimation.

We have located the transducer in 15 different positions in each part, so we allow 48x15=720 samples in each data set. Then we have applied a Chi-square test to the data, with *K*-distributions having a parameter *M* varying from 1 to 10. Figures 1,2 and 3 show the results obtained with each data set. Figures 1a) 2a) and 3a) show the value of the statistic used to perform a Chi-square test for the *M* value range considered. A smaller statistic value implies a better PDF fitting. Figures 1b), 2b) and 3b) show the percentage confidence level of the test for each value; this confidence level is evaluated as the probability of the Chi-square statistic to be greater than the value shown in Figures 1a), 2a) and 3a) respectively. Finally, Figures 1c), 2c) and 3c) show the data set histogram superimposed with the theoretical *K*-type PDF having an *M* parameter value showing the best fitting in the previous figures. That is, *M*=3 in Figure 1c), *M*=7 in Figure 2c) and *M*=6 in Figure

3c). We can observe the general high values of the percentage confidence levels. This is specially true in the last data set (which corresponds to the welding) where the confidence level is 95%. The agreement is also good between the histograms and the theoretical PDF. Also of interest is to note the important variation of the best fitting M value in each part of the block.

Next we have tested the space-time correlation model. The transducer was located in 7 positions forming a straight line. Two consecutive positions were only 2 *mm* appart to guarantee space correlation. In each position we have registered 256 samples taken at 6.4 *MHz* frequency sampling; this will assure time correlation. So we have a matrix having 7x256=1792 samples. Figure 4 shows the space-time correlation estimate of the square envelope. Note the separability of the 2D function as was predicted by (23). The space correlation is shown separately in Figure 5; this corresponds to the particularization of the space-time correlation estimate shown in figure 4 for t=0. The dotted line is the upper limit of the space correlation corresponding to the maximum beamwidth of the beam used. As expected the estimated space correlation is under this limit. In figure 6 we show (dotted line) the corresponding time correlation (particularization of the space-time correlation estimate shown in figure 4 for x=0). The dark line is the normalized pulse used. Clearly, figure 6 is in agreement with expressions (23) and (24).

CONCLUSIONS

We have derived analytical models for the envelope PDF and the space-time autocorrelation functions of the noise due to the grain microstructure which is present in many types of materials. These models can be directly used to built backscattered grain noise simulators for exhaustive testing of processing algorithms. Additionally the noise characterization gives us valuable information about the posible alternatives for noise processing.

Additionally, the envelope PDF model can help in the design of automatic detectors with controlled detection or false alarm probabilities. Another potential use of the models could be material characterization, for example by means of (12).

Finally, the performed experiments encourage the use of the above models in other types of materials or tissues.

REFERENCES

[1] E. Jakeman, P.N. Pusey: "A model for Non-Rayleigh Sea Echo," IEEE AP-24, pp.806-814, Nov. 1976.

[2] J. Saniie et al.: "Statistical Evaluation of Backscattered Ultrasonic Grain Signals," JASA-84, pp. 400-408, July 1988.

[3] J. Saniie, N.M. Bilgutay: "Quantitative Grain Size Evaluation Using Ultrasonic Backscattered Echoes,"JASA-80, pp. 1816-1824, Dec. 1986.

[4] J. Saniie et al. : "Analysis of Homomorphic Processing for Ultrasonic Grain Signal Characterization," IEEE UFFC-36, pp. 365-375, May 1989.

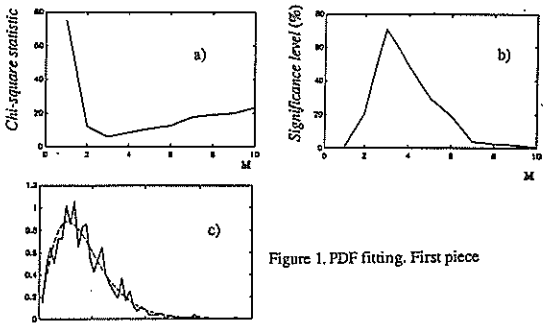


Figure 1. PDF fitting. First piece

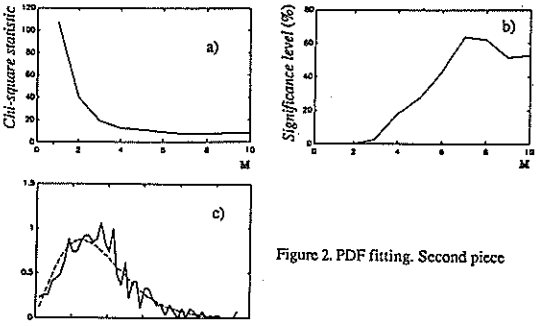


Figure 2. PDF fitting. Second piece

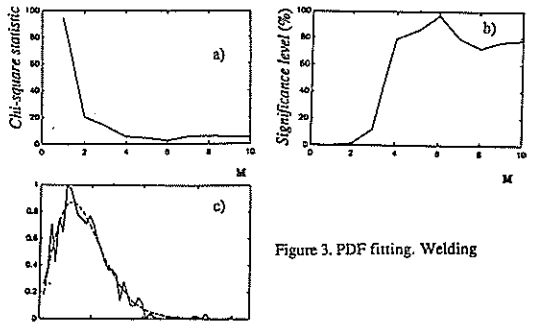


Figure 3. PDF fitting. Welding

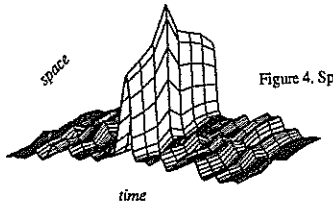


Figure 4. Space-time correlation estimate

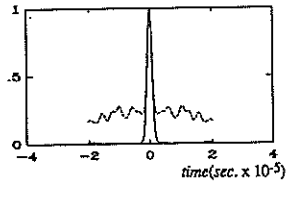


Figure 6. Time correlation estimate

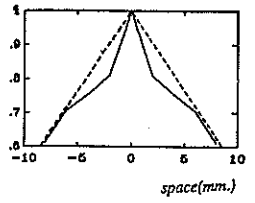


Figure 5. Space correlation estimate

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

MICROSTRUCTURAL PROPERTIES REFLECTED ON THE ENVELOPE AND POWER SPECTRAL DENSITY OF THE RF IMAGE FROM TISSUE-LIKE PHANTOMS

Luigi Landini *, Maria Filomena Santarelli +, Lucio Verrazzani

Institute of Electronics and Telecommunications, University of Pisa, Pisa Italy
Via Diotisalvi, 2 , Pisa, Italy

This paper is devoted to investigate the amplitude distribution and the power spectral density of the RF image in order to extract informations concerning the reflectivity of the medium and its spatial architecture.
It is shown that if the resolution cell holds a few scattering centres a super-Rayleigh or sub-Rayleigh behaviour of the amplitude statistics, well fitted by the Weibull law, comes out in dependence of the homogeneity degree of the scattering cross section. Moreover, starting from a gamma distribution to represent the degree of regularity in the spatial architecture, an expression for the power spectral density of the RF signal is derived. Results from simulation and experiments on phantoms are referred and discussed.

1. INTRODUCTION

When the backscattered or echo pulses are detected for displaying as an image, the resulting picture has the granular structure described as texture or speckle. The quantitative evaluation of the speckle pattern, associated with the microstructure of tissue parenchyma, furnishes the basis for clinical diagnosis. Thus our study is devoted to understand how the microstructural properties of the tissue are reflected on the envelop and the power-spectral density (or autocorrelation function) of the RF signal.
The results known in communication theory [1] establish the following statistical properties of the intensity in the image of a one-dimensional target. If the number of scatterers within the effective image resolution cell is very large, the scattered wavelets have random phases, the scattering region is very rough on the scale of the wavelenght and the scattering structure is too fine to be resolved, then according to the central limit theorem the RF signal Y(t) may be considered as a band-pass process with power density spectrum $W_y(f)$ bandlimited to $(f_0-B<|f|<f_0+B)$ so that the Rice's representation holds:

$$Y(t) = V(t) \cos [2 \pi f_0 t + \Theta(t)]$$

where $V(t) \geq 0$ and $0 < \Theta(t) < 2\pi$ are the input envelope and phase respectively. The output $Z(t) = V^2(t)/2$ of a square law device followed by a low-pass zonal filter (assuming $W_y(f)$ has even simmetry about $f=f_0$), has a probability density function

$$p(Z) = 1/(\sigma_y)^2 \exp [-z(\sigma_y)^2] \quad (Z>0) \qquad (1)$$

and a spectral density

$$W_Z(f) = (\sigma_y)^4 \delta(f) + 2 [W_y(f) \oplus W_y(f)] G_{2B}(f) \qquad (2)$$

that is composed of an impulse term corresponding to the squared input power and a low-pass term corresponding to the random variations.
When the pulsewidth (always a few wavelenghts) and/or the beam aperture are small, however, only a few scattering centers contribute to the field and a deviation from Rayleigh statistics is found which depends on the number and the properties of the individual scattering centers. This problem has been investigated in the literature on high resolution radar operating over the sea, whose performance is often limited by the unwanted returns (clutter) from the sea surface; these works the reader is referred to.

2. PROBABILITY DISTRIBUTION OF THE SQUARED-DETECTED ENVELOPE

The work of E. Jakeman and P.N. Pusey [2] is of particular interest for us since, starting from the model of the scattering phenomenon as a finite two dimensional walk, they demonstrate an important property of the non-Rayleigh echoes: a scaling of the moments of the fluctuations distribution with the target area. Starting from Pearson work [3] they derive the normalized moments of the distribution of the resultant "radar cross section" Z. We rewrite the normalized second moment M_2 given in [2]:

$$M_2 = \langle Z^2 \rangle / \langle Z \rangle^2 = 2(1- 1/N) + (1/N) (\langle \mathcal{A}^2 \rangle / \langle \mathcal{A} \rangle^2)$$

where $\langle \mathcal{A} \rangle^2$ is the average scattering cross section of a single scatterer and $\langle \mathcal{A}^2 \rangle$ its second moment.
This expression suggests the following comments:
- as expected for N large, $M_2 \rightarrow 2$, the second moment of a negative exponential (or Rayleigh amplitude) density function;

L. Landini and M.F. Santarelli are also with the C.N.R. Institute of Clinical Physiology, Via Savi, 8 - Pisa

- even if the second moment has been obtained for a constant number of scatterers in each resolution cell, it may be easily extented by averaging with respect to N so that, retaining the first two terms of E { 1/N } :

$$M_2 \approx 2-(1/N + (\sigma_N)^2/N^3)\,(2- <\mathcal{A}^2>/<\mathcal{A}>^2)$$

where N is now the average number of scatterers per cell and $(\sigma_N)^2$ its variance.
This expression may be rewritten as

$$M_2 = 2 -2(1/N + (\sigma_N)^2/N^3)\,(1-K^{-1}) \qquad (3)$$

where $K^{-1} = <\mathcal{A}^2> / 2<\mathcal{A}>^2$ is a uniformity or homogeneity parameter that accounts for the scattering cross section variations: $1<K<2$ means that the cross section is spatially uniform, while $K<1$ denote inhomogeneous scattering media. The intermediate value $K=1$ corresponds to a ratio standard deviation - mean value of the cross section equal to 1.
Equation (3) shows an explicit dependence of the second moment on the average number of scatterers and on the cross section of each scatterer.
The following cases are of interest:
1) $1<K<2$ or $1 <M_2<2$, that takes place if the cross section of the individual scatterers is spatially uniform in the sense: $[<\mathcal{A}^2>/ 2<\mathcal{A}>^2]<1$. In this case the second moment $M_2 = 2 -(1/N + (\sigma_N)^2/N^3)$ decreases with N and tends to the minimum $M_2 = 1$ for $N \to 1$ and $(\sigma_N)^2 \to 0$; note that only for $N <100$ deviations from exponential statistics will be appreciated. In actual situations this occurs when a small number of scatterers N give rise to an interference of the wavelets scattered with random phases simulating a specular effect (super-Rayleigh conditions). This result is not in accordance with the classical theory, that is based on the central limit theorem and holds for large N($N \to \infty$) and under the hypothesis that the phases of the scattered fields are uniformly distributed over 2π radians (distances of the scatterers larger than the wavelenght)
2) $K =1$ or $M_2 = 2$, that comes true when small variations of the cross section compensate the effect of a low N so that $M_2 =2$ as for negative exponential p.d.f..
3) $K <1$ or $M_2 >2$, that proves correct for scatterers cross sections with large variance, independently of N and σ_N i.e. of the number of the scattering centres.
This last condition entails that, even if N is large, non Rayleigh effects may rise when only a small fraction of scatterers gives a significant contribution, that occurs if the cross section has large fluctuations (sub-Rayleigh conditions).
These results make evident the possibility of an increase in the SNR not imputable to periodicities in the microstructure but to low values of the concentrations associated with uniformity in the scatterers cross section.
The comparison between eq.(3) and the normalized second moment of the Weibull distribution [4]:

$$M_2 = 2b\Gamma(2/b) / \Gamma^2(1/b)$$

indicates that in the previous conditions 1), 2) and 3) the Weibull parameter b is respectively: $b>1$, $b=1$, $b<1$; the corresponding distributions are quite different: indeed for $b<1$ the p.d.f. exhibits long tails that extend further from the mean value than would be expected from a large number of random scatterers (spiky signal with SNR smaller than for exponential statistics), while for $b>1$ larger SNR are found.

Thus the b parameter may be assumed as a characterization parameter:
- $b>1$ is typical of tissues with a low number of scatterers in each resolution cell, uniformly distributed in cross-section;
- $b<1$ is characteristic of tissues with an "effective" number of scatterers greately reduced because the cross section of the individual scatterers has large fluctuations.

3. CORRELATION FUNCTION OF THE RF SIGNAL

The function $W_y(f)$ in eq.(2), which gives a spatial domain description of texture, may be easily derived if a model of the tissue architecture is assumed. Owing to the commutative property of the convolution, the echo formation process may be regarded as an input process, whose sample functions are sequences of delta functions reproducing a particular configuration of the scatterer's spatial distribution, filtered by a pass-band filter H(f) accounting for overall echographic system response (both in transmission and reception). If we consider a field of N scatterers and indicate with a_i the amplitude of reflection from the i-th scatterer and t_i the time delay for a sonic pulse to reach the i-th scatterer and return to the transducer, each sample function of the received process is expressed by:

$$y(t) = [\,\Sigma_i\ a_i\ \delta(t-t_i)] \oplus h(t).$$

The power spectrum of the process becomes:

$$W_y(f) = |H(f)|^2\, E\{\,\Sigma_i\ \Sigma_k\ a_i\, a_k\, e^{-j2\pi f(t_i-t_k)}\} = |H(f)|^2\, W_X(f)$$

with $W_X(f) \Delta E\{\,\Sigma_i\ \Sigma_k\ a_i\, a_k\, e^{-j2\pi f(t_i-t_k)}\}$ the power spectral density of the input process. In the hypothesis that a_i and t_i are statistically independent random variables, it follows:

$$W_X(f) = \Sigma_i\,\Sigma_k\ E\{a_i\, a_k\}\, E\{e^{-j2\pi f(ti-tk)}\} =$$
$$= N\,[\,E\{a_i^2\} - E^2\{a_i\}\,] + E^2\{a_i\}\,\Sigma_i\,\Sigma_k\ E\{e^{-j2\pi f(ti-tk)}\}$$
$$= N\,[\,E\{a_i^2\} - E^2\{a_i\}\,] + E^2\{a_i\}\ \Gamma(f)$$

where $\Gamma(f)$ is the p.s.d. of the stationary-point process.
Let us assume as in [5], that the interarrival time $\tau = t_i - t_{i-1}$, is distributed according to the n-order gamma probability density function, so that as n increases the ratio between standard deviation and mean value of the intervals decreases; thus the random point process sweeps from Poisson process ($n = 1$) to the opposite case of a process of equally spaced impulses ($n = \infty$) and the effects of regularity in space architecture can be investigated.
In this case [5] the p.s.d. is a sum of equally spaced peaks whose number and height increase with n, i.e. with the degree of regularity, while their width decreases; on the contrary the p.s.d. becomes white for quite random process.
Finally

$$W_y(f) = |H(f)|^2 \cdot [\,N\,(\sigma_a)^2 + <a^2>\,\Gamma(f)\,] \qquad (4)$$

This expression is easy to visualize in two particular cases:
- for quite random scatterers $\Gamma(f)$ approaches a constant so that $W_y(f)$ is proportional to $|H(f)|^2$ i.e. the p.s.d. of the RF image equals the p.s.d. of the RF image of a single point reflector; thus the B-mode texture autocorrelation has a lower bound imposed by the RF point response;

- for a regular distribution of scatterers $\Gamma(f)$ and then $W_y(f)$ exihbit a spectral peak (obviously if it falls within the system bandwidth); as a consequence the autocorrelation is greatly widened.
When the non linear detection of the backscattered RF echoes is accounted, which transforms the input voltage in light density, the p.s.d. of the texture bears no resemblance to the p.s.d. of the RF image.
Indeed, from eq. (2) (4) it follows:

$$W_Z(f) = (\sigma_y)^2\delta(f) + 2\,[\,|H(f)|^2\,[N(\sigma_a)^2 + \langle a^2\rangle\,\Gamma(f)] \otimes N(\sigma_a)^2 + \langle a^2\rangle\,\Gamma(f)\,]\,G_{2B}(f)$$

In the extreme cases, previously considered, we have:
- for quite random scatterers $W_Z(f)$, except for a delta function (d.c. component), is proportional to the low-pass component of $|H(f)|^2 \otimes |H(f)|^2$; thus its width is about twice that of $|H(f)|^2$;
- for a regular distribution of scatterers, the periodic peaks in $\Gamma(f)$ give a negligible contribute to the speckle fluctuations owing to the integrative effect of the convolution operator. Thus, in order to extract informations about the architectural properties of a tissue, the p.s.d. or the autocorrelation of the RF signal must be considered.
All the theoretical results have been tested by simulation.

4. EXPERIMENTAL RESULTS

Experiments from tissue-like phantoms provide a mean to differentiate between the effect of the imaging system and the contribute from tissue structure to the texture of the images. Indeed phantoms realize controlled and known, at least statistically, conditions of the scattering medium, so that the influence of the ultrasound system (transducer characteristics, scanning parameters, focusing conditions) can be analyzed while keeping the scattering ensemble constant. On the other hand, by using the same instrumentation and varying the scatterers parameters within the phantom (concentrations, size, degree of homogeneity, compressibility) it is possible to establish how the backscattered signal depends on the tissue structure; thus we can extract diagnostically useful parameters.
With this aim in our laboratory the statistical properties of the ultrasonic backscattering from tissue-like phantoms have been investigated.
Tissue equivalent phantoms were produced with PVC microspheres imbedded in water based gelatine. Measurements were performed by using a broadband pulser-receiver (Panametrics, model 5052PR); RF signal underwent to fast analog-to-digital conversion (Le Croy, model TR 8818, with acquisition rate of 100 MHz and amplitude resolution of 8 bit) and transferred to a computer for processing. Phantoms were investigated in different operating conditions:
- by using transducers with different focus (panametrics, 5 MHz, 13 mm diameter, focused at 7 cm and unfocused; the -3dB resolution cell then ranges from 1 to 1,5 mm^3;
- by varying the scatterers concentration (from 10^4 to 10^6 scatterers per cm^3) .
The amplitude histograms of the envelope-detected backscattered signal, are shown in Fig. 1 and 2 ; they exhibit a deviation from Rayleigh distribution since the resolution cell near the focus, particularly with focused transducers, holds a few scattering centres. As expected the Weibull distribution furnishes a good fitting; indeed if the cumulative histogram of the experimental data are represented in a Weibull paper, so that a distribution

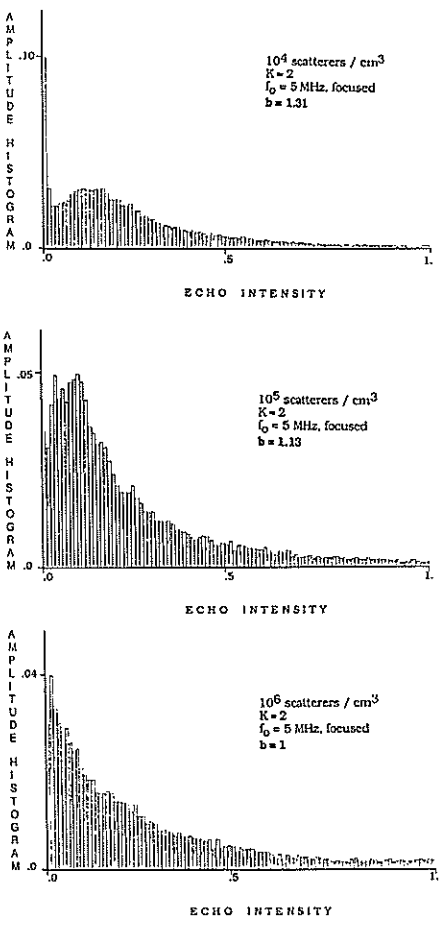


Figure 1

Experimental amplitude histograms obtained from tissue-like phantoms at different concentrations, with focused transducer.

which is Weibull will result in a straight line, small deviations appear, according to the minimum-mean-square-error. The dependence of the distribution parameters on the transducer characteristics and scatterers properties can be summarized as follows:
- since all the experiments have been performed on phantoms with selected particles of 30μm diameter, the parameter K can be assumed equal to 2; thus the p.d.f. are super-Rayleigh;
- as concentration rises the deviation from Rayleigh decreases;
- when a focused transducer is used larger deviations are found, as a consequence of the reduction in the resolution cell.
In Fig. 3 the p.s.d. of the RF signal are shown: as far as the dependence of the p.s.d. on the scatterers concentration, the experimental results allow to conclude that the spectral width increases with the

scatterers concentration, in accordance with the results of the p.d.f. which shows a decrease of the SNR that must be ascribed to low values of N.

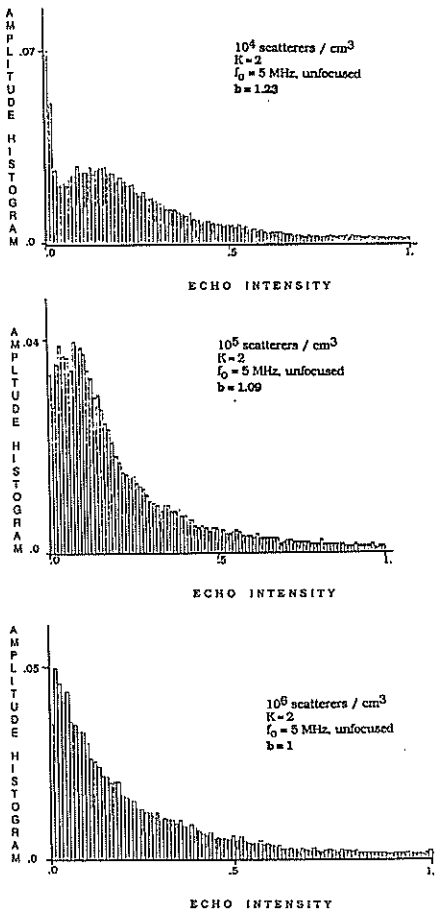


Figure 2

Experimental amplitude histograms obtained from tissue-like phantoms at different concentrations, with unfocused transducer.

5. CONCLUSIONS

The second moment analysis of the square-law detected echoes evidentiates deviations from the Rayleigh distribution. In the classical literature these deviations are encountered both when a specular component is added to a diffuse component (Rician distribution) and when the scatterers are resolvable (sub-Rayleigh distribution).
For low concentrations, as it occurs in many practical situations when the resolution cell is reduced so that a few scatterers contribute to the scattered field, a sub-Rayleigh or super-Rayleigh behaviour occurs which

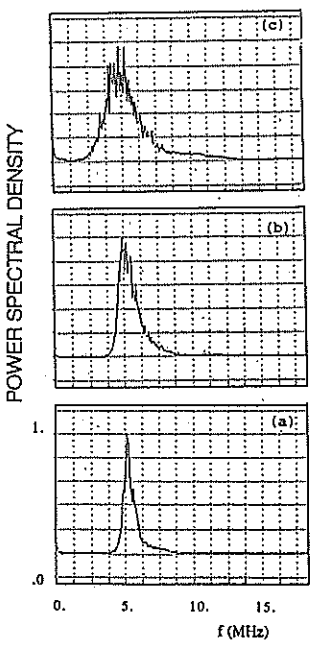


Figure 3

Experimental p.s.d. obtained from tissue-like phantoms at different concentrations: a: 10^4 *scatterers/ cm*^{3}; b : 10^5 *scatterers / cm*^{3}; c: 10^6 *scatterers/cm*^{3} .

differentiates heterogenous from homogeneous media. In these conditions the Weibull distribution well fits the simulated and experimental data so that the b-parameter of the p.d.f. can be correlated to the homogeneity properties of the scattering medium. As far as the architectural properties are concerned, the degree of regularity, which affects also the SNR of the intensity, can be recognized through the presence, of spectral peaks in the p.s.d. of the RF image.

REFERENCES

[1] Davenport, W.B, Jr, and W.L. Root, An introduction to the theory of Random Signals and Noise, (McGraw-Hill Book Company, Inc., New York, 1958) Chapter 12.

[2] Jakeman, E. and Pusey, P.N., A model of non-Rayleigh sea echo, (IEEE Trans. Antennas Propagat., Vol. AP-24, 6, 1976) pp. 806-814

[3] Pearson, K., A mathematical theory of random migration, (Dreper's Company Research Memoirs, Biometric Series III, n. 15, 1906).

[4] Boothe, P.R., The Weibull distribution applied to the ground clutter backscatter coefficient (U.S. Army Missile Command, Report n. RE. TR. 69-15, AD 91-109, 1969) pp. 1-19.

[5] Landini, L., Verrazzani, L., Spectral characterization of tissues microstructure by ultrasound: a stochastic approach. (IEEE Trans. Ultrasonics, Ferroelectrics and Frequency Control), in press.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ANISOTROPIC DIFFUSION AND MORPHOLOGICAL APPROACHES FOR ECHOCARDIOGRAPHY IMAGE PROCESSING

Claudio Lamberti', Fiorella Sgallari"

'Dipartimento di Elettronica,Informatica e Sistemistica. "Dipartimento di Matematica. Universita' di Bologna. Bologna, Italy.

Automatic image analysis allows reduced manual operations and, above all, ensures objectivity and repetition of analysis. In this paper, we present an approach for image preprocessing that utilizes the anisotropic diffusion technique, and afterwards morphological techniques are utilized for ventricular contour detection. Results of automatic detection of left ventricle contours on 2-D echocardiographic images are presented and discussed.

1. INTRODUCTION

Imaging techniques in medical diagnostics are used more and more frequently and their application field is expanding rapidly. Particularly within the cardiological field, echocardiography has proven a high level of effectiveness at reasonably low cost, even if the image quality obtained by using such technique is not as good as those obtained by using others. Moreover, important deductions as to heart size and function can still be drawn from echo images.
Automatic contour detection in 2-D echocardiography presents serious difficulties because of the poor image quality, the possibility of drop-outs, high noise level. valve and papillary muscle interference in the images, etc.
A common approach to edge detection of 2-D monochromatic digitized echocardiographic images provides a three step processing [1,2,3,4]:
a) Image regularization; b) Edge enhancement; c) Ventricular contour detection.

In this paper, we present a new approach for image preprocessing that utilizes the anisotropic diffusion technique. This approach, described in Section 2, has the advantage of reducing the effect of noise superimposed to the image and emphasizing the edge of the objects. In Section 3, after a short review of some definitions in mathematical morphology, we describe our ventricular contour detection method based on morphological operations. Experimental results and conclusions are given in Section 4 and 5, respectively.

2. ANISOTROPIC DIFFUSION TECHNIQUE

The anisotropic diffusion technique was proposed by Perona and Malik [5] to reduce the effect of the noise superimposed to the image and to emphasize the edges of the objects. By convolving the original image with a low-pass kernel, the high frequencies are damped and a generalized smoothing is obtained so that noise is reduced but the contours are blurred and the algorithms for contour detection produce poor results. The anisotropic diffusion technique, on the other hand, estimates the luminous intensity gradient locally and referring to such a value produces a smoothing only inside the semantically homogeneous regions, while the contrast is stressed close to the contours. For a detailed explanation of how the technique operates, we refer to [5,6].
The anisotropic diffusion technique produces different homogeneous regions so that segmentation is now much easier. In fact, if the ventricular contour has to be detected, as in our case, a binary image is obtained by comparing the pixel value with a suitable threshold.

3. THE MORPHOLOGICAL APPROACH

Up to now in the next step towards ventricular contour detection we use derivatives of Gaussian -shaped filters of different sizes [1,3]. Gaussian filter is powerful because it does not introduce zero-crossing as one move to coarser scale, but the zero-crossing scheme is sensitive to noise [1,7]. It is known that the morphological opening filter, also, possesses

that monotonic property, and moreover has some other advantages such as to be computationally less expensive or to allow simple and parallel implementation [7,8]. Moreover the images preprocessed by the anisotropic diffusion technique constitute an ideal starting point for the use of morphological approach to ventricular contour detection.

Mathematical morphology is based on set theory by considering objects in an image as sets [7-10]. Its basic operation are usually defined between two sets, X and B, where X is the set we want to examine and B is called the structure element. The structure elements can have different shapes and sizes according to our purpose. Each specific structure element B filters out further a specific kind of information we are not interested in. The four basic morphological operations are erosion, dilation, opening and closing. Let X, B be subsets in the space E^N . The erosion and dilation of X by B are denoted by $X \ominus B$ and $X \oplus B$, and are defined by

$$X \ominus B = \{z \in E^N \quad z + b \in X, \text{ for every } b \in B\}$$

$$X \oplus B = \{z \in E^N \quad z = x + b, \text{ for some } x \in X \text{ and } b \in B\}$$

The opening and closing of X by B are denoted by $X \circ B$ and $X \bullet B$, and are defined by

$$X \circ B = (X \ominus B) \oplus B$$
$$X \bullet B = (X \oplus B) \ominus B$$

The opening and closing can be viewed as smoothing operations. They cause the object to be shrunk and reexpanded to essentially its original size, but with its contours smoothed. In particular, the opening causes details which project outward from the object and are smaller than half the size of the structure element to disappear. Similarly the closing causes small details inside or projecting into the object to be smoothed. It is worth noting that the operations involved in the morphological trasformations are simple; they require only a comparison of a pixel with its neighbors, as defined by the structuring element, and computing that amounts to a logical AND/OR. From the previous operations it is simple to obtain edge operators by considering the difference between the original object and its dilated or erosed version, that is:

$$E1 = X - (X \ominus b)$$
$$E2 = (X \oplus B) - X$$
$$E3 = (X \oplus B) - (X \ominus B)$$

If we want to improve the effectiveness of the edge operator, we may substitute the image X by a prefiltered one by means of opening and closing smoothing operator. The effect of the opening is to eliminate small spikes in the image function by eroding them away with a larger structuring element and then dilate back the rest of the image. Similarly the closing operation will fill in small holes in the image function.

4. RESULTS

We present here some experiments for contour detection of the left ventricular cavity in a parasternal view. This view is commonly used by the cardiologist to evaluate the functionality of the left ventricle: quantitative evaluations of the characteristics of contraction and relaxation of the various regions of the ventricular wall are made on the basis of the analysis of the ventricular detected contours in a sequence of frames, . These images constitute a portion of 256 x 256 pixels taken from the original echo images. Images in Figures 1 to 6 refer to a left ventricular end-systole frame taken from a parasternal view sequence. In Figure 1 the results obtained by filtering the original image (upper left part) with a median filter 5 x 5 (upper right part) are shown. Such prefiltered image has been processed by means of the anisotropic diffusion technique. We note that in the resulting image (lower right part) a smoothing had been produced inside the homogeneous regions, while the contrast close to the contours had been stressed. A thresholding operation has been performed on that image to obtain a binary one (lower left part) to enhance the ventricular cavity. Such binary image has been processed by using the morphological approach. Experimentation results have suggested to use a 3 x 3 pixel structuring element. Figure 2 reports the binary image (upper left part), the result of erosion (upper right part), the successive dilation (to obtain an opening operation) (lower right part), and the subtraction of the eroded image from the opened one (lower left part). Figure 3 shows the same binary image (upper left), the dilated image (upper right), the erosed one (lower right), and the subtraction of the eroded image from the dilated one (lower left).The obtained contour superimposed to the original image is reported in Figure 4 ; we observe that the latter procedure does not eliminate little artifact inside the ventricle, but the resulting contour

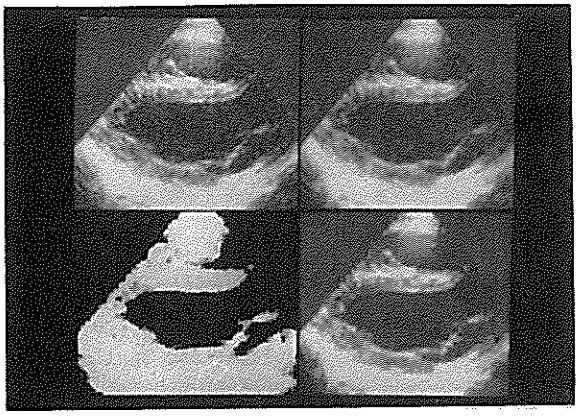


Figure 1. Parasternal view of the left ventricle (systole) (upper left). Median filtering 5x5 (upper right). Anisotropic diffusion technique (lower right). Binary image (lower left).

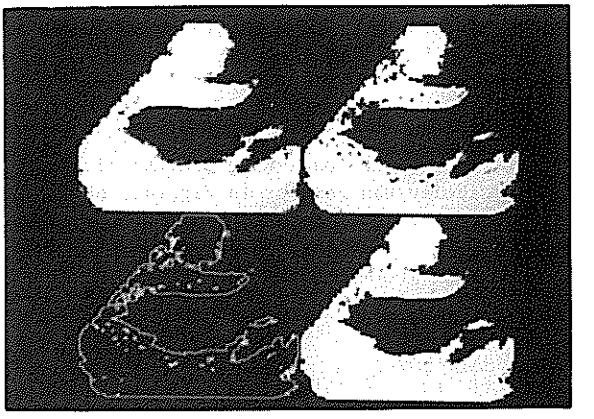


Figure 2.Binary image (upper left). Erosion (upper right). Opening (lower left). Edge detection (lower right).

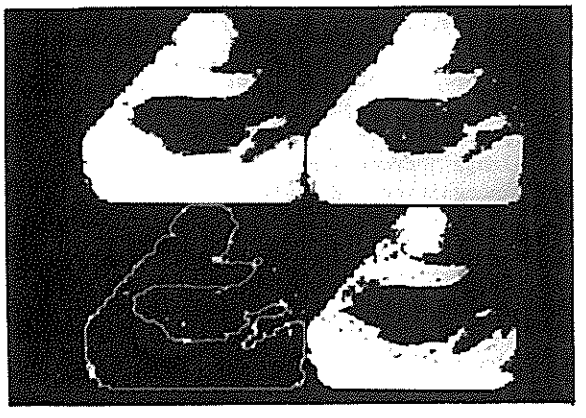


Figure 3.Binary image (upper left).Dilation (upper right).Erosion (lower right). Edge detection (lower left).

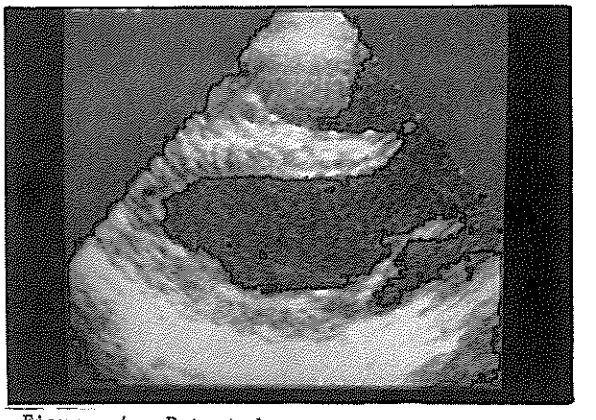


Figure 4. Detected contour superimposed to the original image.

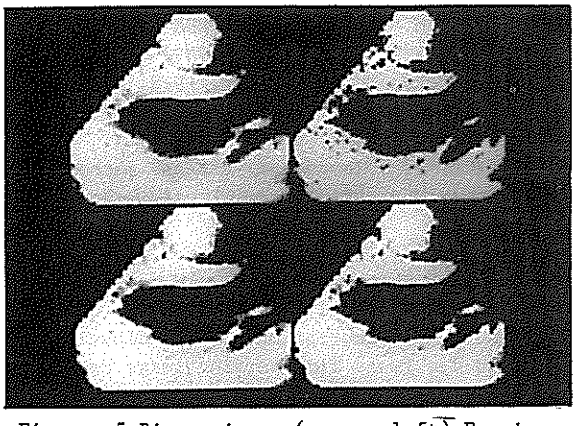


Figure 5.Binary image (upper left).Erosion (upper right). Opening (lower right). Gaussian filtering (lower left).

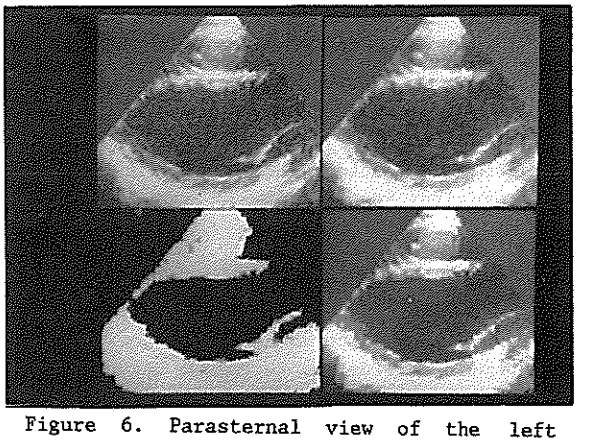


Figure 6. Parasternal view of the left ventricle (diastole) (upper left). Median filtering 5x5 (upper right). Anisotropic diffusion technique (lower right). Binary image (lower left).

is more regular than that reported in Figure 2. Figure 5 shows how the opening procedure (upper & lower right) produces results similar to those obtained by using a Gaussian filter (lower left), as known [7]. Figures 6 to 8 show results obtained by analyzing a diastolic frame. In particular the contour in Figure 8 has been obtained after prefiltering with closing.

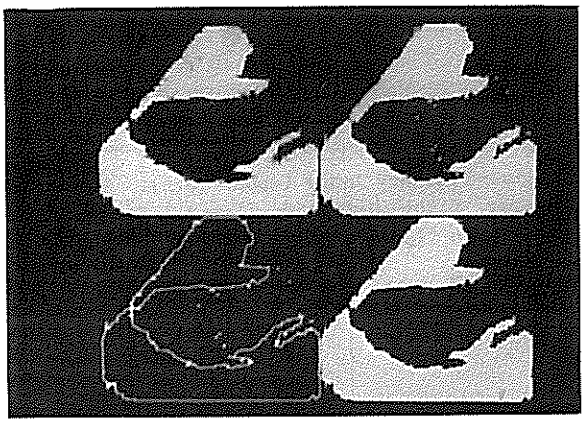


Figure 7. Binary image (upper left). Dilation (upper right). Closing (lower right). Edge detection (lower left).

Figure 8. Detected contour superimposed to the original image.

5. CONCLUSIONS

Our preliminary experimentations show that the anisotropic diffusion technique together with the morphological approach may give good results with short run time also in a 3-D environment, that is taking into account sequences of echo-frames. Moreover, it will be worthwhile to devise a parallel implementation.

ACKNOWLEDGEMENTS

The authors would like to thank Cecilia Fariselli, Gianluca Massalini and Stefano Capellini for their active cooperation in software development.

REFERENCES

[1] Torre,G. and Poggio,T.A.,IEEE Trans. Pattern Anal. Machine Intell. 8 (1986) 147.

[2] Lamberti,C.,Guidazzoli,A. and Sgallari, F., Image regularization for echocardiography digital processing, in: Cappellini V. (ed.), Proc. Time Varying Image Processing and Moving Object Recognition (Elsevier, Amsterdam, 1990) pp. 173-180.

[3] Marr, D. and Hildreth, E. Proc. Royal Soc. Lond. B., 207 (1980) 187.

[4] Torres,L.,Sangra',E.,Gasull,A. and Sallent, S., A new algorithm for automatic border detection of two-dimensional echocardiographic images , in: Cappellini V. (ed.), Proc. Time Varying Image Processing and Moving Object Recognition (Elsevier, Amsterdam, 1990) pp.181-188.

[5] Perona, P. and Malik, J., A network for multiscale image segmentation, in: Proc. 1988 IEEE Intern. Symp. on Circuits and Systems (IEEE Press, New York, 1988) pp. 2565-2568.

[6] Lamberti, C., Lusvardi, S. and Truzzi, C., Workstation for 2-D echocardiography image processing, in Ripley K.L.(ed.), Computers in Cardiology (IEEE Comp. Soc. Press, New York, 1989) in print.

[7] Chen, M. and Yan, P. ,IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 694.

[8] Haralick, M., Stenberg, S.R. and Zhuang, X., IEEE Trans. Pattern Anal. Mach. Intell. 9 (1987) 532.

[9] Serra, J., Image analysis and mathematical morphology (Academic Press, New York, 1982).

[10] Cappellini, V., Daniels, N.J., Raspollini, C. and Venetsanopulous, A.N., Applications of morphological filters in edge detection and shape description , in: Cappellini V. (ed.), Proc. Time Varying Image Processing and Moving Object Recognition (Elsevier, Amsterdam, 1990) pp.57-67.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

IMAGE REGISTRATION OF EYE FUNDUS ANGIOGRAMS

Ana Maria R. S. Faria de Mendonça *, Aurélio J. C. Campilho *
Francisco J. O. Restivo *, José M. Rodrigues Nunes **

* DEEC / Faculdade de Engenharia da Universidade do Porto
** Hospital Geral de S.António - Serviço de Oftalmologia

In this paper a new technique for image registration is presented. This technique was specially developed by the authors for the registration of eye fundus angiograms. The new procedure is confronted against methods based on image correlation and sequential similarity detection and some results are presented and discussed.

1. INTRODUCTION

The automatic comparison of two images for the detection of changes is a common procedure in Image Processing and Pattern Recognition. The comparison is only possible if the images are registered.

In recent years, several image registration techniques have been developed. Their importance can be evaluated by the large range of imaging applications where they are required. Radar images, images of aerial photography, "Landsat" images and different fields of Medical Imaging, like Radiology, Nuclear Medicine, Dermatology and Ophthalmology, are just some examples.

Image registration is not an easy task. To detect "similar" regions in two images, it is necessary to have an appropriate similarity criterion. An ideal criterion, to be effective, must have the following features: insensible to image translation, rotation and scaling; insensible to variations of local intensities and small changes of local contrast; computational time as small as possible, to be compatible with interactivity.

To attend to all these features simultaneously is very difficult, even impossible. In each specific situation, the choice of a similarity criterion which is best adapted to the case in study is required, taking in consideration the type of images as well as the registration purpose.

In Ophthalmology, the study of several pathologies requires the analysis of time evolutions, based on images taken at different time instants. Images may be used to detect slow changes, such as those that occur in drusen [1], nerve fiber layer [2], optic disc cupping and pallor [3], or quick changes, as in measuring dilution curves in fluorescein angiography [4]. Although time intervals between images may vary from several years to just a few seconds, the characterization of time evolutions always implies a comparative study, that is simplified if images are registered.

Correlation and Sequential Similarity Detection (SSD) are the most used methods for eye fundus image registration [4], [5]. Both employ similarity measures based on pixel values of the areas to compare. These techniques give good results when applied to some retinal images. However, in the registration of angiograms that kind of measures is useless because of the intensity changes in vessels during an angiography.

In this paper a new method for image registration is presented. This technique was specially developed by the authors for eye fundus angiograms registration. A similarity criterion has been conceived that tries to avoid the referred restriction of correlation and sequential similarity detection, by using intensity independent information based on the position of characteristic pixels.

The new registration method is described in the next section. The results from the new procedure and two

other methods (Correlation and SSD) are presented and discussed in sections 3 and 4.

2. EDGE / POSITION SIMILARITY DETECTION

Image registration techniques that have been proposed so far, differ essentially in the criteria used to evaluate the "similarity" between areas to be compared. However, the registration procedure is common to several techniques. A JxK reference window is defined in the first image. Another area, of dimension MxN (search area) is also selected in the second image. This window is usually larger than the reference window and is supposed to include the "similar" region to be detected. The reference window is moved across the search area, and the similarity measure is calculated for every point, according to the criterion previously established.

The methods based on correlation use the correlation coefficient (normalized or unnormalized) as the similarity measure, and the registration point is associated with the maximal calculated value [7], [8], [9].

In sequential similarity detection the registration position corresponds to the minimum of the sum of the absolute differences between pixel values of the reference and correspondent search areas. The differences may be calculated for all pixels or just for those that verify some predefined conditions [5], [6].

All these techniques have a common characteristic: the similarity measure depends on pixel intensities of the areas to be compared. So, their application is limited to images pairs where the relations between each pixel value and its neighbourhood are approximately maintained.

The specific characteristics of eye fundus angiographic images and particularly the intensity changes in vessels along the sequence, prevent the above referred methods from being useful in registration .

In the proposed method, the similarity criterion, is invariant with changes of local intensity in vessels, and uses information dependent on the position of the retina vascular edges.

This edge/position similarity detection procedure can be described by the following steps:

. interactive definition of reference and search areas;
. image filtering of selected areas, followed by the identification of characteristic points (those with stronger edges);
. calculation of similarity measures, based on the number of coincident localizations of characteristic points;
. best match selection, corresponding to the maximum of the number of coincidences.

The interactive definition of reference and search areas is illustrated in figure 1a). On the left, the first image and the selected reference area are represented; on the right, the search area, defined in the second image, can be observed. Figure 1b) shows the result of the registration procedure, with the reference area inserted into the "similar" region of the second image.

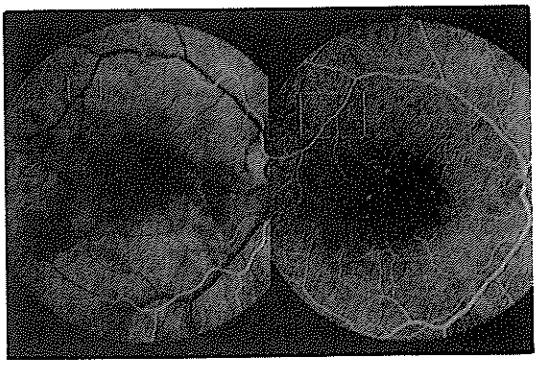


Figure 1a) - Reference and search areas selection.

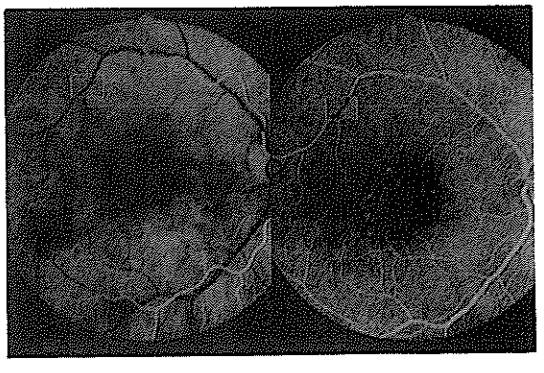


Figure 1b) - Reference area inserted into the "similar" region.

Registration performance among methods

TABLE I

Meth. / Img.	Corr. |Δx|,|Δy|	SSD |Δx|,|Δy|
I.1	0, 0	0, 0
I.2	0, 0	0, 1
I.3	0, 0	0, 0
I.4	2, 2	1, 3
I.5	16,15	4, 0
I.6	12,11	10,14

Within an angiographic sequence

TABLE II

Meth. / Img.	Corr. |Δx|,|Δy|	SSD |Δx|,|Δy|
II.1	0, 0	1, 2
II.2	0, 0	0, 0
II.3	0, 0	1, 2
II.4	0, 1	0, 0
II.5	1, 1	0, 0
II.6	0, 0	2, 2

Between different angiographic sequences (same phase)

TABLE III

Meth. / Img.	Corr. |Δx|,|Δy|	SSD |Δx|,|Δy|
III.1	16,14	19, 12
III.2	28,11	4, 1
III.3	14,16	3, 3
III.4	7, 5	2, 5
III.5	14,13	3, 3
III.6	4, 5	11, 18

Between different angiographic sequences (different phases)

3. RESULTS

The new registration procedure was evaluated on several angiographic sequences, some of them taken at time instants separated by a few years. The results were confronted against those obtained with the other methods (correlation and SSD). The correlation method uses the normalized correlation coefficient as a measure of similarity [9]. Sequential similarity detection algorithm is based on the method referred in [5].

For validation of the proposed method a manual registration procedure has been applied to the region of interest. In all areas very small differences (± one pixel) have been observed, even in regions where there are significant intensity changes.

Aiming a comparative study, x and y absolute offsets have been measured with different methods (Tables I, II and III), using as reference the final image position obtained with the new method (where $|\Delta x|=0$ and $|\Delta y|=0$). Three situations were evaluated: a) registration in the same angiographic sequence (Table I); b) registration in the same phase of different angiographic sequences (Table II); c) registration in different phases of different angiographic sequences (Table III). Image I.1 was used as a reference for all methods.

Figures 2 and 3 show typical results of angiographic image registration using a) the new method, b) correlation and c) sequential similarity detection.

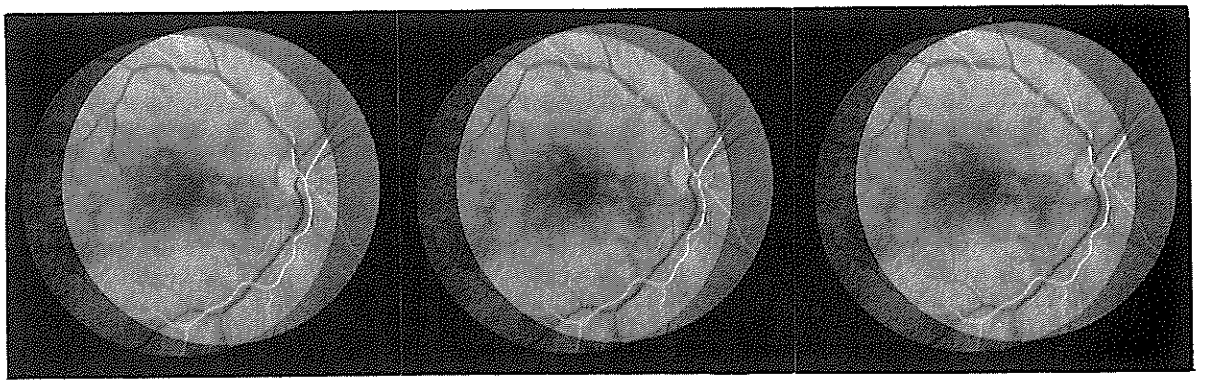
As can be observed in figure 2, the three methods present similar results in the registration of images of the same angiographic phase. In figure 3 only the new method produced an acceptable result because images belong to different phases.

4. DISCUSSION

The registration of eye fundus angiograms - the main goal of this work - has been fully achieved. Furthermore, the analysis of Tables I, II and III shows the superior performance of the proposed method.

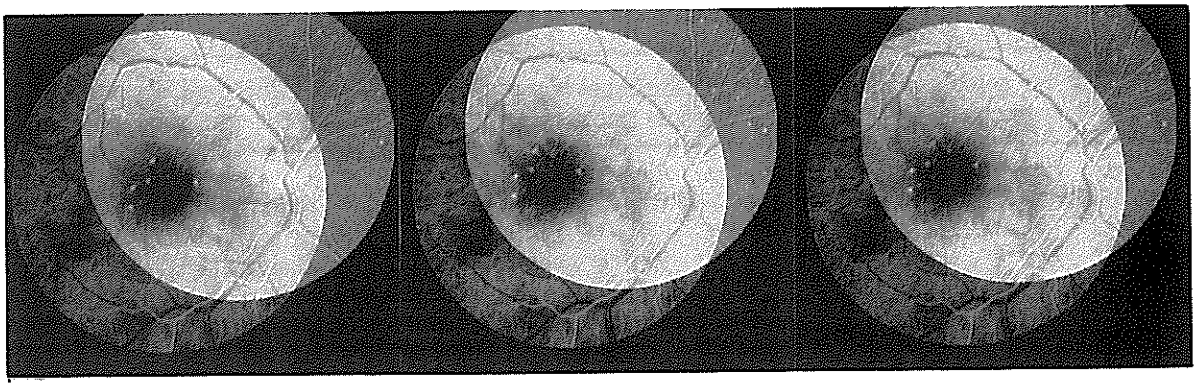
Although the new method was specially developed for eye fundus angiograms registration, its range of application is very wide. In fact, it can be directly used with retinograms and other types of images.

The developed algorithm has shown to be very efficient, with execution times less than those achieved with sequential similarity detection. Correlation is much more time consuming.



a) New method b) Correlation c) SSD

Figure 2 - Registration of angiograms of the same angiographic phase.



a) New method b) Correlation c) SSD

Figure 3 - Registration of angiograms of different angiographic phases.

The new registration procedure is simple and does not use complex operations, like rotation or scaling. The translational registration has demonstrated to produce good results, particularly in the search area and its neighbourhood. Nevertheless, in a few cases, in areas far from the comparison region, small misalignments have been observed. Their compensation would involve the use of more elaborated transformations, that are not realistic for routine use.

ACKNOWLEDGMENTS

This work has been partially funded by INIC under grant CEEUP/LA3 and by JNICT under contract 900-86-41.

REFERENCES

[1] Peli, E. and Lahav, M. , Ophthalmology (1986) 1575.

[2] Peli, E., Hedges, T.R. and Schwartz, B.,Acta Ophtalmol. (1986)113.

[3] Nagin, P., Schwartz, B. and Raynolds, G., Ophthalmology (1985) 547.

[4] Nagin, P., Schwartz, B. and Nanba, K., Ophthalmology (1985) 243.

[5] Peli, E., Angliere, R.A. and Timberlake, G.T., IEEE Trans. on Medical Imaging (1987) 272.

[6] Barnea, D.I. and Silverman, H.F., IEEE Trans Computers (1972) 179.

[7] Pratt, W.K., Digital Image Processing (Wiley, New York, 1978).

[8] Aggarwal, J.K, Davis, L.S. and Martin, W.N., Proceedings IEEE (1981) 562.

[9] Gonzalez, R.C. and Wintz, P., Digital Image Processing, (Addison-Wesley , 1987).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A STATISTICAL APPROACH TO THE DETECTION AND TRACKING OF MOVING OBJECTS IN AN IMAGE SEQUENCE

Patrick Lalande and Patrick Bouthemy

IRISA / INRIA-Rennes
Campus de Beaulieu, 35042 Rennes Cedex, France

This paper describes an original framework to solve the basic issue of motion detection in an image sequence. It is based on statistical models, i.e. spatio-temporal Markov fields, and a global bayesian formulation. This method does not need any knowledge on the intensity distributions of respectively moving object projections and background. It can handle objects of any size and any motion. All the resulting computations are very local. Besides this scheme can lead to an attractive tracking procedure. A lot of experiments on real image sequences have been performed.

1. Introduction

The analysis of apparent motion in an image sequence includes several aspects. Indeed four main topics can be distinguished : motion detection, motion estimation (computation of velocity fields), motion segmentation (edge-based or region-based), motion interpretation (derivation of quantitative 3D motion parameters or determination of qualitative labels), [1-4]. This paper addresses the basic issue of motion detection in an image sequence. If the output of a motion detection module corresponds to the simplest motion information level (binary map: moving parts versus static ones in the image plane), nevertheless this early processing step is of great use in a large class of problems related to dynamic scene analysis. Indeed, when the camera is static, detecting moving objects in the scene comes to detecting moving regions in the image plane. A variety of applications are concerned with this problem, such as traffic control, remote surveillance of industrial areas, biomedical studies or target tracking, [5].
This paper describes an original framework to solve the motion detection issue based on statistical models: i.e. *spatio-temporal Markov fields.* This method generalizes a first attempt of this kind we have recently presented in [6]. Substantial modifications have been introduced, which contribute to clear up several shortcomings of the previous algorithm. This new version is able to handle textured moving objects and overlapping cases. By this last term, we mean situations where successive projections in time of a moving object overlap each other in the image plane, and occlusion situations between different moving objects. The ability to cope with overlapping cases avoid to pay attention to the time sampling rate of the processed image sequence with respect to the size and speed of moving objects of interest. Besides the new way of modeling and using temporal contextual information enables to easily track moving regions along the image sequence.

2. Problem statement

If moving objects are present in the scene, changes in time will obviously occur in the image intensity array. In turn, when the camera is static, temporal intensity changes can be related mainly to motion. Nevertheless, motion detection cannot be reduced to temporal change detection. In particular a moving object gives raise to three kinds of change regions; first one corresponding to uncovered background, second to covered background, third to the overlap of object projections (by the way, this last sub-class is often partially extracted by a temporal change detector). But what is sought for (only but completely) in every image are projections of moving objects, (also called moving object masks). Usual methods first extract successive temporal change maps, then try to recover projections of moving objects by applying some heuristics, [3,7]. We have adopted a quite different approach. The main idea was to consider the motion detection issue as a whole and to rely on a properly defined modeling step. To this end this issue is stated as a *statistical labeling* problem according to a global bayesian formulation. The modeling step consists in expressing local contextual properties of moving masks by making use of Markov random field models. Such an approach has already been proved relevant to other topics related to motion analysis, as reported in [8] for scene segmentation according to motion information, and in [9,10] for optic flow estimation. In a labeling problem, two sets of elements must be defined: *observations*, (i.e. data to be considered); *labels* (i.e. primitives to be extracted).

3. Observations and primitives

The remarks in Section 1 naturally lead to take as observation the temporal derivative of the intensity function f, that is the "temporal variation signal". Let o_t denote the observation array at time t, and $p = (x, y)$ a pixel, we have: $o_t(p) = \partial f(p)/\partial t$. In fact temporal derivatives will be approximated by finite differences between time t and time $t - dt$:

$$o_t(p) = \tilde{f}_t(p) = f(x, y, t) - f(x, y, t - dt) \qquad (1)$$

Moreover another set of information is considered: the logical map of temporal changes between time t and

time $t - dt$: $\bar{o}_t$. $\bar{o}_t(p) = 1$ if a temporal change of the intensity function is validated at point p, $\bar{o}_t(p) = 0$ otherwise. It is obtained by an operator able to detect even weak temporal changes. The intensity is locally modeled by a linear function with an additive white gaussian noise of constant variance. Changes in the model parameter values between two windows centered at the same point but at two successive times, are validated by a likelihood test similar to the one reported in [11]. This kind of change detector is relevant in the context of motion detection, if we assume that no illumination variation occurs between time $t - dt$ and t. Nevertheless another temporal change detector could be chosen as the one described in [12] if such situations have to be handled. Then we deal with two sets of observation arrays o_t and $\bar{o}_t$, but they take place in the statistical framework in different manners as explained further on. On the other hand we consider primitives directly tied to the type of image content we want to delineate. Therefore the set of label values consists of two symbols, $\Omega = \{a, b\}$, a for belonging to a moving object mask, b for belonging to static background. Spatio-temporal models are associated with these primitives. These models must express what properties the solution is supposed to have (that forms the *regularization effect*). Let us first give some intuitive insights to the modeling step. All masks of moving objects obviously share some intrinsic basic spatio-temporal properties. They must show sufficient spatial coherence and their successive positions in time obey a certain law. This can be expressed in terms of required spatio-temporal contextual configurations. Markov field models represent very efficient and well-defined means to mathematically formulate this problem, [13]. Let us denote the label field at time t by e_t. This field is modeled as a Markov field in space *and* time; besides the markovian property in time is assumed in both directions along the time axis. This will be explained in more details in the two next sections.

4. The decision criterion

Indeed we need contextual information from the close past and from the near future to identify the label field at time t. That is the reason why we consider label fields in pairs in the identification process. The solution to the labeling problem is formulated according to the maximization of the joint likelihood of observed and unknown variables:

$$\max_{e_{t-dt}, e_t} \quad \xi(e_{t-dt}, e_t, o_t, \bar{o}_t) \tag{2}$$

The best interpretation in terms of moving object projections must have the greatest a posteriori probability given the observations at hand. This statistical approach also permits to properly deal with noise-corrupted observations. One attractive aspect of this approach is that we can build an explicit version of ξ using the equivalence between Gibbs distributions and Markov fields, as primarily emphasized in the context of image processing in [13]. More precisely, this means that the distribution ξ can be expressed as follows:

$$\xi(e_{t-dt}, e_t, o_t, \bar{o}_t) = \frac{1}{Z} exp[-W(e_{t-dt}, e_t, o_t, \bar{o}_t)] \tag{3}$$

where Z is a normalizing factor. The so-called energy function W is given by: $W = \sum_{c \in C} V_c$,
where C denotes the set of cliques associated to the chosen neighborhood system describing interactions between the different variables and V_c is called a potential function which is locally defined on the cliques. A clique is a subset of sites (here sites are pixel locations) which are mutual neighbors.Then finding the Maximum A Posteriori estimate comes to the problem of minimizing the global energy function W, which in turn can be decomposed into local potentials.

5. The energy function

Local contextual models can thus be related to potential functions defined on so-called cliques. We consider the following spatio-temporal neighborhood system: a 3x3 spatial neighborhood centered in (p, t) and one-to-one temporal connections from (p, t) toward $(p, t - dt)$ and $(p, t + dt)$. As far as spatial cliques c_s are concerned, we only take into account the four ones comprising two sites. Spatial potentials V_{c_s} have been defined in such a way as to favouring homogeneity of the label field, that is to have a spatial regularization effect, (e.g. to eliminate isolated points and to fill in missing points inside the masks). They are of logistic kind; that is equal to a predefined level, (resp. $-\beta_s$ and β_s), according to the label configuration of the clique at hand, (resp. same labels and different labels), knowing that a negative value encourages the corresponding configuration. It is not necessary here to introduce a complementary edge-site system to take into account discontinuities as in [13], because of the existence of the temporal cliques. These potentials summed over spatial cliques form a first energy function W_s.

The temporal cliques also contain two sites. The potential functions tied to these cliques are far more specific to the problem at hand. Nevertheless they are again of logistic kind. They contribute to determine which temporal configurations between (a, a), (a, b), (b, a), (b, b), at pixel p, at two successive times, are encouraged and which are discouraged, according to $\bar{o}_t(p)$ considered as a deterministic external information. They are described in Table 1. The table content requires the following comments. First, the preference given to the configuration $(a, a, 1)$ over $(a, a, 0)$ must be related to the fact that the temporal change detector which is used, is claimed to be sensitive to even weak temporal changes. Second, the high positive potential β'_τ, with $\beta'_\tau \gg \beta_\tau$, assigned to the configuration $(a, b, 1)$ may be surprising. As a matter of fact, these two points contribute to make overlapping situations be correctly handled. As long as a temporal change is detected the site is still supposed to belong to a moving object mask, as far as the temporal clique is concerned. The situation where a site actually belongs again to the static background (label value b) of course does not remain unsolved, but its treatment is only postponed to the next time. Indeed the output of the algorithm at time t is the estimated label field $\hat{e}_{t-dt}$; or in other terms the final estimate $\hat{e}_t$ is only available at time $t + dt$. These temporal potentials V_{c_τ} lead to a second energy function W_τ.
The third energy function W_ϵ will express the consis-

tency between observations and current estimates of the primitives. It is easily derived knowing that the relation between these two sets is defined by:

$$o_t(p) = \psi(e_{t-dt}(p), e_t(p)) + n(p) \qquad (4)$$

where ψ can take a value among three possible ones: 0 for (b,b), m_1 for (a,a) and m_2 for (a,b) or (b,a). These parameters can be either predefined or locally estimated on-line. n is a white (in space and time) zero-mean Gaussian noise of constant variance σ^2. The corresponding energy function W_c is then given by:

$$W_c = \frac{1}{2\sigma^2} \sum_p [\tilde{f}_t(p) - \psi(e_{t-dt}(p), e_t(p))]^2 \qquad (5)$$

The noise variance is estimated once at the beginning of the processed image sequence.
Finally we get the total energy function:

$$\begin{aligned} W(e_{t-dt}, e_t, o_t, \bar{o}_t) &= W_c(e_{t-dt}, e_t, o_t) + W_s(e_{t-dt}) \\ &+ W_s(e_t) + W_\tau(e_{t-dt}, e_t, \bar{o}_t) \end{aligned} \qquad (6)$$

The optimal map of moving object masks will correspond to the lowest value of energy W.
To minimize W, we use an iterative deterministic method as in the early version described in [6]. It yields a very good trade-off in our case between computation speed and result quality the more so as we can achieve an appropriate initialization stage as explained below. Moreover the optimization method is completed by an efficient procedure for iteratively selecting sites to be visited, as suggested in [14]. This procedure allows to constantly focuse on ill-labeled sites. Let us point out that all computations are very local. As already pointed out, (in particular the necessity to introduce information from the past, the present and the future before concluding), two successive label fields are always simultaneously considered. This leads to an optimization process of every label field e_t done in a two-pass manner. First we derive a first estimate $\check{e}_t$ when considering pair (e_{t-dt}, e_t). Second, considering pair (e_t, e_{t+dt}), we update $\check{e}_t$ and we get the final estimate $\hat{e}_t$; of course the first estimate $\check{e}_{t+dt}$ is simultaneously obtained (knowing that e_{t+dt} is also initialized by $\check{e}_t$). Afterwards the same holds for e_{t+dt} and so on. This prediction scheme, associated with the markovian property of the label field (in space and time) enables to define a very simple kind of mask tracking as described in the next section.

6. Mask labeling for tracking purpose

As a matter of fact, this approach is not limited to delivering the binary map of moving masks at each time. The labeling process can take place at a upper level. Up to now we have dealt with the binary early level concerning each pixel site. We can also consider the mask level. The ability to label each mask entity along the sequence could then naturally lead to a tracking procedure. Indeed the mask labeling level is reachable; it is essentially a simple matter of recursive number allocation to subsets of points connected in space and time. First, an initializing step is needed; that is, given the first estimated binary label field $\hat{e}_{t_0}$ a different number is assigned to each connected subset of a-labeled points. Let us assume that the mask labeling process has been achieved until time $t-dt$; it is pursued at time t as follows. Let q_k be a reference point among points with number k in the image at time $t-dt$, (q_k can be for instance the mass centre of these points, but a more judicious choice may have to be done). The same number k is assigned to the point p_q in the image t belonging to the temporal clique of q_k. This assumes that the intersection of two successive projections in time of a given moving object is not empty. Then starting from point p_q, the assignment of number k in the image t is propagated to the a-labeled points connected through the spatial cliques. When this is achieved, number $k+1$ is taken into account, and so on.
Two specific cases may happen: the merging and the splitting of moving masks. They can correspond for instance to the crossing of two different moving objects. The first case will be detected when two different numbers k and k' will be present in a spatial clique. This conflict situation is easily solved by "equaling" k' to k (i.e. by creating a link). The second case will provide the same situation as the one encountered when a new moving object is appearing. A subset of connected a-labeled points in the image at time t will remain without any number assigned at the end of the process. It will then receive a new number. This tracking-like stage has been separately presented to make the explanation easier. Actually numbers k can be considered as supplementary primitives which can be straightforwardly included in the modeling step. Therefore detection and tracking can be nearly simultaneously carried out. Indeed what is worth noticing in this scheme is not the number assignment procedure as it is, which is a common one, but the way the markovian framework can directly manage it. Another important feature is the following. The tracking process is here performed on the complete silhouette of moving projections. Hence a global description of moving entities (for solid as well as articulated or deformable objects) is directly reachable, contrary to other tracking techniques relying on other primitives as edge lines for instance. Of course a long-term memory must be somehow added to this scheme to performe a complete tracking process.

7. Results

Numerous experiments with several image sequences depicting outdoor scenes have been processed. Results are quite fine. We present here one example. Fig. 1 shows three (not successive) images from the input sequence taken at times t_1, t_2, t_3. The camera is static and is looking at an highway scene. All the cars are moving. In the results obtained with the markovian approach, Fig.2, the stationary background is free of spuriously detected points and the extracted regions rather well correspond to the real masks of the moving objects (shadows included). The time interval dt which is considered here corresponds to the standard video rate (i.e., $1/25s$). Let us point out that the method has been able to cope with objects of very different size (from large mask in the foreground to small ones for the cars in the distance) and of different apparent motion magnitude. This has been largely confirmed by the experiments carried out on an important set of data (more

than one thousand images). It has also been found that the parametrization of the model is not a critical problem for this motion detection issue. The same set of parameter values can be used for the different image sequences. The output results vary very smoothly and slightly when parameter values are changed. Moreover the number of parameters is rather small.

We have described a general, modular, model-based and robust method for motion detection in an image sequence, which besides can lead to an attractive tracking procedure. This method does not suppose any a priori knowledge on the respective intensity values of moving object projections and background; it does not require anymore any identification of the background intensity distribution. It can also handle objects of different size and different motion; hence it is completely independent of the temporal sampling rate of the image sequence to be processed. All the resulting computations are very local and an efficient fast implementation is indeed reachable, which allows an effective use in practical situations.

References:

[1] **H.-H. Nagel**, Image sequences -ten (octal) years- from phenomenology towards a theoretical foundation, *Int. Jal of Pattern Recognition and Artificial Intelligence*, Vol. 2, No 3, 1988, pp.459-483
[2] **J.K. Aggarwal** and **N. Nandhakumar**, On the computation of motion from sequences of images - a review, *Proceedings of the IEEE*, Vol. 76, No 8, Aug. 1988, pp.917-935
[3] **R. Jain**, Dynamic scene analysis, in *Progress in Pattern Recognition 2*, L. Kanal and A. Rosenfeld (eds.), North-Holland, 1985, pp.125-167
[4] **P. Bouthemy**, Modèles et méthodes pour l'analyse du mouvement dans une séquence d'images, *Technique et Science Informatiques*, Vol. 7, No 6, 1988, pp.527-546
[5] **T.S. Huang** (ed.), Image sequence processing and dynamic scene analysis, *NATO-ASI Series*, Vol. F2, Springer-Verlag, 1983
[6] **P. Bouthemy** and **P. Lalande**, Motion detection in an image sequence using Gibbs distributions, *Int. Conf. on Acoustics, Speech and Signal Processing*, Glasgow, May 1989
[7] **J. Wiklund** and **G.H. Granlund**, Image sequence analysis for object tracking, *Proc. 5th Scandinavian Conf. on Image Analysis*, Stockholm, June 1987, pp.641-648
[8] **D.W. Murray** and **B.F. Buxton**, Scene segmentation from visual motion using global optimization, *IEEE Trans. on PAMI*, Vol. 9, No 2, March 1987, pp.220-228
[9] **J. Konrad** and **E. Dubois**, Multigrid Bayesian estimation of image motion fields using stochastic relaxation, *Proc. 2nd Int. Conf. on Computer Vision*, Dec. 1988, pp.354-362
[10] **F. Heitz** and **P. Bouthemy**, Multimodal motion estimation and segmentation using Markov random fields, *Proc. 10th Int. Conf. on Pattern Recognition, Computer Vision Conf.*, Atlantic City, June 1990
[11] **Y.Z. Hsu, H.-H. Nagel** and **G. Rekers**, New likelihood test methods for change detection in image sequences, *Computer Vision, Graphics and Image Processing*, Vol. 26, 1984, pp.73-106
[12] **K. Skifstad** and **R. Jain**, Illumination independent change detection for real world image sequences, *Computer Vision, Graphics, and Image Processing*, Vol. 46, 1989, pp.387-399
[13] **S. Geman** and **D. Geman**, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. on PAMI*, Vol. 6, No 6, Nov. 1984, pp.721-741
[14] **P.B. Chou** and **R. Raman**, On relaxation algorithms based on Markov random fields, *Technical Report* No 212, Computer Science Dpt, Univ. of Rochester, July 1987, 28p.

This work was partly supported by the French CNRS Program, "PRC Man-Machine Interface, Vision group" under contract PMFE 88F1 548

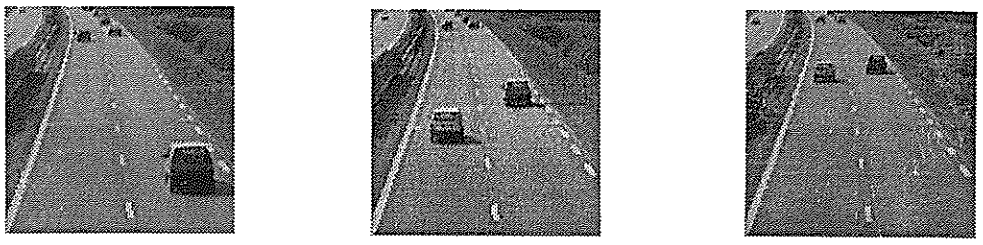


Figure 1: *Three original images (not successive) out of the sequence at times t_1, t_2, t_3.*

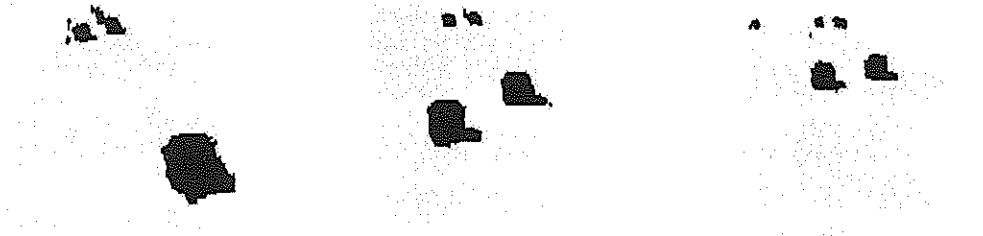


Figure 2: *The moving object masks at times t_1, t_2, t_3; ($\beta_\tau = 100$, $\beta'_\tau = 1000$, $\beta_s = 10$) (The time interval dt indeed corresponds to the video rate).*

$(e_{t-dt}, e_t, \bar{o}_t)$	(b,b,0)	(b,b,1)	(a,b,0)	(a,b,1)	(b,a,0)	(b,a,1)	(a,a,0)	(a,a,1)
$V_\tau(e_{t-dt}, e_t, \bar{o}_t)$	$-\beta_\tau$	$+\beta_\tau$	$+\beta'_\tau$	$+\beta'_\tau$	$+\beta_\tau$	$-\beta_\tau$	$+\beta_\tau$	$-\beta_\tau$

Table 1: *The temporal potentials ($\bar{o}_t = 1$, means temporal change, $\bar{o}_t = 0$ no change)*

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Moving Object Segmentation Based on Adaptive Reference Images

Klaus-Peter Karmann, Achim v. Brandt, Rainer Gerl
Siemens AG, Corporate Research and Development,
Information and Knowledge Processing, Image Processing,
Otto-Hahn-Ring 6, D-8000 München 83, Fed. Rep. Germany.

Abstract:

For applications of computer vision in robotics and traffic monitoring, where several moving objects have to be detected in a static environment, various methods for change detection have been proposed. Early approaches based on consecutive frame differences are sensitive to noise, require ambiguous postprocessing and yield poor object segmentations. Recent time recursive methods, based on adaptive reference images, are not robust and may become unstable in certain situations. In this paper an algorithm is presented, which overcomes these difficulties because the computation of the reference images does not depend directly on the decisions made during the detection process.

Introduction

For applications of computer vision in robotics and traffic monitoring, where several moving objects have to be detected in a static environment, various methods for change detection have been proposed. The most widely used method for this task is change detection by differences of consecutive frames [Wik87], where the intensity I(k-1,p) of frame k-1 at image point p is subtracted from I(k,p). Imposing a suitable threshold on the resulting difference D(k,p) yields a binary object mask M(k,p) which indicates the regions where moving objects have been detected. While the detected regions in general do not reflect the true shapes and boundaries of the moving objects if this method is used, it is highly desirable (e.g. for a subsequent motion estimation or object tracking) that the boundaries of the connected regions of M(k,p) coincide with the true object boundaries [Kar89, Hoe88] (moving object segmentation).

To overcome this problem, other authors proposed change detection methods [Don88, Gun88, Kar89] where the current image is compared to a reference image B(k,p) (background), which adapts to slow variations of the environment (changing illumination, e.g. dawn and dusk), and from which the fast variations (especially the moving objects) are eliminated as far as possible.

$$D(k,p) = I(k,p) - B(k,p) \tag{1}$$

$$M(k,p) = \begin{cases} 1 \text{ if } |D(k,p)| > \text{threshold} \\ 0 \text{ else} \end{cases} \tag{2}$$

A common feature of these adaptive reference image methods [Don88, Gun88, Kar89] is the time recursive scheme by which the reference sequence is computed from the original images. The latter are essentially low-pass filtered in time to eliminate all fast variations. The above methods differ, however, in the way the time-dependence of the filter coefficients is determined and use filters of different complexity.

Donohoe [Don88] proposes a time-independent filter that computes the intensities of the reference image as a moving average of the intensities of the images in the original sequence, using the same time-independent filter coefficient for all image points.

$$B(k+1,p) = B(k,p) + g \cdot (I(k,p) - B(k,p)) \quad (3)$$

In his approach moving objects are treated the same way as the nonmoving environment. Consequently the moving objects are eliminated from the reference images only to a certain degree. This can be improved by switching the filter gain to zero [Gun88] (or to a small, nonzero value

$$B(k+1,p) = B(k,p) + g(k,p) \cdot (I(k,p) - B(k,p)) \quad (4)$$

$$g(k,p) = \alpha \cdot M(k,p) + \beta \cdot (1-M(k,p) \quad (5)$$

[Kar89]) where a moving object has been detected in the preceding frame. In [Kar89] a predictor is proposed which increases the adaptivity of the filter and thus improves the performance of the method if faster illumination changes (clouds, snow, rain, etc.) occur.

While substantially improving their predecessors [Wik87, Don88], the proposed procedures [Gun88, Kar89] are not entirely satisfactory, because of serious unexpected problems that arise in certain situations.

If switched on in the presence of moving objects or if applied to situations, where an object, which has currently been regarded as part of the static environment, suddenly starts to move, the procedures of Gunzinger [Gun88] and Karmann [Kar89] may be trapped in a deadlock mode. In these cases, erroneous detections (noise, sudden changes in the environment) do not decay in time, but prevent the system from observing the original image sequence. This way, detection errors may accumulate in time, and the system may eventually become blinded by increasing detection errors.

An improved reference image algorithm

In this paper an algorithm is presented, which overcomes these difficulties because the *computation of the reference images does not (directly) depend on the decisions made during detection process* (object detected or not). Such a scheme can be represented in the same way as the former approaches. We now choose

$$g(k,p) = \frac{V(k,p)}{(D(k,p))^2 + V(k,p)} \quad (6)$$

where the variance V(k,p) is estimated recursively by help of

$$V(k+1,p) = (1 - g(k,p)) \cdot V(k,p) + S(k) \quad (7)$$

where S(k) is an estimate for the system noise.In contrast to the algorithms in [Gun88] and [Kar89] the filter gain in equ. (6) now depends on the difference image instead of the binary object mask. The estimated variance V(k,p) is calculated recursively and measures the local noise power in the reference image. (Since this quantity is also used by the detection process, it has to be computed anyway [Don88, Gun88] and no computational overhead is introduced by this choice). Unlike in former approaches [Gun88, Kar89], decisions made in the detection step do not directly influence the computation of the reference image, which is controlled by the instantaneous difference D(k,p) and the estimated variance V(k,p).

The proposed algorithm is an exact Kalman filter [Lew86], where the measurement noise variance R(k,p) has been chosen as

$$R(k,p) = (D(k,p))^2 \quad (8)$$

and the system noise variance has been denoted by S(k). This quantity has to be specified according to the special performance features desired in various applications. In the following section, the choice of S(k) will be discussed in more detail.

Equ. (8) may be justified by two arguments: If no moving object occludes the background at image point p at time k, the original image I(k,p) and the background image B(k,p) differ only by the random noise of the image acquisition system (camera). In this case, the measurement noise R(k,p) of the Kalman filter should be chosen

equal to the estimated instantaneous power of this random camera noise. The optimal instantaneous estimate for this noise power is given by equ. (8). In the other case, in which image point p corresponds to a moving object at time k, the background intensity B(k,p) cannot be measured at (p,k). This situation may appear to correspond to a complete lack of information, which should be represented by a (practically) infinite measurement noise variance R(k,p). Equ. (5) was found on the grounds of this argument. Since, however, the original image I(k,p) differs from the background B(k,p) just by the difference between the intensities of the object and the background respectively (plus camera noise), the choice of equ. (8) is more reasonable in this case too.

Estimation of the system noise variance

In Kalman filter theory [Lew86] the system noise variance expresses the degree of accuracy (or validity) of the dynamical model used for the time update. The dynamical model leading to equ. (4) can be characterized by

$$B(k+1,p) = B(k,p) + w(0,S(k)) \qquad (9)$$

which means that we expect the background intensity not to change with time, and that all unexpected changes can be described by a random noise w of zero mean and variance S(k). Consequently, S(k) should be varying in time according to the illumination changes and should be independent of the image point coordinate p, because the validity or accuracy of the deterministic part of equ. (9) should not depend on p and there is no reason to expect a dynamical model different from equ. (9) at different times k.

Since S(k) measures the reliability of the time update of a Kalman filter, it should be related to $(D(k,p))^2$(equ. (1)) at every image point, where no moving object occludes the background (or, in other words, where I(k,p) is a good measurement for B(k,p)). This reasoning leads to the choice

$$S(k+1) = S(k) + \alpha \cdot (\mu \cdot Q(k) - S(k)) \qquad (10)$$

for S(k), where α and μ are suitably chosen constants (α and $\mu < 1$) and

$$Q(k) = \left\langle (1-M(k,p)) \cdot (D(k,p))^2 \right\rangle = \frac{\sum_p (1-M(k,p)) \cdot (D(k,p))^2}{\sum_p (1-M(k,p))} \qquad (11)$$

The system noise S(k) is therefore estimated from the image sequence by averaging $(D(k,p))^2$ over all image points p where M(k,p)=0 and subsequently low-pass-filtering this spatial average in time with filter coefficient α. The constant μ in eq. (10) provides a suitable normalization of Q(k). The spatial and temporal averaging process is a suitable way to remove the random noise parts in $(D(k,p))^2$, whereas the restriction of the spatial averaging to the non-moving regions ensures that S(k) does not contain any measurement noise contributions from the moving objects.

Moving object detection

Once the moving objects have succesfully been eliminated from the reference (background) image sequence by use of eqs. (1), (4), (6), (7), (10) and (11), the objects can be detected by application of equ. (2). These equations define the procedure visualized in figure 1. The quality of the detection

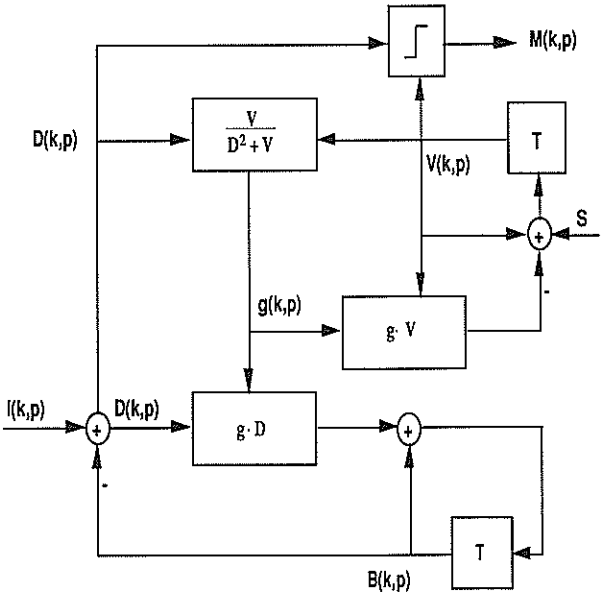


Fig. 1: Data frow chart for the improved algorithm

process depends seriously on a proper choice of the detection threshold in equ. (2). [Don88] uses Bayesian estimation theory to derive an expression for the threshold in terms of the camera noise variance V

$$\text{threshold} = 2 \cdot V \cdot \ln \frac{W}{\sqrt{2\pi \cdot V}} \qquad (12)$$

under the assumption of gaussian noise statistics. [Don88] and [Gun88] give recursive schemes to estimate the noise variance from the image sequence.

From the general results of Kalman filter theory we know that V(k,p) as calculated by means of equ. (7) represents a reasonable estimate of the noise (uncertainty) in B(k,p). Since V(k,p) contains contributions from the large measurement noise for p inside a moving region, V(k,p) usually overestimates the noise inside these regions. Outside the moving regions, V(k,p) tends to underestimate the camera noise and thus would lead to noisy pixels in the object masks M(k,p) if it were used to calculate the detection threshold. To get rid of this difficulty, a spatial average of V(k,p)

$$V = \frac{\sum_p V(k,p)}{\sum_p 1} \qquad (13)$$

over all image points should be used in equ. (12), whose value interpolates between the extremal values of V(k,p).

Discussion and simulation results

From the above arguments we conclude that the proposed algorithm should be able to overcome the deficiencies (wrong or ambiguous object boundaries, doubling of objects, long or even unstable start-up phase, instabilities due to self-enhancing detection errors) of earlier approaches.

Computer simulations using urban traffic scenes confirm our theoretical expectations. The shortcomings and deadlock situations associated with the former approaches were not observed at all using the algorithm proposed in this work. At the same time, the elimination of the moving objects from the reference image sequence is superior compared to the two procedures presented in [Gun88] and [Kar89]. Figure 2 display a typical difference image obtained from the original and the background image in the present algorithm from which the shapes of the moving objects can be easily determined.

All simulations employed the detection scheme proposed above, where the threshold is controlled by a spatially averaged variance. The decay times of the mean squared difference in the start-up phase are equal to the decay times of perturbations and detection errors in all three algorithms. The results clearly indicate that the procedure presented in this work is a robust, simple and efficient algorithm for moving object segmentation.

References

[Don88]
G.W. Donohoe, D.R. Hush and N. Ahmed, "Change Detection for Target Detection and Classification in Video Sequences", Proceedings of the ICASSP 88, 1084-1087.

[Gun88]
A. Gunzinger, S. Mathis, W. Guggenbühl, "Datenflußrechner zur Echtzeitbildverarbeitung", Proc. 10th DAGM-Symposium, Zürich 1988, 76-82.

[Hoe88]
M. Hötter, R. Thoma, "Image Segmentation Based on Object Oriented Mapping Parameter Estimation", Signal Processing 15 (1988) 315-344.

[Kar89]
K.P Karmann, A. v. Brandt, "Detection and Tracking of Moving Objects by Adaptive Background Extraction", Proceedings of the 6th SCIA, Oulu, Finland, June 1989, 1051-1058.

[Lew86]
F. L. Lewis, "Optimal Estimation", Wiley & Sons, New York 1986.

[Wik87]
J. Wiklund, G. H. Granlund, "Image Sequence Analysis for Object Tracking", Proceedings of the 5th Scand. Conf. on Image Analysis, Stockholm, June 1987, pp. 641-648.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Change Detection with Moment Invariants under Time–Varying Illumination Case

Chang–Wu Fu and Shyang Chang

Dept. of Electrical Engineering, National Tsing Hua University
Hsin Chu, Taiwan, Rep. of China

In this paper, an illumination independent change detection method is proposed. Based on the defined moments, we can distinguish the changes caused by motion from the ones by illumination. Furthermore, the amount of calculation of the moments can be reduced by the introduction of the j–th order circular shift moments. Hence, the time required for our method is no more than the shading model method in the worst case. Examples are given to illustrate its performance.

1. Introduction

In time–varying imagery, change detection is fundamental to many machine vision applications. For instance, an accurate and robust change detection algorithm is necessary for tracking systems of moving objects [1]–[2], the segmentation of dynamic scenes [3] and traffic flow analysis [4]. The advantage of detecting changes before any further processing is the reduction of the amount of raw image data to be processed. Because of its importance in the preprocessing of dynamic scene analysis, many researches have been devoted to the development of more robust algorithms [5]–[6].

Recently, Skifstad and Jain [7] proposed a shading model (SM) method to detect change in time–varying illumination case. According to Phong's shading model [8], the SM method uses the ratio of the intensities recorded in the corresponding AOIs of the two frames in a image sequence to detect changes. However, the amount of computation of the shading model method is huge because of its number of divisions. It is therefore very time–consuming to implement it.

In this paper, an illumination independent change detection method is proposed. Based on the property that the defined j–th order circular shift moments will be preserved in stationary regions in the consecutive frames when the illumination varies over time, we can detect the changes caused by motion. The remainder of this paper is organized as follows. In section II, we first define the j–th order circular shift moment (CS_j) functions and derive the moment–invariants based change detection algorithm under time–varying illumination. In section III, some examples are given to illustrate the performance of the proposed methods. Finally, the conclusions are given in the last section.

2. Change Detection with Moment Invariants

It is well–known that the brightness of the scenes can be written as the product of the illumination and the reflectance of the object surface for the reason of mathematical tractability [9]. Since the mapping factor between the brightness and the intensity recorded in the corresponding imaging sensor is fixed and irrelevant to our work, we may assume it to be 1. Hence the intensity recorded at the position (x,y) on the k–th frame in the image sequence can be expressed as

$$f(x,y,k) = i(x,y,k) \cdot r(x,y,k), \qquad (1)$$

where $f(x,y,k)$ is the intensity function, $i(x,y,k)$ the illumination one and $r(x,y,k)$ the reflectance one which depends on the physical surface itself and is independent of the illumination. Since the reflectance function $r(x,y,k)$ is independent of the illumination, it is straightforward to detect change by comparing the values of $r(x,y,k)$ and $r(x,y,k-1)$. Unfortunately, it is difficult to estimate the reflectance function without any additional information about the scenes. In order to avoid the mentioned problem, the idea is to define some characteristic function of intensity in the AOI and use it to detect change. The defined function should possess the property that it keeps the same value if no change of the scenes occurs in the corresponding AOIs between two frames when the illumination is time–varying.

Definition 1:
Let

$$F_i(k) = \sum_{(x,y)\in A_i} f(x,y,k),$$

the x– and y–direction moments $m^0_{i,x}(\cdot)$ and $m^0_{i,y}(\cdot)$ of some area of interest (AOI) A_i are defined as:

$$m^0_{i,x}(k) = \sum_{(x,y)\in A_i} f(x,y,k)\cdot x/F_i(k) \qquad (2a)$$

and

$$m^0_{i,y}(k)= \sum_{(x,y)\in A_i} f(x,y,k)\cdot y/F_i(k). \qquad (2b)$$

□

Definition 2:
The variation factor $\alpha(\cdot)$ of the illumination at position (x,y) in the k– and (k–1)–frame is defined as:

$$\alpha(x,y,k) = i(x,y,k)/i(x,y,k-1). \qquad (3)$$

□

If the AOI A_i is small enough, the values of the variation factor $\alpha(\cdot)$ of the illumination are approximately the same in A_i. It is reasonable to assume that $\alpha(x,y,k)=\alpha(z,w,k)=\alpha_k$, for (x,y), (z,w) $\in A_i$ when A_i is small enough. Therefore, (2a) becomes

$$m^0_{i,x}(k)$$

$$= \sum_{(x,y)\in A_i} i(x,y,k)\cdot r(x,y,k)\cdot x/F_i(k)$$

$$= \sum_{(x,y)\in A_i} i(x,y,k-1)\cdot r(x,y,k)\cdot x/F_i'(k),$$

where

$$F_i'(k)= \sum_{x,y)\in A_i} i(x,y,k-1)\cdot r(x,y,k).$$

Similarly, (2b) becomes

$$m^0_{i,y}(k)= \sum_{x,y\in A_i} i(x,y,k-1)\cdot r(x,y,k)\cdot y/F_i'(k). \qquad (4b)$$

It is obvious that the defined moment function suits the purpose of detecting change under time–varying illumination. This can be seen from that whenever there are no scenes changed in A_i between two frames, i.e., $r(x,y,k)=r(x,y,k-1)$ for $(x,y) \in A_i$, $m^0_{i,x}(k)=m^0_{i,x}(k-1)$ and $m^0_{i,y}(k)=m^0_{i,y}(k-1)$. However, it is not sufficient to detect change with only the defined moment functions. As a matter of fact, there are ambiguities in some cases. Figure 1 depicts one of the examples. In the constant illumination case, although the physical scenes in Figure 1(a) and 1(b) are substantially different, the values of the corresponding moment functions are the same. We therefore make a wrong decision that there is no change in the AOI between frames. To deal with this problem, some other characteristic functions must be defined in order to assist in detecting changes. The most intuitive way may be to define the higher order moment functions as in [10]. By the same reasoning, the higher order moments are also invariant when the illumination varies over time. Therefore, the illumination independent change detection can be performed by checking these moments between frames. However, the amount of computation to calculate them is huge. In fact, it is desirable to detect change as fast as possible. To reduce the amount of computation as well as assist in detecting change, the j–th order circular shift (CS_j) moment which is also invariant in the time–varying illumination case is introduced. Before we put forward the definition of the

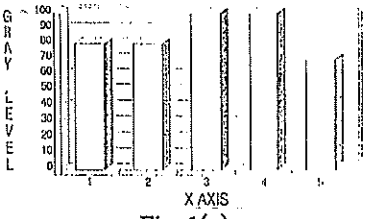


Fig. 1(a)

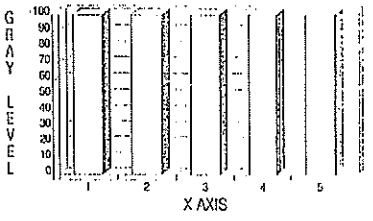


Fig. 1(b)

Figure 1. Example of two different physical scenes in the AOIs of size 1×5–pixel in the consecutive frames. Both of them possess moments of the same value 3.

CS_j moment function, we partition the image frame into N×N–pixel, square blocks for our use. Let A_{uv} denote the (u,v)–th block where $(x,y) \in A_{uv}$ if $u\cdot N \le x < (u+1)\cdot N$ and $v\cdot N \le y < (v+1)\cdot N$.

Definition 3:
The j–th order, x– and y–direction circular shift (CS_j) moment functions in A_{uv} are defined as:

$$m^j_{uv,x}(k)= \sum_{x,y=0}^{N-1} f(x',y',k)\cdot (x-j)_{\text{mod } N}/F_k \qquad (5a)$$

and

$$m^j_{uv,y}(k)= \sum_{x,y=0}^{N-1} f(x',y',k)\cdot (y-j)_{\text{mod } N}/F_k, \qquad (5b)$$

where $(y)_{\text{mod } N}=y-v\cdot N$ with n chosen such that $0 \le (y)_{\text{mod } N} < N$, $x'=x+u\cdot N$ and $y'=y+v\cdot N$.

□

With such defined moment functions, the reduction of the number of computations comes from the interrelation between the moment function (2) and the CS_j's.

Proposition:
Let

$$f_x(x',k) = \sum_{y=0}^{N-1} f(x',y',k),$$

moment functions and CS_j's have the following properties:

$$m^j_{uv,x}(k)=m^0_{uv,x}(k)-j+\sum_{x=0}^{j-1} f_x(x',k)\cdot N/F_k. \qquad (6)$$

$$m^{j}_{uv,x}(k)=m^{i}_{uv,x}(k)+(i-j)+\sum_{x=i}^{j-1} f_x(x',k)\cdot N/F_k. \tag{7}$$

Proof:
Multiply both sides of (5a) with F_k, we can get

$$F_k\cdot m^{j}_{A_{(m.n),x}}(k)$$

$$=\sum_{x=0}^{N-1} f_x(x',k)\cdot(x-j)_{\bmod N}$$

$$=\sum_{x=0}^{j-1} f_x(x',k)\cdot(N+x-j)+\sum_{x=j}^{N-1} f_x(x',k)\cdot(x-j)$$

$$=\sum_{x=0}^{j-1} f_x(x',k)\cdot x+(N-j)\cdot\sum_{x=0}^{j-1} f_x(x',k)$$

$$+\sum_{x=j}^{N-1} f_x(x',k)\cdot x-j\cdot\sum_{x=j}^{N-1} f_x(x',k)\cdot$$

$$=\sum_{x=0}^{N-1} f_x(x',k)\cdot x-j\cdot\sum_{x=0}^{N-1} f_x(x',k)+N\cdot\sum_{x=0}^{j-1} f_x(x',k).$$

Dividing both sides of the above equation with F_k, equation (6) is proved. As for the proof of equation (7), it follows from (6) directly.

□

Similarly, the properties (6) and (7) also hold for the y–direction moment function and CS_j's. With (2), (4) and (7), the moment invariants (MI) based change detection algorithm can be stated as follows.

MI based Algorithm:
In the AOI A_j,
Step 1. Pick the thresholds, $\delta_{1,x}$, and employ (2) to calculate the values of CS_0 moments associated with the AOI in the (k–1)– and k–th image frames.
If $|\ m^{0}_{j,x}(k)-m^{0}_{j,x}(k-1)\ |\le\delta_{1,x}$, go to step 2;
else, it is claimed that changes occur in A_j.
Step i. (i=2,3$\cdots$N) Choose the threshold, $\delta_{i,x}$, and calculate the corresponding values of CS_{i-1} moments by (7).
If $|\ m^{i-1}_{A_j,x}(k)-m^{i-1}_{A_j,x}(k-1)\ |\le\delta_{i,x}$, go to next step;
else, it is claimed that changes occur in A_j.

□

It is obvious that the accuracy of the proposed algorithm increases as the step of process progresses. That is the accuracy depends on the time the process of change detection takes. From this point of view, it is the same as the progressive image transmission technique which transmits image data with resolution varies from coarse to fine. Since the time divided for the preprocessing is usually short, such a progressive change detection algorithm can check the full image frame once during a limited time interval. As for the SM method, it takes more time than ours and may not be able to check the full image frame once during the limited time interval. Therefore the MI based algorithm is more suitable than the SM method for applications when the time divided for the preprocessing is limited.

3. Examples

In this section, we will utilize the proposed algorithm to detect change between the consecutive frames in an image sequence. In our experiments, the images used are 255×255 pixels and are partitioned into 5×5–pixel, square blocks. The results of applying the proposed algorithm to the test image pair will be compared with the corresponding ones to which the SM method is applied. Figure 2 shows the test image pair taken in a laboratory setting containing one block standing on a table with a new object (the toy car) added to the scene in the second frame. As can be seen, there is a significant change in the illumination on the scene. We implement the SM method and our proposed one in C language and the programs are run on PC–AT. Figure 3 shows the results of applying the SM method to the first test pair in Figure 2. And the results of the proposed algorithm being terminated at step 1 and 5, are depicted in Figure 4(a) and 4(b), respectively. As can be seen, both the SM method and the proposed one can isolate the toy car accurately from the background. Very little of the background is estimated as having changed. Note that the time required for the proposed method is less than the SM method. The SM method takes 22.46 seconds; as for the proposed one, it takes 15.65 and 20.36 seconds, respectively.

4. Conclusions

A moment invariant based change detection algorithm has been presented in this paper. Based on the defined j–th order circular shift moments, the proposed algorithm can detect changes when the illumination is time–varying. Furthermore, the amount of computation has been reduced and the time required has been shown less than the SM method. Experimental results have been given to illustrate the benefits of the proposed method not at the cost of degradation of the performance in comparison with that of SM method.

Reference:

[1] Fu, C.W. and Chang, S., Pattern Recog. Letters, (1989) 195.
[2] Lai, S.H. and Chang, S., Pattern Recog. Letters, (1988) 341.
[3] Jayaramamurthy, S.N. and Jain, R., Comput. Vision, Graphic, Image Processing., (1983) 239.
[4] Nagle, H.H., ICASSP, (1982) 1179.
[5] Jain, R., Militzer, D. and Nagle, H.H., IJCAI, (1977) 612.
[6] Hsu, Y.Z., Nagle, H.H., and Refers, G., Comput.

Fig. 2

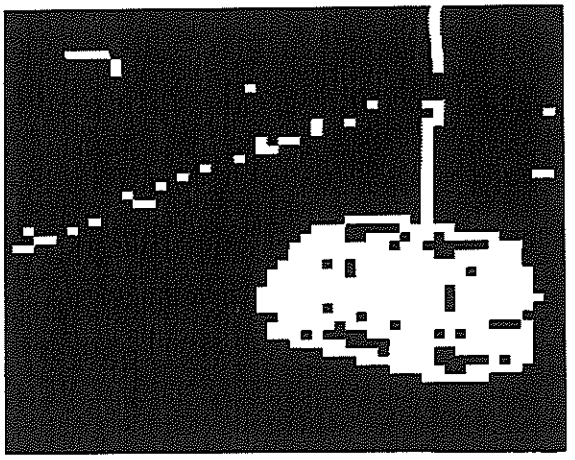


Fig. 3

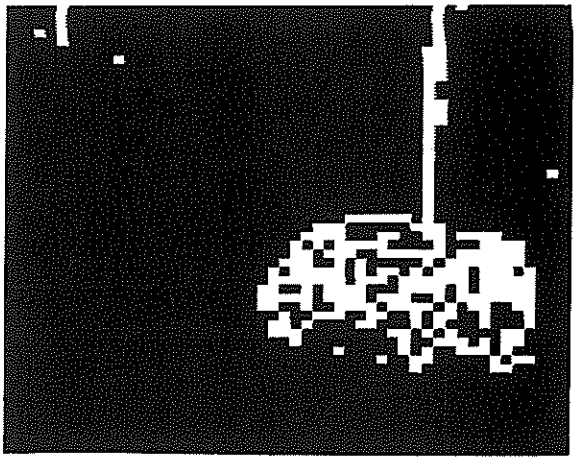


Fig. 4(a)

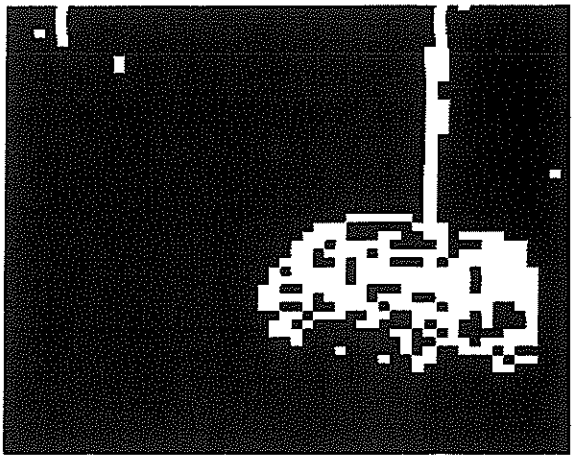


Fig. 4(b)

Vision, Graphic, Image Processing., (1984) 73.
[7] Skifstad, K. and Jain, R., Comput. Vision, Graphic, Image Processing., (1989) 387.
[8] Phong, B.T., Commun. ACM, 311.
[9] Rosenfeld, A. and Kak, A.C., Digital Picture Processing (U.S.A., New York, 1982).
[10] Mingfa, Z., Hasani, S., Bhattarai, S. and Singh, H., Pattern Recog. Letters, (1989) 175.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

RECURSIVE MOTION ESTIMATION BASED ON A MODEL OF THE CAMERA DYNAMICS

Achim v. Brandt, Klaus-Peter Karmann, Stefan Lanser
Siemens AG, Corporate Research and Development
Information and Knowledge Processing, Image Processing
Otto-Hahn-Ring 6, D-8000 München 83, F. R. Germany
e-mail: brandt@ztivax.uucp

The problem of estimation and suppression of global movements in image sequences is addressed. A procedure is presented that calculates a stabilized version of the given image sequence and at the same time determines the position and shape of moving objects. The global movements are assumed to be caused by camera jitter where the rotation and translation parameters are generated by a multichannel AR process representing the mechanical properties of the camera system. Based on this model of the system dynamics local image matching, Kalman filtering and AR parameter estimation are combined for obtaining an efficient time recursive motion estimation scheme.

1. INTRODUCTION

In image sequences taken from a "stationary" camera for the purpose of moving object detection, the image is often subject to *oscillatory global movements* (mainly translation and rotation) due to *camera vibrations*. An example is *traffic monitoring* with the camera being mounted on a high pole which is subject to wind forces. These movements may give rise to false alarms in the object detection process. So they should be duly recognized and compensated.

Motion compensation is a well-known technique in image sequence coding where its main application is motion compensated prediction of an image frame by its predecessor (cf. [1, 2]). This is accomplished by calculating a motion vector field (or optical flow field) from every two successive frames (I_k and I_{k+1}, say) consisting of a displacement vector for every picture element. By shifting all the pixels of I_k according to the motion vector field, an approximation to I_{k+1} is obtained.

Here the situation is slightly different. Given an image sequence obtained by a "stationary" camera which is, however, subject to oscillations resulting in some jitter in the images, we like to turn every image into a normalized image in "zero position" by suitable motion compensation.

So for every frame I_k we need a motion vector field V_k indicating the transformation between I_k and some position-normalized (non-moving) predicted image $B_{k|k-1}$. V_k should contain the *global translations and rotations* caused by camera motion without being confused by any moving object occurring in the scene.

For every I_k a corresponding V_k could in principle be obtained by any motion estimation procedure that calculates global motion parameters from two given frames (e. g. [2] or [3]). However, the computational load can be reduced while increasing the performance by taking the temporal correlation of the motion parameters into account. Every motion vector field V_k (or its characterizing parameters, respectively) can be predicted from the past by suitable *modelling of the mechanical properties* of the camera and its support, i. e. the camera dynamics. After an initialization phase, every V_k can be predicted from the previous ones and the present frame I_k is just used for *correcting* the predicted global motion parameters rather than for motion estimation from scratch.

This results in *less noisy estimates* of the global translation and rotation parameters. Moreover the number of features in I_k and $B_{k|k-1}$ that need to be matched for updating the predicted motion

vector field is much less than for non-predictive motion estimation. The procedure is explained in the next section followed by simulation results in section 3.

2. RECURSIVE MOTION ESTIMATION

2.1. System Model

The current global image transformation at time k can be derived from the current values of the six parameters ω_x, ω_y, ω_z, t_x, t_y, and t_z, which denote the camera rotation angles and the camera translation in three dimensions, respectively. So these are the desired parameters to be recursively estimated. At any time there are two sources of information for these parameters: (1) the measurable displacements between the current frame and the reference image, and (2) the predicted parameter values based on previous measurements.

Formally this can be expressed by a system model and a measurement model which are the basis for recursive state space estimation (*Kalman filter* [6]). Since the system model should reflect the typical oscillatory behaviour of the camera system, the state vector of the Kalman filter must not only include the six parameters to be estimated but also their temporal derivatives or, equivalently, their previous values. Therefore we define the state vector as

$$\mathbf{x}(k) = \left[\mathbf{x}_0^T(k),\ \mathbf{x}_0^T(k-1)\right]^T \tag{2.1}$$

where

$$\mathbf{x}_0(k) = \left[\omega_x(k),\omega_y(k),\omega_z(k),t_x(k),t_y(k),t_z(k)\right]^T \tag{2.2}$$

Based on this definition the *system model* can be stated:

$$\mathbf{x}(k+1) = \mathbf{A}_k\,\mathbf{x}(k) + \mathbf{u}_k \tag{2.3}$$

where $\mathbf{A}_k$ is the system matrix and $\mathbf{u}_k$ is a white noise process with covariance matrix $\mathbf{Q}_k$ [6]. This equation defines a multichannel autoregressive (AR) process. For ease of computation this multichannel process is split into separate scalar AR processes, one for every element of $\mathbf{x}_0(k)$. So $\mathbf{A}_k$ and $\mathbf{Q}_k$ are structured in such a way that eq. 2.3 is equivalent to a set of six equations of the form

$$\omega_x(k+1) = a_1(k)\omega_x(k) + a_2(k)\omega_x(k-1) + u_k \tag{2.4}$$

These second order AR processes are able to generate oscillatory parameter variations provided the AR coefficients $a_i(k)$ are suitably adjusted (see below).

2.2. Measurement Model

The coefficients of the state vector $\mathbf{x}(k)$ must be related to some measurable quantities. In our case these are the displacement vectors $\mathbf{v} = [u,v]^T$ measured at some image coordinates $\mathbf{p} = [x,y]^T$. They are preferably obtained by block matching at predefined coordinates or by feature tracking. This is done in the "local motion estimation" unit in fig. 1. For small camera rotations and translations the displacement vectors depend on the motion parameters in $\mathbf{x}_0(k)$ by [5]:

$$u = \frac{-t_x + x t_z}{Z(x,y)} + \omega_x xy - \omega_y(x^2+1) + \omega_z y \tag{2.5}$$

$$v = \frac{-t_y + y t_z}{Z(x,y)} - \omega_y xy + \omega_x(y^2+1) - \omega_z x \tag{2.6}$$

where $Z(x,y)$ is the depth at $[x,y]^T$, i. e. the distance of the object from the camera, and the focal length has been normalized. This can be written as the linear superposition

$$\mathbf{v}_k(x,y) = \begin{bmatrix} u_k(x,y) \\ v_k(x,y) \end{bmatrix} = \sum_{i=1}^{6} w_i(k)\,\mathbf{w}_i(x,y) \tag{2.7}$$

where k is the time index; w_1 to w_6 are the coefficients ω_x, ω_y, ω_z, t_x, t_y, and t_z, and $\mathbf{w}_i(x,y)$ are elements of the basic motion vector fields:

$$\mathbf{w}_1(x,y) = [xy,\ y^2+1]^T \tag{2.8.1}$$

$$\mathbf{w}_2(x,y) = [-(x^2+1),\ -xy]^T \tag{2.8.2}$$

$$\mathbf{w}_3(x,y) = [y,\ -x]^T \tag{2.8.3}$$

$$\mathbf{w}_4(x,y) = \left[-\frac{1}{Z(x,y)},\ 0\right]^T \tag{2.8.4}$$

$$\mathbf{w}_5(x,y) = \left[0,\ -\frac{1}{Z(x,y)}\right]^T \tag{2.8.5}$$

$$\mathbf{w}_6(x,y) = \left[\frac{x}{Z(x,y)},\ \frac{y}{Z(x,y)}\right]^T \tag{2.8.6}$$

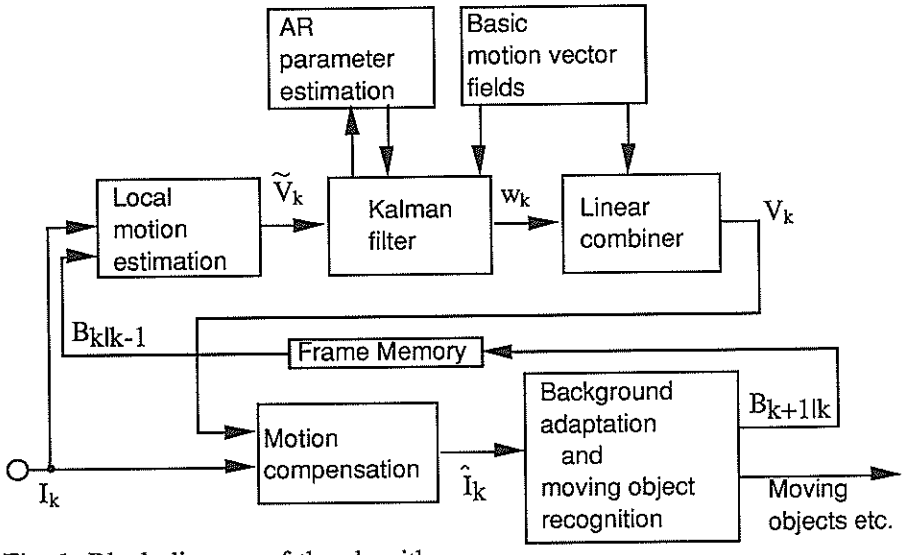


Fig. 1: Block diagram of the algorithm.

Obviously the motion vector fields which are related to camera rotation (eqs. 2.8.1-2.8.3) are independent of the depth value Z(x,y). So these can be applied without any additional information. However, for the determination of the remaining motion vector fields which are related to camera translation (eqs. 2.8.4-2.8.6) a terrain model is required for the derivation of Z(x,y). With a fixed camera position and a static background, this terrain model is known and can be used as apriori knowledge for the motion estimation procedure. Otherwise we assume that the depth is inversely proportional to the vertical pixel coordinate, i. e. we use $Z(x,y) = 1/y$, resulting in the following depth-independent equations for $w_4(x,y)$, $w_5(x,y)$ and $w_6(x,y)$:

$$w'_4(x,y) = [-y,\ 0]^T \tag{2.9.1}$$

$$w'_5(x,y) = [0,\ -y]^T \tag{2.9.2}$$

$$w'_6(x,y) = [xy,\ y^2]^T \tag{2.9.3}$$

These basic motion vector fields are applied for definition of the *measurement model* . Let $\{p_n = [x_n,y_n]^T,\ n = 1...N\}$ be the set of pixel coordinates for which displacement vectors $v_n = [u_n, v_n]^T$ have been measured. Hence

$$z(k) = [v_1^T, v_2^T, ..., v_N^T]^T \tag{2.10}$$

is the measurement vector to be used in the measurement model

$$z(k) = H_k\, x(k) + r_k \tag{2.11}$$

where H_k is the measurement matrix consisting of the entries

$$h_{n,i} = w_i(p_n) \tag{2.12}$$

for n = 1...N and i = 1...6, and $h_{n,i} = 0$ for i>6 as can be derived from eq. (2.7). Again, r_k is a white noise process with covariance matrix R_k. The entries of R_k should reflect the variances or uncertainties of the measured displacement vectors v_n.

2.3. Kalman Filter

The system and measurement models defined above (eqs. (2.3) and (2.11)) can be uniquely transformed into the well-known recursive state estimation equations (cf. [6], p. 69). For calculation of the Kalman gain the inversion of a matrix with dimension six is required.

2.4 AR Parameter Estimation

The AR coefficients $a_1(k)$ and $a_2(k)$ (eq. (2.4)) must be adapted to the actual camera dynamics. This is accomplished by calculating the optimal predictor for the entries of the state vector $x_o(k)$. Using the estimated entries of $x_o(k)$ as the given input signal a large number of recursive AR parameter estimation scheme is available [7]. In most cases LMS estimation will be sufficient.

2.5 Motion compensation

From the estimated global motion parameters a complete displacement vector field V_k can be obtained using eq. (2.7). This is used for transforming the given image I_k into a position normalized image $\hat{I}_k$ similar to motion compensated prediction [2].

2.6 Background / Object Segmentation

Since the global motion parameters can only be determined from the motion of the "stationary" background, a procedure for background/object segmentation [4] is applied. The result of this procedure is a segmented image as well as a predicted background image $B_{k+1|k}$ for use in the next step of the motion estimation procedure.

3. SIMULATION RESULTS AND DISCUSSION

Fig. 2 shows a scene from traffic monitoring where the image has been rotated artificially corresponding to a camera rotation around the z-axis. The global motion vector field that has been obtained by the recursive state estimation procedure based on local block matching (block size 8*8 pixels) has been overlaid.

Fig. 3 displays the same image after motion compensation and another local motion estimation operation followed by moving object segmentation. Due to the compensation process the shape of moving objects is clearly visible in the optical flow field.

The results obtained so far have demonstrated the robustness and efficiency of the state space approach for motion estimation. The combined use of a hierarchy of estimators where the state variables of one estimator are the input signal of the following will be studied further.

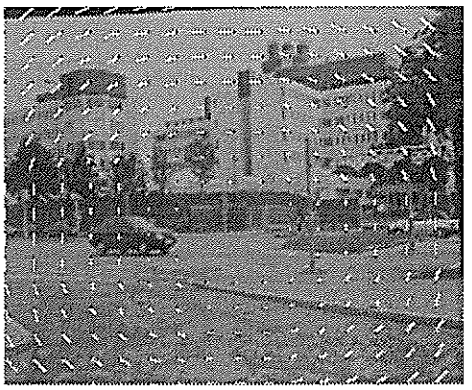


Fig. 2: Estimated global motion

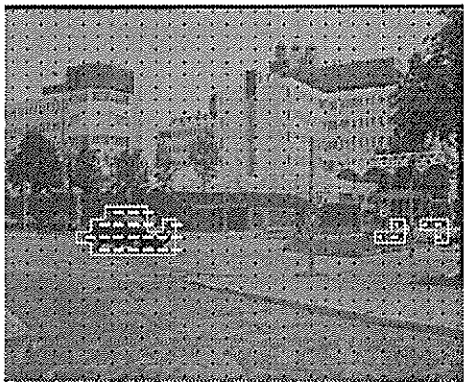


Fig. 3: Estimated local motion after global motion compensation

REFERENCES

[1] A. v. Brandt, "Motion estimation and subband coding using Quadrature Mirror Filters", Proc. EUSIPCO-86, vol. 2, pp. 829-832, 1986

[2] M. Hoetter, "Differential estimation of the global motion parameters zoom and pan", Signal Processing, vol. 16, no. 3, pp. 249-265, 1989

[3] G. Keesman, "Motion estimation based on a motion model incorporating translation, rotation and zoom", Proc. EUSIPCO-88, vol. 1, pp. 31-34, 1988

[4] K.-P. Karmann, A. v. Brandt, R. Gerl, "Moving object segmentation based on adaptive reference images", this volume

[5] J. Heel, "Dynamical systems and motion vision", AI Memo 1037, MIT Artificial Intelligence Laboratory, April 1988

[6] L. Lewis, Optimal Estimation. Wiley 1986

[7] P. Strobach, Linear Prediction Theory. Springer 1990

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

REAL TIME TOKEN TRACKER

S. DE PAOLI (LTIRF - ITMI)
A. CHEHIKIAN (LTIRF)
P. STELMASZYK (ITMI)

LTIRF
INPG
46 av Felix VIALLET
38031 GRENOBLE-CEDEX, France

ITMI
Chemin des Prés
38240 MEYLAN, France

The use of depth from motion for 3-D scene analysis requires a system which can accurately and reliably measure image motion. Such measurements may be obtained by tracking the position of edge lines in a closely spaced monocular sequence of images. The reliability of such a technique is greatly enhanced by maintaining a model of image flow composed of the position and velocity of tokens constructed from edge lines.
In this paper we present some algorithmic aspects of such a system and show how the complexity can be reduced to optimize hardware aspects. We also present a hardware architecture and discuss some results based on real data.

1. INTRODUCTION

The information of motion does not exist in a static image but can be calculated by tracking the positions of tokens in a sequence of images. The Token Tracker is to be integrated into a real time vision-machine which contains a token extraction process. The token extraction process is performed by an edge detector, a chaining operator and a polygonal approximation operator. Extensive descriptions of these operators are given in [DER-87],[DIS-90].

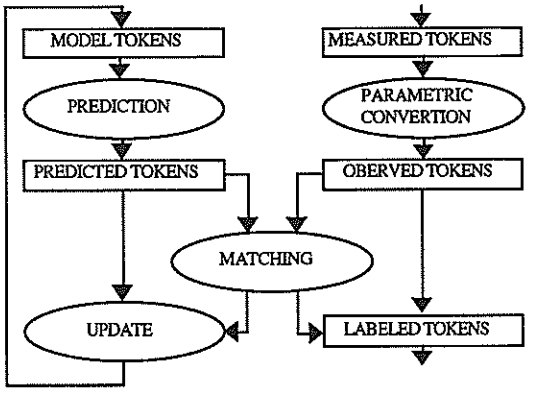


figure 1: the "flow model" process

To have reliable results, the Token Tracker has to cope with different sources of noise. These noises are issued from the vibrations of the mechanical means which support the camera if the camera is moving, from the instabilities of the token extraction process, from the occlusions and from the photometric effects. An actively updated "flow model" based on the principle of the Kalman filter provides a technique for minimizing the effects of these degradations although the noises are not gaussian. An overview of such a process that matches tokens and maintains the flow model is shown in figure 1.

The flowchart works as follows: Newly observed tokens arrive continuously from the token extraction process and the parametric representation is computed. From model tokens updated at time t the new attributes are predicted at time $t+\delta t$, using Kalman equations. The observed tokens are then matched with the predicted ones, and receive a label and a confidence factor before being restituted. Finally, model tokens are updated at time $t+\delta t$ using the predicted tokens and the matching results. The update is performed by: i) model token estimating from predicted and matched tokens, ii) adding newly observed tokens, iii) deleting unmatched model tokens, [CRO-88].

2. PARAMETRIC REPRESENTATION

Although the coordinates of the end points are sufficient to describe a token, these parameters are not well suited for efficient matching as well as for Kalman filter implementation. The reason for this is that these parameters are closely dependent and highly affected by any change in the length of the token. We have chosen the set (c,θ,l,x,y) to represent line segments. In this set (refer to figure 2): c is the perpendicular distance from the origin to the line

This work is performed as a part of ESPRIT project P940.

supporting the token, θ is the orientation of the token, l is the length of the token and (x,y) are the cartesian coordinates of the middle point. A more complete description of this parametric representation is given in [STE-88].

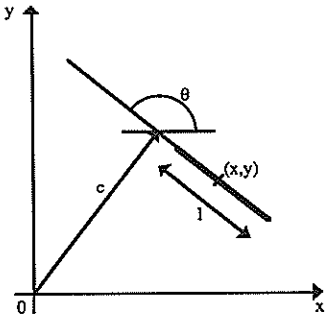


figure 2: the parametric representation

3. A HARDWARE IMPLEMENTATION FOR THE KALMAN FILTER

With respect to the prediction and to the model update phase, the requirement of fast and simple hardware imposes the need to reduce the complexity of Kalman filter's equations. This is possible by making the following assumptions:

a) We assume that an observed token can be expressed as a vector:

$$T^t = \{c_t, \theta_t, l_t, x_t, y_t\}$$

If we assume that the parameters (c,θ,l,x,y) are not correlated, each parameter can be processed independently, and the flow model of a token can be expressed as a set of flow models, each one representative of only one parameter. We can then save time by implementing 5 independent Kalman filters, one for each parameter. In fact, the assumption of parameter independence is not true (x and y are not independent). Considering the parameters as independent is equivalent to making a model error which commonly leads the filter to diverge, but this problem can be avoided by increasing the estimated state covariance matrix.

b) We assume that the transition matrix of the Kalman filter is not time dependent, i.e. we consider a stationary model updated at constant time intervals. Thus, if a model token cannot be matched with an observed token, we update the model using the predicted value.

c) We do not estimate acceleration, i.e. each parameter λ of a model token is expressed as a set (X,V) in which X represents the position and V the velocity. Making such an assumption is equivalent to making a model error, but we have seen that this problem can be overcome.

A complete description of these equation is given in [CHE-89].

4. THE MATCHING PROCESS

For any parameter of a model token, the Kalman filter defines a search area in which its corresponding observed token must be found. For each model token, the matching process compares it with all the observed tokens and the most similar observed token is then selected as a match. If N_m is the number of model tokens and N_o is the number of observed tokens then the number of comparisons needed for the matching process is $N_m \times N_o$. As a consequence, despite the fact that the matching process involves simpler computations than Kalman filtering, care must be taken in terms of computation time.

The first step in the matching process consists of tests on parameters c,θ,l according to:

$$Z - \tilde{X} \in [-\delta\tilde{X}, \delta\tilde{X}] \qquad (1)$$

$\delta\tilde{X}$ is the width of the search area. The test on (c,θ) permits one to determine the straight line on which the token belongs, while the test on l determines if the length of the matching candidate is correct.

The x and y parameters are used to check overlap between the predicted and observed token:

$$|x-\tilde{x}| + |y-\tilde{y}| < \frac{\tilde{l}+l}{2} \qquad (2)$$

In the case where there still remain several matches for a model token, the best match is computed in a second step, using a cost function based on a form of the Mahalanobis distance between the model token and the observed token:

$$BM = \frac{\tilde{l}-l}{\sigma_l} + \frac{\tilde{\theta}-\theta}{\sigma_\theta} \qquad (3)$$

The best match is simply the one having the minimum BM value. In fact, equations (2) and (3) need some complex computations, so they are only applied to those observed tokens which lie in the predicted search area of a given model token.

5. SPEEDING UP THE MATCHING PROCESS

With 250 observed tokens and a model memory containing up to 300 tokens, about 75000 matching computations have to be performed every 0.1 of a second. If we consider that N_i is the number of necessary instructions for one computation, the processor board which supports the algorithm should have the capacity to process up to $75000*10*N_i$ instructions per second, only for the matching process.

We can speed up the matching process by dividing observed tokens into several subsets using a convenient criterion, and matching only the observed tokens belonging to the selected subset. The orientation is well suited for dividing observed tokens because most of the time the histogram of θ covers completely the space of $[0,2\Pi]$. The following figure shows the histogram of θ in the sequence of 89 images shown in figure 6.

Thus, we have chosen to divide the observed tokens into four subsets corresponding to:
$[0,\Pi/2]$, $[\Pi/2,\Pi]$, $[\Pi,3\Pi/2]$, $[3\Pi/2,2\Pi]$,
and four overlapping subsets corresponding to:
$[7\Pi/4,\Pi/4]$, $[\Pi/4,3\Pi/4]$, $[3\Pi/4,5\Pi/4]$, $[5\Pi/4,7\Pi/4]$.

It is necessary to create the overlapping subsets which contain the observed tokens, located near the borders of a main area. Notice that tokens which are found near the border of a main subset are in the center of the corresponding overlap subset. Then as the matching is performed in a single subset, the matching area of a predicted token falls in a single subset.

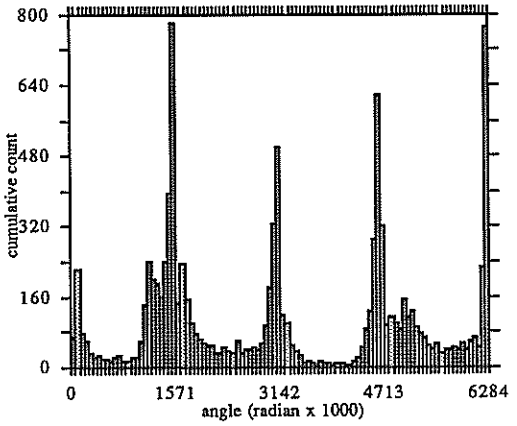


figure 3: histogram of θ

6. THE HARDWARE IMPLEMENTATION

We have designed a single-processor board with an original memory organisation. Such a design has several advantages over a multi-processor board since it is less complex and cheaper. The memory which contains the observed tokens is divided into different pages. The pages are the hardware representation of the subsets described in chapter 5 and are selected with a page register.

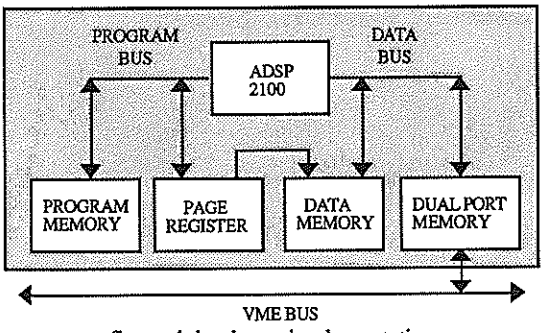


figure 4: hardware implementation

For hardware implementation we have to take into account the environment in which the token tracker works. In our case, it communicates with the other boards of the vision-machine via the VME bus. The adopted architecture is illustrated in figure 4. The processor used is the digital signal processor ADPS 2100 of Analog Devices. This DSP, because of its separate data and program buses is very convenient with regard to the imposed speed requirements.

The Token Tracker is a slave board and the dialogue with a host computer is performed via a dual-port memory. With such an interface, the VME bus only sees a memory and a status/control register.

7. SOME EXPERIMENTAL RESULTS

Figure 5 shows the computation times that have been obtained with a prototype of the presented Token Tracker. The times are measured between the time an image is taken by the Token Tracker and the time when the results are available. The number of token per image is about 170 and the sequence (shown in figure 6) contains 89 images.

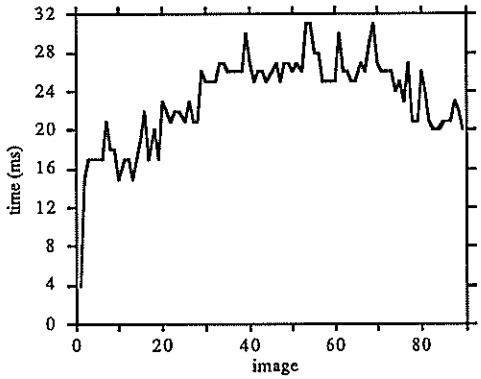


figure 5: computation times

One can notice that the maximum computation time is about 30 ms, with about 170 tokens. By extrapolation, we can estimate that the computation time for 250 tokens is lower than 100 ms, so the Token Tracker is able to work at a rate higher than 10 images per second.

Figure 6 shows the image sequence we are working with. In this figure, the sequence is shown in 10 images.

Figure 7 illustrates some of the motion information one can obtain with the Token Tracker. In this figure, the host computer selects some tokens observed for a certain amount of time. This figure shows all tokens seen more than 30 times and thus having a maximum confidence factor: figure 7.1 shows the evolution of these tokens in the sequence and figure 7.2 shows the motion of the middle point of these tokens and their label.

8. CONCLUSION

The Token Tracker is a simple process which provides an elegant and reliable solution to the problem of image flow measurement and image correspondence. The resulting flow model is solid, able to tolerate image disturbance due to noise. Token tracking is made possible by maintaining an explicit model of the token motion using Kalman filtering. Making some assumptions, a Kalman filter can be easily implemented with such a hardware and a suitable representation of the tokens allows a simple and fast matching. The process is implemented on a single processor

VME board and is able to track up to 250 tokens per image at the rate of 10 images per second.

BIBLIOGRAPHY

CHE-89
A.Chehikian, P.Stelmaszyk, S. de Paoli. Hardware evaluation for tracking edge-lines. Proceedings of the international workshop on industrial applications of machine intelligence and vision. pp 332-335, IEEE, Tokyo, April 89.

CRO-88
JL. Crowley, P.Stelmaszyk, C.Discours. Measuring image flow by tracking edge-lines. 2nd Int. Conf. on Computer vision. Tampa, Florida, Dec 88.

DER-87
R. Deriche. Optimal edge detection using recursive filtering. First International conference on computer vision. London (UK), June 87.

DIS-90
C.Discours. Analyse du mouvement par mise en correspondance d'indices visuels. These INPG Feb 90.

STE-88
P.Stelmaszyk, C. Discours, A. Chehikian. A fast and reliable token tracker. IAPR workshop on computer vision, Tokyo, Oct 88.

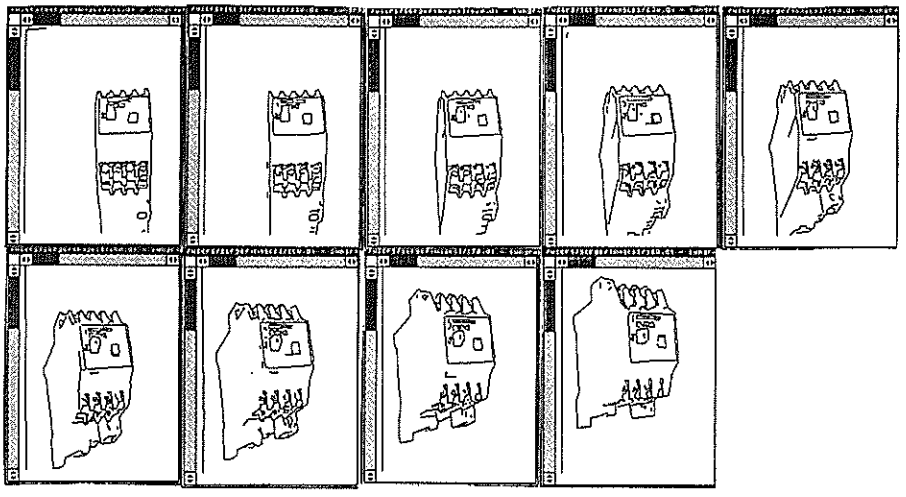


figure 6: The image sequence

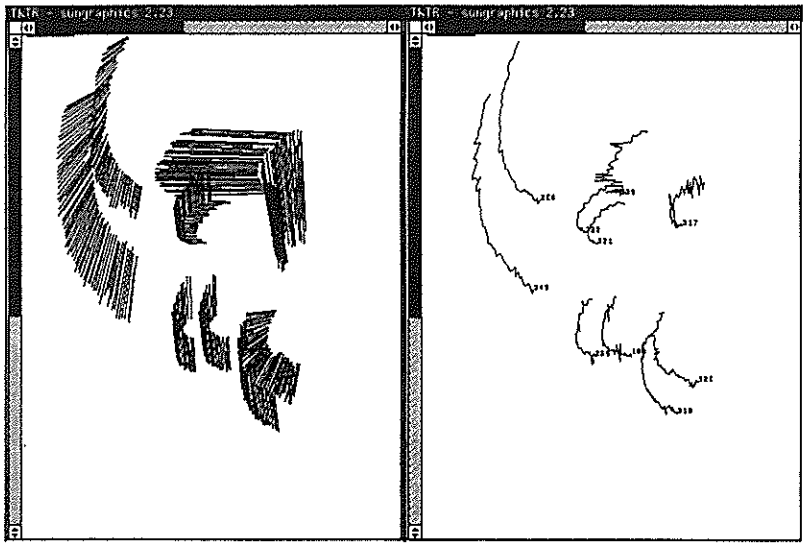


figure 7.1: max CF > 30 times (CF: Confidence Factor)

figure 7.2: motion

Figure 7: Tracking results

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

SMOOTHING THE DISPLACEMENT FIELD FOR EDGE-BASED MOTION ESTIMATION

Georges TZIRITAS

Laboratoire des Signaux et Systèmes, Ecole Supérieure d'Electricité, Plateau de Moulon
91192 Gif-sur-Yvette Cedex, FRANCE

In this article we present a method for estimating the two-dimensional velocity field on moving edges. We assume that the 2D velocity field is an affine transformation of the point coordinates. This model takes into account many different types of motion, and allows us to obtain a recursive relation on the 2D velocity field. We use this relation to determine the optimal smoother based on the measure of the normal component of the velocity vector. The well-known two-filter formula [1] is used.

1. INTRODUCTION

Motion estimation in a sequence of images is an important challenge in image processing and scene analysis. Two approaches can be considered: a 2D region-based and an edge-based. In this paper we are interested in edge-based estimation. Indeed, edges may constitute relevant features for motion robust estimation. The points on the edges usually correspond usually to orientation discontinuities between different surfaces or to object boundaries. It can be assumed that the observed 2D edges or boundaries are the geometrical projection of the 3D edges on the image plane. The estimated velocity field is not dense, but there are many reasons which allows us to consider that this velocity field is more significant than the region-based one. The principal reason is that in these points the apparent motion (or optical flow) can be considered as the projection on the image plane of the 3D velocity field of a moving scene. A.Verri and T.Poggio [8] have shown that the apparent and the real velocity field are very close where the image gradient is sufficiently strong. They conclude that to recover the 3D velocity field, edge-based algorithms seem more suitable than algorithms based on spatial and temporal derivatives of the image brightness. In the following we suppose that edges correspond to features in the scene.

The first operation of an edge-based motion estimator is the edge detection. In order to realize this operation we use J.Canny's method [2], as it is implemented by R.Deriche [3]. The result of this operation is the localization of the edge points, and the estimation of the orientation of the edge segment at each point.

In order to estimate the displacement vector, it is also necessary to determine, independently of the method used for the estimation, connections between edge points. In this paper we use a simple method to test connections and to link edge points. At the end of this operation a link of the points which constitute an entire edge is determined, the orientation of the edge at each point is obtained and the normal component of the displacement vector is estimated.

The result of the feature extraction processing is a set of lists of points belonging in different edge elements. An edge is described by a list of points $\mathbf{p}_k$ with 2D coordinates (x_k, y_k),

$$\{\mathbf{p}_k : k = 1, 2, \ldots, N\}$$

for an edge containing N points.

Concerning motion, only one component of the displacement vector can be measured from edge positions in successive images. This is the perpendicular to the edge

component. It is the well-known aperture problem. To measure the normal component a displacement on the perpendicular direction from the first contour to the second is considered. This measurement may introduce errors if the edge is not locally a straight line [7].

If $\mathbf{w_k} = [\ u_k\ \ v_k\ ]^T$ is the 2D velocity vector at point $\mathbf{p_k}$, and $\mathbf{n_k}$ the normal vector at the same point, with $\| \mathbf{n_k} \| = 1$, then the normal component of the velocity vector is given by $\mathbf{n_k^T w_k}$. If $w_k^{\perp}$ is the measured normal component, then we can write

$$w_k^{\perp} = \mathbf{n_k^T w_k} + z_k$$

where z_k is a random noise, here supposed to be zero-mean and white with variance equal to R_k.

To estimate the other component of the displacement vector some smoothness constraints must be used. E.Hildreth [4] proposed a regularization method which search for a compromise between the closeness on the data and the smoothness of the displacement vector. E.Hildreth [4] proposed to minimize the following criterion

$$\sum_{k=1}^{N-1} \| \mathbf{w_{k+1} - w_k} \|^2 + \alpha^2 \sum_{k=1}^{N} (w_k^{\perp} - \mathbf{n_k^T w_k})^2$$

and to solve for $\{\mathbf{w_k} : k = 1, 2, \ldots, N\}$ using the conjugate gradient algorithm. Here we propose to use an optimal smoother, also optimizing a quadratic criterion, and based on the same measures, but in another type of smoothing, which is presented in the following Section. The resulting smoother is presented in Section 3.

2. MODEL OF THE VELOCITY FIELD

In a precedent article [7] we considered some simple geometrical assumptions in order to obtain a model of the 2D velocity field in the case of a rigid 3D motion projected on the image plane. If this projection is orthographic, the 2D velocity fiels is an affine transformation of the point coordinates. G.Mailloux et al. [5] use the same model in a different domain of application concerning two-dimensional echocardiograms and heart motion. We present this model in the following, and we use it for obtaining neighbourhood relations on the velocity field.

The velocity vector $\mathbf{w}$ at a point $\mathbf{p}$ is modelized by

$$\mathbf{w} = \mathbf{t} + \mathbf{A}\mathbf{p} \qquad (2)$$

where $\mathbf{t}$ corresponds to a translation vector and matrix A takes into consideration rotation and some deformation of the pattern of the edge. The criterion (1) takes into account only a pure translation vector (A = 0). Let us suppose for simplicity, that for the definition of the velocity the time unit is equal to the temporal sampling period. Then, if $\mathbf{p'}$ and $\mathbf{p}$ are corresponding points, in two successive frames, we have

$$\mathbf{p'} = \mathbf{t} + (\mathbf{I}+\mathbf{A})\mathbf{p} \qquad (3)$$

We can assume for applicable models that eigenvalues of matrix are in modulus small in comparison with 1, and therefore matrix I+A is always supposed non-singular. Under these hypotheses, it is easy to demonstrate that a straight line is transformed by (3) into a straight line, and a polygon into a polygon. Indeed, if $(\mathbf{p'}, \mathbf{p})$, $(\mathbf{p_1'}, \mathbf{p_1})$ and $(\mathbf{p_2'}, \mathbf{p_2})$ are corresponding points and $(\mathbf{p}, \mathbf{p_1}, \mathbf{p_2})$ are aligned, we have

$$\det [\ \mathbf{p'-p_1'}\ \ \mathbf{p_1'-p_2'}] = \det(\mathbf{I}+\mathbf{A})\ \det [\ \mathbf{p-p_1}\ \ \mathbf{p_1-p_2}] = 0$$

which means that points $(\mathbf{p'}, \mathbf{p_1'}, \mathbf{p_2'})$ are aligned.

We also remark that, if $\mathbf{t} = 0$, and A is a positive-definite matrix, then the motion modelized by (2) is a dilation. If A is a negative-definite matrix, then the motion modelized by (2) is a contraction. In conclusion, we can say that the model proposed here can take into consideration several different types of motion.

Let us note the elements of matrix A as following

$$\mathbf{A} = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix}$$

For two successive points $\mathbf{p_{k+1}}$ and $\mathbf{p_k}$ we can write, in accordance with (2),

$$\mathbf{w_{k+1} - w_k} = \mathbf{A}(\mathbf{p_{k+1} - p_k})$$

From this last equation, and considering separately the two components of the velocity vector, we can write for component u, taking into account four successive points

$$\begin{bmatrix} u_{k+1}-u_k & x_{k+1}-x_k & y_{k+1}-y_k \\ u_k-u_{k-1} & x_k-x_{k-1} & y_k-y_{k-1} \\ u_{k-1}-u_{k-2} & x_{k-1}-x_{k-2} & y_{k-1}-y_{k-2} \end{bmatrix} \begin{bmatrix} -1 \\ a_1 \\ b_1 \end{bmatrix} = 0 \quad (4)$$

We also can write similar equations for the components v. A consequence of (4) is that the determinant of the above matrix must be zero,

$$\begin{vmatrix} u_{k+1}-u_k & x_{k+1}-x_k & y_{k+1}-y_k \\ u_k-u_{k-1} & x_k-x_{k-1} & y_k-y_{k-1} \\ u_{k-1}-u_{k-2} & x_{k-1}-x_{k-2} & y_{k-1}-y_{k-2} \end{vmatrix} = 0$$

We can then write

$$D_{1,k}(u_{k+1}-u_k)-D_{2,k+1}(u_k-u_{k-1})+D_{1,k+1}(u_{k-1}-u_{k-2}) = 0$$

with $$D_{1,k} = \begin{vmatrix} x_k-x_{k-1} & y_k-y_{k-1} \\ x_{k-1}-x_{k-2} & y_{k-1}-y_{k-2} \end{vmatrix}$$

and $$D_{2,k+1} = \begin{vmatrix} x_{k+1}-x_k & y_{k+1}-y_k \\ x_{k-1}-x_{k-2} & y_{k-1}-y_{k-2} \end{vmatrix}$$

If the three points $\mathbf{p}_k$, $\mathbf{p}_{k-1}$ and $\mathbf{p}_{k-2}$ are not aligned, then $D_{1,k}\neq 0$, and we can write

$$u_{k+1}=(1+\frac{D_{2,k+1}}{D_{1,k}})u_k-\frac{D_{1,k+1}+D_{2,k+1}}{D_{1,k}}u_{k-1}+\frac{D_{1,k+1}}{D_{1,k}}u_{k-2} \quad (5)$$

which is an autoregressive relation on the velocity. The same relation is valid for the other component of the velocity vector.

3. ESTIMATION OF THE VELOCITY FIELD

We propose to use the autoregressive relation given in the precedent section for estimating the 2D velocity field. Let us consider the equation (5) and write the autoregressive relation for the two velocity components

$$u_{k+1} = \beta_k u_k + \beta_{k-1} u_{k-1} + \beta_{k-2} u_{k-2}$$
$$v_{k+1} = \beta_k v_k + \beta_{k-1} v_{k-1} + \beta_{k-2} v_{k-2}$$

The identification of coefficients $\{\beta_k\}$ is obvious according to (5); they depend on edge line curvature. Of course the above model is not perfect and we have to take into acount a model noise. We designate ξ_k the state vector given by

$$\xi_k = [\, u_k \;\; u_{k-1} \;\; u_{k-2} \;\; v_k \;\; v_{k-1} \;\; v_{k-2} \,]^T$$

The state equation according to the above recursive relations is given below

$$\xi_{k+1} = \begin{bmatrix} \Phi_{k+1|k} & 0 \\ 0 & \Phi_{k+1|k} \end{bmatrix} \xi_k + \omega_k \quad (6)$$

where the noise vector ω_k is zero-mean with covariance matrix $\mathbf{Q}_k$ and the transition matrix $\Phi_{k+1|k}$ is

$$\Phi_{k+1|k} = \begin{bmatrix} \beta_k & \beta_{k-1} & \beta_{k-2} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

The equation of measurement is given by

$$y_k = [\, c_{1k} \;\; 0 \;\; 0 \;\; c_{2k} \;\; 0 \;\; 0 \,]\, \xi_k + z_k$$

where y_k is the measured projection of the velocity vector on the perpendicular to the contour vector and z_k is a measurement noise which is assumed zero-mean with variance R_k. We have

$$[\, c_{1k} \;\; c_{2k} \,]^T = \mathbf{n}_k$$

$\mathbf{n}_k$ being the normal vector. We can write

$$(c_{1k}, c_{2k}) = (f_x(x_k,y_k), f_y(x_k,y_k))$$

if the equation of the contour is known: f(x,y)=0. The system and measurement noise are assumed to be independent and independent between different points.

The problem to solve is the following: given the observation of $\{y_k\,;\, 1\leq k\leq N\}$ on the edge, how to estimate the field of $\{(u_k, v_k) : 1 \leq k \leq N\}$. This is a smoothing problem and we propose to use the two-filter smoothing formula [1], which gives the optimal solution with a quadratic criterion

$$\hat{\xi}_{k|N} = \mathbf{P}_{k|N}(\mathbf{P}^{-1}_{f,k|k-1}\hat{\xi}_{f,k|k-1} + \mathbf{P}^{-1}_{b,k|k}\hat{\xi}_{b,k|k})$$

$$\mathbf{P}^{-1}_{k|N} = \mathbf{P}^{-1}_{f,k|k-1} + \mathbf{P}^{-1}_{b,k|k}$$

where $\hat{\xi}_{f,k|k-1}$ is a forwards optimal prediction and $\hat{\xi}_{b,k|k}$ is a backwards optimal filtering, both based on the same state vector equation (6). $\mathbf{P}_{f,k|k-1}$ and $\mathbf{P}_{b,k|k}$ are the corresponding error covariance matrices. The initial covariance matrix, at k=1, for the forwards filter, and at k=N, for the backwards filter, are assumed sufficiently great in the above formula. A similar approach using a different state equation is presented in [6] where some results concerning the motion of simulated edges are given.

4. CONCLUSION

We have introduced a model for the 2D velocity field, which is adaptable in many domains of applications, and many types of 2D motion or 3D motion projected on the image plane. We have shown how this model may be used to estimate the 2D velocity field on points belonging on edges detected from a sequence of images. The evaluation of the performance of this method in natural images is currently under investigation.

REFERENCES

[1] R.ACKNER and T.KAILATH "Discrete-time complementary models and smoothing" *Int.J.Control*, Vol.49, No.5, 1989, pp.1665-1682.

[2] J.CANNY "A computational approach to edge detection" *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, No. 6, Nov. 1986, pp. 679-698.

[3] R.DERICHE "Using Canny's criteria to derive an optimal edge detector recursively implemented" *Inter J. Comp.Vision*, Vol. 1, No. 2, 1987.

[4] E.HILDRETH, *The Measurement of Visual Motion*, MIT Press, Massachusetts, 1983.

[5] G.MAILLOUX, F.LANGLOIS, P.SIMARD and M.BERTRAND "Restoration of the velocity field of the heart from two-dimensional echocardiograms" *IEEE Trans. Med. Imaging*, Vol.MI-8, No.2, June 1989, pp.143-153.

[6] A.ROUGEE, B.LEVY and A.WILLSKY "Reconstruction of two-dimensional velocity fields as a linear estimation problem" *Proc. of the 1st Int.Conf. on Computer Vision*, London, June 1988, pp.646-650.

[7] G.TZIRITAS "Recursive and/or iterative estimation of the two-dimensional velocity field and reconstruction of three-dimensional motion" *Signal Processing*, Vol.16, Jan. 1989, pp. 53-72.

[8] A.VERRI and T.POGGIO "Against quantitative optical flow" *Proc. of the 1st Int.Conf. on Computer Vision*, London, June 1988, pp.171-180.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

The Flow Analysis Using the Flow Visualization Images with Fuzzy Reasoning

Moriyuki MATSUO

Department of Microwave Telecommunication Technical University of Budapest
Budapest Goldmann tér 3 H-1111 HUNGARY

In order to analyze the flow visualized image,the distance and the direction of the motion of the tracer particle in a flow at a certain interval is obtained using an image processing technique with Fuzzy reasoning. For this purpose two pictures,one is a primary images taken at the time "t",another one is a secondary image taken at the time "t+dt" is used.

1. INTRODUCTION

In the case of a flow visualized image and its analysis using image processing technique,several tracer particles are put into the flow. According to this method a track of the moving tracer particle in a flow in a certain interval. Some methods for obtaining the track of a tracer using image processing technique are developed up to the present, these methods are to be obtaining the track from the primary picture of the particles in a flow at the time "t" and the secondary picture of the particles at the time "t+dt". Generally,each tracer particles,however,do not have a correspondence one to one in two pictures,because many tracer particles are in the flow.
This paper describes to determine each tracer in the primary picture correspond to it in the secondary picture using Fuzzy reasoning using character agree with the path of the tracer stream line and streak line when the flow situation is not change at the time. The flow situation is not uniform, because the flow is affected by a flow hindering thing. But,localy the flow in set of small region and the boundary of the Fuzzy region.A tracer is estimated which Fuzzy region is it belong to.On the other hand,the situation in each Fuzzy region is characterized by moving tracer of approximate angle and distance in a certain interval.

2. THE APPLICATION OF FUZZY LOGIC TO THE FLOW

2.1. The initial point and the end point of the tracer

The tracers threw in to the flow are defined as p_k (k=1,2,..,n) at the time "t" and its position are defined as P_k .The tracers are defined as q_k (k=1,2,..,n) at the time"t+dt"and its positions are defined as Q_k .
If P_k is able to have one to one correspondence to q_k from two still picture of the flow at the time "t" and "t+dt", the moved distance and the angle of the tracer is obtained by to minimize the "dt" easily. That is a velocity vector in the flow is obtain by two still pictures. Generally, however the p_k is not easy to have one to one correspondence to the q_k from two still pictures because a complicated change of the flow situation is produced by each position and it is impossible to have separate marks for every tracers.Because a lot of tracers are threw in to the flow.
For given p_k , to find the q_k correspond to the p_k , the inference rule for forecasting tracer p_k at the time "t" is obtained by Fuzzy reasoning.The P'_k resulting of Fuzzy reasoning is not showed the determinate position. It shows the result position.The p_k and the q_k have one to one correspondence by the determinate position which P'_k is Q_k, when Q_k a position of q_k , in uncertain region.

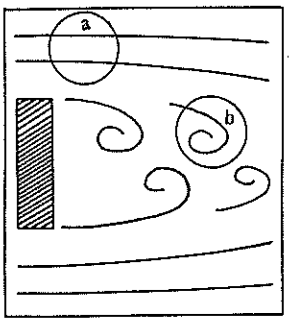
2.2. A region in the flow

It is controlled by the flow situation that the distance and the angle of moving p_k at an interval dt. Because the situation depends upon some physical conditions, for instance there is an object which is injurious to the cause of a stream, the flow situation is different from every other individual sub-region in the flow.The flow division is able to make between a region having straight stream and another region having the vortex, having if there is an obstacle standing squarely in the way of the flow as shown Fig.2.1.The whole flow is is divided the number of m sub-regions and let ω_i (i=1,2,...m) be every sub-regions. On the other hand, it can be to consider that the flow are dependent upon in each region, which are affected the flow of neighboring sub-regions. Accordingly, the boundary of the individual region is not clear each other. The boundary is regarded to be Fuzzy state.So that let Ω be whole space that deal with this paper. It is given by

$$\Omega = \bigcup_{i=1}^{m} \omega_i \qquad (2.1)$$

2.3 Fuzzy logic for a moving tracer in a region

In this section a sub-region $\omega_s \subset \Omega$ in the flow is considered. To find the nearest



a:straight the flow
b:vortex the flow
Fig.2.1 An example of the flow.

position of tracer $p'_k \in \omega_a$ moving from the original position of the tracer $p_k \in \omega_a$ in the interval "dt",n Fuzzy set to obtain p'_k is defined using membership function μ_{A^j} as follows:

$$A^j = \int_R \mu_{A^j}(x^j)/x^j \quad (2.2)$$
$$(j=1,2,\cdots,n,\ A^j \subset x^j)$$

where A^j is the Fuzzy set required to obtain p'_k and x^j is whole set of A^j.
Let $\mu(p)$ be ratio at with P_k belongs to ω_a then $p \in \omega_a$ .The Fuzzy reasoning for to be anticipate the position of the tracer $p'_k \in \omega_a$ is as follows:

[Fuzzy reasoning 1]
Antecedent condition:
If p is ω_a then x^1 is A^1
Antecedent:p is ω'

consequence: x^1 is $A^{1'}$

[Fuzzy reasoning n]
Antecedent condition:
If p is ω_a then x^n is A^n
Antecedent:p is ω'

consequence: x^n is $A^{n'}$

The Fuzzy sets denote by p'_k can be finded from the results($A^{1'}$,$\cdots$,$A^{n'}$) which is obtained this reasoning.

2.4 Fuzzy logic for a tracer moving over the several regions

In this section the number of m sub-regions $\omega_i \subset \Omega$ (i=1,2, $\cdots$,m) in the flow is considered. As mentioned in the previous section,the position p'_k for the potion p_k will be forecasted. The number of mn Fuzzy sets necessary to determine p'_k in each sub-region are defined as follows:

$$A_i^j = \int_{x^j} \mu_{A_i^j}(x^j)/x^j \quad (2.3)$$
$$(i=1,2,\cdots,m,j=1,2,\cdots,n,\ A_i^j \subset x^j)$$

where A_i^j is the Fuzzy set which is necessary to obtain p'_k and x^j is whole set of A_i^j.
And let $\mu(p)$ be ratio at which p_k belongs to ω_i .The Fuzzy reasoning for the anticipation the position of the tracer p'_k is as follows:

[Fuzzy reasoning 1]
Antecedent condition 1:
If p is ω_1 then x^1 is A_1^1 else
Antecedent condition 2:
If p is ω_2 then x^1 is A_2^1 else
Antecedent condition m:
If p is ω_m then x^1 is A_m^1
Antecedent:p is ω'

consequence: x^1 is $A^{1'}$

[Fuzzy reasoning n]
Antecedent condition 1:
If p is ω_1 then x^n is A_1^n else
Antecedent condition 2:
If p is ω_2 then x^n is A_2^n else
Antecedent condition m:
If p is ω_m then x^n is A_m^n
Antecedent:p is ω'

consequence: x^n is $A^{n'}$

Fuzzy conditional proportion "If p is ω_i then x^i is A^i " is called Fuzzy relation R as follows:

$$R = \omega \rightarrow A = \omega \times A. \quad (2.4)$$
($\times$ is a Cartesian product)

Although method to define a Cartesian product of Fuzzy set is "min" operation of the membership function or algebraic product, in this paper, the "min" operation is used. The positional rule of the reasoning consequence from Fuzzy condition and antecedent as follows:

$$\omega' * (\omega \rightarrow A) = \omega' * (\omega \times A) = A' \quad (2.5)$$

where the symbol * means max-* composition. Furthermore various operation is able to define as the symbol "*",according to the result of the inference,a operation can select advantageous to obtain the result.

3. THE INFERENCE FOR THE END POINT CORRESPONDING TO THE INITIAL POINT

3.1. Fuzzy logic for the determination of magnitude and the direction of the vector

The vector direction from the position of the initial point P'_k to the position of the end point P'_k in the flow is considered.
The vector is defined as V_k and denotes as follows.

$$V_k = r_k \exp(j\theta_k) \quad (3.1)$$

Thus,let $A_i \subset R$ be Fuzzy set to denote an approximate distance and let $B_i \subset \Theta$ be Fuzzy set to an approximate angle, whose the membership functions are defined as $\mu_{Ai}(r)$ and $\mu_{Bi}(\theta)$.Furthermore,let the membership function be $\mu_{wi}(p)$ which is the ratio at p_k belong to ω_i .Fuzzy reasoning are application to following reasoning by making n equal to 2:

[Fuzzy reasoning 1]
Antecedent condition 1:
If p is ω_1 then r is A_1 else
Antecedent condition 2:
If p is ω_2 then r is A_2 else

:
Antecedent condition m:
If p is ω_m then r is A_m
Antecedent:p is ω'

consequence: r is A '

:
:

[Fuzzy reasoning 2]
Antecedent condition 1:
If p is ω_1 then θ is B_1 else
Antecedent condition 2:
If p is ω_1 then θ is B_2 else
:
:
Antecedent condition m:
If p is ω_m then θ is B_m else
Antecedent:p is ω'

consequence: θ is B'

Fuzzy relation of there reasonings can be shown that

$$R_1 = \omega \rightarrow A = \omega \times A. \quad (3.2a)$$
$$R_2 = \omega \rightarrow B = \omega \times B. \quad (3.2b)$$

Furthermore,the composition rule of Fuzzy condition and antecedent to infer from consequence make following expression:

$$\omega' * (\omega \rightarrow A) = \omega' * (\omega \times A) = \omega \circ A' (\omega \times A) = A' \quad (3.3a)$$
$$\omega' * (\omega \rightarrow B) = \omega' * (\omega \times B) = \omega \circ B' (\omega \times B) = B' \quad (3.3b)$$

where Fuzzy relations R_1 ,R_2 are defined as product of the membership functions. A symbol 'o' means max-min composition. Furthermore, 'else' in Fuzzy conditional proposition correspond to U(join). Let max be symbol 'V' let min be symbol '$\wedge$' . Composition rule is as follows:

$$A' = \omega' \circ (\omega_1 \times A_1) \cup \omega' \circ (\omega_2 \times A_2) \cdots \cup \omega' \circ (\omega_1 \times A_1) \cdots \cup \omega' \circ (\omega_m \times A_m) \quad (3.4)$$

$$\mu A'(r) = \{\mu_W(p) \wedge (\mu_{W1}(p) \cdot \mu_{A1}(r))\} \vee \{\mu_{W'}(p) \wedge (\mu_{W2}(p) \cdot \mu_{A2}(r))\} \cdots \vee \{\mu_{W'}(p) \wedge (\mu_{Wi}(p) \cdot \mu_{Ai}(r))\} \cdots \vee \{\mu_{W'}(p) \wedge (\mu_{Wm}(p) \cdot \mu_{Am}(r))\} \quad (3.5)$$

$$B' = \omega' \circ (\omega_1 \times B_1) \cup \omega' \circ (\omega_2 \times B_2) \cdots \cup \omega' \circ (\omega_1 \times B_i) \cdots \cup \omega' \circ (\omega_m \times B_m) \quad (3.6)$$

$$\mu B'(\theta) = \{\mu_W(p) \wedge (\mu_{W1}(p) \cdot \mu_{B1}(\theta))\} \vee \{\mu_{W'}(p) \wedge (\mu_{W2}(p) \cdot \mu_{B2}(\theta))\} \cdots \vee \{\mu_{W'}(p) \wedge (\mu_{Wi}(p) \cdot \mu_{Bi}(\theta))\} \cdots \vee \{\mu_{W'}(p) \wedge (\mu_{Wm}(p) \cdot \mu_{Bm}(\theta))\} \quad (3.7)$$

The approximate position of the end poit can obtained by the above expression.

4. EXPERIMENT

4.1. A preprocess of two images of the initial point the end point

In a practical flow, the initial point of the tracers in the primary picture have done one correspondence to the end point or the tracers in the secondary pictures using Fuzzy reasoning.There is a quadrangular prisim as an obstacle in the flow (Photo.4.1,4.2). And the four circles are in the each corner as the marker for taking in agreement with two images The preprocess to extract position of the tracer using Fuzzy reasoning is given the following procedure.

[Procedure 1] A picture of an initial point in the flow are feeded by TV-camera to the image processor. The given image is performed by binarization. Furthermore, noise is removed by using the eight-neighbors of an isolated point.

[Procedure 2] The connecting components in image given by Procedure [1] are labeled and obtained a size of an area and the center of the gravity. The connecting components can be classified according to the area because the size of each areas is different from each others.And then,let each center of the gravity is coordinated.

[Procedure 3] [Procedure[1] and [2] are performed in regard to the picture of the end point and it should be obtained a gap of a position to the end point. And the position of the end point image is corrected by transferring and rotating of the image.

Photo.4.1 and 4.2 are showed the position of the initial point and the end point in Table 4.1 after the performing preprocess and computing.

4.2. Inference of the end point using Fuzzy reasoning

The end point corresponding to the initial point is obtained using Fuzzy reasoning from the position of the initial point and the end point of the tracer after the preprocessing. The flow is divide into sixsub-regions.The membership functions($\mu_{wi}(p)$,i=1,2,$\cdots$,6) meaning ratio at an initial point belongs to the region must be defined,since sub-regions are Fuzzy regions.In practice ,the regions are selected by a joystickas watching the image on the monitor and the regions are divided. $\mu_{wi}(p)$ in each region is given as follows,that is an effect of some neighborhoods is in proportion to the distance. Where x and y are the position of the initial point respectively. The membership functions on the distance and the angle in each region ($\mu_A(r)$,$\mu_B(\theta)$) are given depending on the flow velocity without an obstacle and the shape of the obstacle in the flow. The position of the end point is estimated by using equations (3.5) and (3.7), and by these membership functions and the position of the initial point of the tracers. And then, the practical end point being with correspondence.

4.3. A result of an experimental

Table 4.2 is a result the end point an image of the initial point in Photo 4.1 and is with correspondence to the end point. And the result is indicated by the graphical display as shown in Fig.4.1. Twenty-four tracers being with one to one correspondence,seven initial points be with miss correspondence,for the initial points of thirty-one tracers.

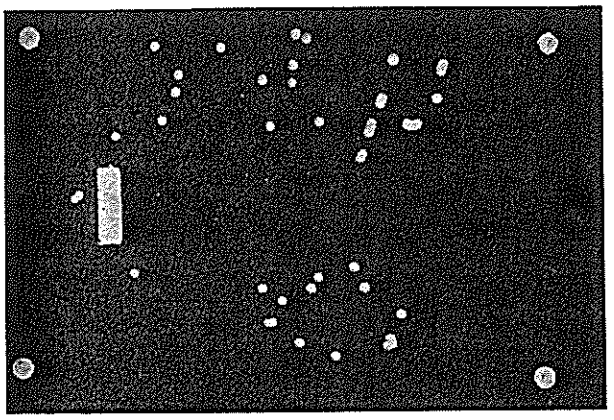


Photo.4.1
A primary picture of the tracer perticles.

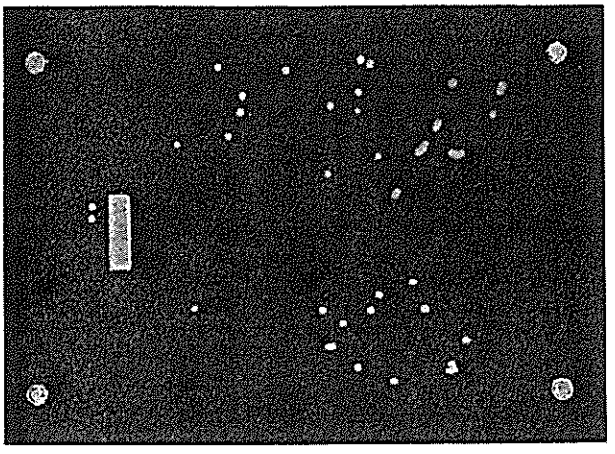


Photo.4.2
A secondary picture of the tracer perticles.

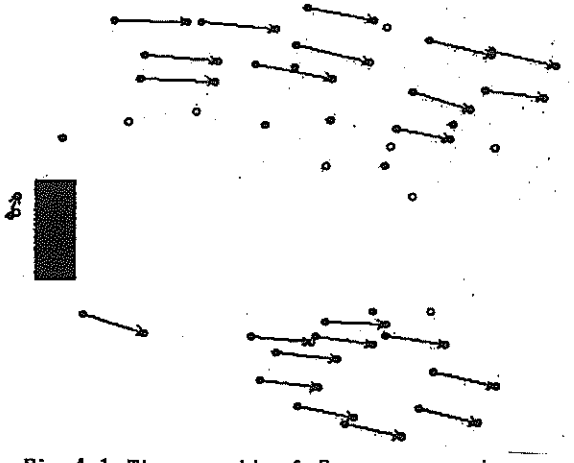


Fig.4.1 The result of Fuzzy reasoning.

Table 4.1
The tracer position of the initial point and the end point.

No.	X	Y	No.	X	Y
1	137	136	1	147	200
2	193	137	2	157	334
3	339	149	3	74	258
4	254	152	4	75	247
5	253	166	5	184	137
6	228	164	6	190	194
7	157	159	7	241	142
8	154	173	8	264	340
9	104	210	9	202	175
10	234	202	10	273	228
11	276	199	11	277	173
12	376	181	12	281	351
13	118	322	13	292	388
14	304	320	14	301	163
15	273	327	15	204	162
16	226	336	16	304	342
17	267	336	17	312	141
18	312	336	18	312	329
19	242	346	19	315	216
20	343	359	20	322	400
21	256	381	21	269	369
22	286	392	22	341	320
23	261	128	23	304	136
24	380	156	24	350	342
25	329	182	25	329	247
26	319	205	26	366	193
27	355	202	27	371	391
28	311	228	28	380	159
29	71	259	29	383	368
30	232	364	30	414	185
31	333	382	31	421	165
			32	353	211
			33	382	217

Table 4.2
The result of Fuzzy reasoning.

No.		No.	
I.P.	E.P.	I.P.	E.P.
1	5	17	16
2	7	18	24
3	28	19	12
4	14	20	29
5	—	21	13
6	11	22	20
7	15	23	23
8	9	24	31
9	—	25	26
10	—	26	32
11	—	27	—
12	30	28	—
13	2	29	4
14	—	30	21
15	18	31	27
16	8		

5. CONCLUSIONS

In order to analyze the flow visualized image, the distance and the direction of the motion of the tracer particle in a flow at a certain interval is obtained using an image processing technique with Fuzzy reasoning.
Firstly,the flow is divided into the Fuzzy regions and Fuzzy reasoning is giving the definition at the each Fuzzy region.
Second, the vector directed from an initial point to an end point is supposed, and the inferential method of the distance and the angle of the vector make it clear.
As processing the two pictures of the initial points and the end points the position of the tracers are detected.
After the flow is divided into the several region and Fuzzy reasoning is inferred by given membership functions which is expressed an approximate distance and angle. Consequently the initial points can be corresponded to the end points exactly in the region which is a little change of the flow. However the initial points do not be corresponded the end points in the region which is in vortex.
In the future, it is necessary to study more how to infer Fuzzy reasoning and to divide the region which is in vortex.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

Motion Field Estimation by 2-D Kalman Filtering *

J.N. Driessen and J. Biemond

Delft University of Technology,
Department of Electrical Engineering, Information Theory Group,
P.O. Box 5031, 2600 GA Delft, The Netherlands.

Abstract

In this paper a new pixel-recursive estimator is derived for the extraction of the motion field from two consecutive images. The estimator is based on a recursive 2-D AR vector model for the motion field and a nonlinear equation for the observation of the motion field in the time-varying intensities. Special cases of such models are predictors used in earlier proposed pixel-recursive algorithms. To reduce the computational complexity involved in the true 2-D Kalman filter, a sub-optimal one is proposed based on reduced-order model strategies. The update equation for the first two state elements in the filter is quite similar to earlier proposed updates. Theoretically, the advantage over the earlier proposed pixel-recursive estimators is the explicit assumption of motion field models and the incorporation of estimation error covariances. In practice some care has to be taken in updating these error covariances due to imperfect linearization of the nonlinear observation model.

1 Introduction

A motion field is the projection of 3-D object motion *onto* the image plane and it represents local displacements *in* the image plane. In an intensity image sequence the motion field relates intensities in consecutive frames along the so-called *motion trajectory* and in motion compensated processing of image sequences such as coding, filtering or interpolation the motion field is essential information. Prior to this processing, the motion field has to be estimated from the consecutive frames which is a difficult problem for two main reasons. A scene often consists of differently moving objects which causes the motion field to exhibit discontinuities at object boundaries and even regions in which the motion field is undefined due to uncovered object parts. Moreover, even within a moving object the estimation of the motion field is locally an ill-posed problem since only one intensity value variation is observed for two unknown vector components.

A possible solution to these problems is provided by a parameterization of the motion field as a function of 3-D object motion and surface parameters [1,2]. Such an approach turns the problem into a simultaneous parameter estimation and image segmentation problem. In such models, motion field discontinuities are included implicitely as the boundaries of segmented objects; uncovered regions can be found by testing the model hypothesis. Unfortunately, parametric models are more or less restricted to cover rigid motion of very smooth 3-D object surfaces. Another solution is the explicit assumption of smoothness of the motion field within a moving object [3,4]. Discontinuities are dealt with by adapting the smoothness constraint to intensity gradient information [3]. Intuitively, these techniques capture a more general class of motions than the parametric techniques. Since the algorithms developed here are intended to be used for real life sequences attention is focussed on these smoothness-based techniques.

The scope of this paper is to present a pixel-recursive motion field estimator technique that is based on a causal stochastic motion field model, a nonlinear observation equation and Kalman filter solution techniques. Originally, pixel-recursive techniques were introduced as adaptive estimators without explicit underlying modeling assumptions about the motion field: they consisted of a horizontal displacement vector predictor and a vector update that was based on the minimization of a very local functional [5]. In the following years, attempts were reported to incorporate more a priori model-knowledge into the estimators such as the assumption that the motion field can be modeled as a one-dimensional first order AR-process and estimated by a 1-D Kalman filter [6], that the update is a stochastic variable [7] and that the one-dimensional prediction can be improved by a two-dimensional prediction [8]. The estimator presented in this paper can be viewed as a generalization of these approaches.

2 Modeling Assumptions

In this section, the models which the estimator is based upon are presented. At first, the motion field model is described which is essentially a recursive stochastic model to cover structured but non-rigid motion. Finally, the motion field observation model is discussed which is essentially a non-linear model to cover large local displacement vectors.

2.1 Causal AR Motion Field Models

The motion field is assumed to be generated by the following causal 2-D AR vector process:

$$\vec{d}(x,y) = \sum_{(i,j)\in S} A_{ij}\vec{d}(x-i,y-i) + \vec{v}(x,y), \qquad (1)$$

with $\vec{d}(x,y)$ a motion field vector at location (x,y), with A_{ij} 2×2 matrices representing the AR-model parameters, with S a causal support region and with $\vec{v}(x,y)$ the driving noise which is assumed to be a zero-mean Gaussian random process with covariance matrix P_v. The AR-model matrices A_{ij} together

*This research is supported by the Netherlands Technology Foundation (STW).

with the noise covariance matrix P_v determine the smoothness of the motion field.

A convenient and elegant choice for the model matrices arises if it is assumed that the motion field has a decoupled separable autocorrelation function for each vector field component:

$$\begin{aligned} R_d(r,s) &= E[\vec{d}(x,y)\vec{d}^T(x-r,y-s)] \\ &= \begin{pmatrix} \sigma_x^2 \rho_{hx}^{|r|} \rho_{vx}^{|s|} & 0 \\ 0 & \sigma_y^2 \rho_{hy}^{|r|} \rho_{vy}^{|s|} \end{pmatrix}. \end{aligned} \tag{2}$$

Interestingly, a similar autocorrelation model has been proposed very recently by Namazi and Lee [9].

In natural scenes there is no reason for a difference between the correlation of the horizontal or vertical motion field component, so the correlation coefficients and the variances are assumed to be equal for both components:

$$\begin{aligned} \sigma^2 &= \sigma_x^2 = \sigma_y^2, \\ \rho_h &= \rho_{hx} = \rho_{hy}, \\ \rho_v &= \rho_{vx} = \rho_{vy}. \end{aligned} \tag{3}$$

If the horizontal and vertical sampling frequencies are equal, there is also no reason for a difference between the horizontal and vertical correlation coefficient in each vector component:

$$\rho = \rho_h = \rho_v \tag{4}$$

The AR model matrices for such a model are given by:

$$\begin{aligned} A_{10} &= \rho I, \quad A_{-11} = 0, \\ A_{01} &= \rho I, \quad A_{11} = -\rho^2 I \end{aligned} \tag{5}$$

and the noise covariance matrix is given by:

$$P_v = \sigma^2(1-\rho^2)^2 I. \tag{6}$$

The predictor of Tziritas [8] results from this model as a special case for the correlation coefficient equal to one and a zero noise covariance:

$$\vec{d}(x,y) = \vec{d}(x-1,y) + \vec{d}(x,y-1) - \vec{d}(x-1,y-1). \tag{7}$$

2.2 A Nonlinear Observation Equation

The motion field is observed as displaced intensities in an intensity image according to the following nonlinear observation model:

$$f(\vec{x},t) = f(\vec{x} - \vec{d}(x,y), t-dt) + w(\vec{x},t), \tag{8}$$

where $f(\vec{x},t)$ is the observed intensity at time t at spatial location (x,y), dt is the temporal distance between consecutive frames and where $w(\vec{x},t)$ is the observation noise which is assumed to be zero-mean Gaussian random process with variance σ_w^2. The assumption underlying model equation (8) is that the intensities along the motion trajectory are more or less constant. This assumption is not true in general and the noise term $w(\vec{x},t)$ accounts for small errors in this assumption and is *not* intended to account for noise present *in* the consecutive frames. This is indeed a simplification , however, it prevents the difficult combined motion estimation and image sequence filtering problem. Finally, notice that Eq. (8) provides only one equation in two unknown motion field vector components, which shows the ill-posedness of the estimation problem without *a priori* motion field model knowledge.

3 2-D Kalman Filter Solutions

In this section, we present a Kalman filter based on the models previously discussed. At first, to formulate the problem as a Kalman filtering problem, the motion field model and the observation equation are presented in a state-space formulation. Next, to reduce the computational complexity, the optimal Kalman filter has to be approximated based on sub-optimal approaches. Finally, it is shown that the update equation proposed here covers earlier proposed pixel-recursive update formulas as a degenerated case.

3.1 State-Space Formulation

To formulated the state-space equations, the image is assumed to be scanned line-wise starting at the most upper line and each line is assumed to be scanned from the left to the right. The state is defined as the set of vector elements of the motion field belonging to all previously scanned pixel locations that determine future vector elements of the motion field according to model (1). For a first-order AR vector model this results in the following state vector:

$$\begin{aligned} \vec{s}(x,y) &= [\vec{d}^T(x,y), \vec{d}^T(x-1,y), \cdots, \vec{d}^T(1,y), \\ &\quad \vec{d}^T(M,y-1), \cdots, \vec{d}^T(x,y-1)]^T, \end{aligned} \tag{9}$$

with M the number of pixels on a row. The state-space evolution equation is given by:

$$\vec{s}(x+1,y) = A\vec{s}(x,y) + C\vec{v}(x,y), \tag{10}$$

where the matrices A and C are the system matrices that have to be chosen appropriately. For a first-order AR motion field model, A is given by:

$$A = \begin{pmatrix} A_{10} & 0 \cdots 0 & A_{-11} & A_{01} & A_{11} \\ & & A_{sub} & & \end{pmatrix}, \tag{11}$$

where the matrices A_{ij} are the AR model matrices and where A_{sub} is a matrix containing zeroes and ones that only perform shifts of state elements. The driving noise $\vec{v}(x,y)$ only affects the first two state elements so the matrix C is given by:

$$C = \begin{pmatrix} I & 0 & \cdots & 0 \end{pmatrix}^T. \tag{12}$$

with I the 2×2 identity matrix and 0 the 2×2 zero matrix.

The state-space observation equation is given by:

$$f(\vec{x};t) = f(\vec{x} - H\vec{s}(x,y); t-dt) + w(\vec{x},t), \tag{13}$$

where H performs the extraction of the actual displacement from the state vector $\vec{s}(x,y)$:

$$H = \begin{pmatrix} I & 0 & \cdots & 0 \end{pmatrix}. \tag{14}$$

In the optimal extended Kalman filter, outlined in App. A, the first-order derivatives of the observation equation with respect to the state elements is needed. These derivatives are evaluated as follows:

$$\begin{aligned} F(\vec{x},\vec{s}(\vec{x})) &= \nabla_{\vec{s}} f(\vec{x} - H\vec{s}(\vec{x}); t-dt) \\ &= -H^T \nabla_{\vec{x}} f(\vec{x} - \vec{d}(\vec{x}), t-dt). \end{aligned} \tag{15}$$

Although the true extended Kalman filter is straightforward to derive, its implementation is computationally expensive due to the large dimension of the state which is in the case of a first-order model of order $O(2M)$.

3.2 Sub-Optimal Solutions

In the field of image restoration, two recursive sub-optimal 2-D Kalman filters that do not pose restrictions on the stationarity of the state-space models have been proposed [10,11]. The reduced update Kalman filter (RUKF) [10] assumes that the gain in the optimal Kalman filter is relatively close to zero outside a small region called the *local state*. The number of gain elements outside the local state is set to zero and since this number is quite large this reduces the computational load by applying a smart implementation. The reduced-order model Kalman filter (ROMKF) [11] is based on a reduction of the dimension of the state in the true state-space equations and Kalman filtering applied to the resulting state-space equations. The RUKF is a sub-optimal Kalman filter based on the original 2-D models, where the ROMKF is an optimal Kalman filter based on simplified state-space equations. The computational complexity of the ROMKF is less than the complexity of the RUKF with only a minor decrease in performance.

For a first-order AR motion field model the reduced-order model state is defined, similar to [11], by:

$$\begin{aligned}\vec{s}_R(x,y) &= [\vec{d}^T(x,y), \vec{d}^T(x-1,y), \vec{d}^T(x+2,y-1), \\ &\quad \vec{d}^T(x+1,y-1), \vec{d}^T(x,y-1)]^T. \end{aligned} \tag{16}$$

The state evolution equation is given by:

$$\vec{s}_R(x+1,y) = A_R\vec{s}_R(x,y) + B_R\vec{u}(x,y) + C_R\vec{v}_R(x,y), \tag{17}$$

where $\vec{u}(x,y)$ is an input variable that represents the most recent estimate of the displacement vector at the spatial location $(x+3, y-1)$:

$$\vec{u}(x,y) = \vec{d}_e(x+3,y-1) + \vec{v}_u, \tag{18}$$

where $\vec{v}_u$ accounts for the uncertainty in the estimate. The noise term $\vec{v}_R$ includes the driving noise $\vec{v}$ from the original model together with $\vec{v}_u$:

$$\vec{v}_R = \begin{pmatrix} \vec{v} \\ \vec{v}_u \end{pmatrix}. \tag{19}$$

Under the assumption that both components in the noise term are independent, the covariance matrix equals:

$$P_{v_R} = \begin{pmatrix} P_v & 0 \\ 0 & P_u \end{pmatrix}. \tag{20}$$

The model matrices are given by the following equations:

$$A_r = \begin{pmatrix} A_{10} & 0 & A_{-11} & A_{01} & A_{11} \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \end{pmatrix}, \tag{21}$$

$$B_r = \begin{pmatrix} 0 & 0 & I & 0 & 0 \end{pmatrix}^T. \tag{22}$$

$$C_r = \begin{pmatrix} I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \end{pmatrix}^T. \tag{23}$$

$$H_r = \begin{pmatrix} I & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{24}$$

with again I the 2×2 identity matrix and 0 the 2×2 zero matrix.

In this specific case the state dimension is of order 10 and the implementation can be done very efficient due to the large number of zeroes in the matrices and the large number of shift operations performed by the identity matrices in the system matrices. The computational complexity of this ROMKF is reduced substantially compared with the original Kalman filter.

3.3 Relation with Existing Algorithms

To show the relation with existing pixel-recursive algorithms, the updating equation for the first two state elements is considered into detail. The equation for the gain matrix is given by:

$$K = -P_b H^T \nabla f [\nabla^T f H P_b H^T \nabla f + \sigma_w^2]^{-1}, \tag{25}$$

where the spatial coordinates are omitted for convenience. Notice first that the inversion is a simple scalar inversion. A closer look at the matrix $HP_b(x,y)H^T$, tells us that the operation $H \cdot H^T$ extracts the most upperleft 2×2 submatrix of the covariance matrix. This matrix is refered to a P_d, since it represents the prediction error covariance matrix of, the actual displacement vector. The post multiplication of the matrix P_b by H^T extracts the first two columns of the state error covariance matrix. The first two elements of the gain, refered to as K_d, are thus given by:

$$K_d = -P_d \nabla f [\nabla^T f P_d \nabla f + \sigma_w^2]^{-1}. \tag{26}$$

The update of the first two elements results by multiplying $K_d(x,y)$ by the innovation term refered to as $dfd(\vec{x}, \vec{d}_b, t)$:

$$\begin{aligned} dfd(\vec{x}, \vec{d}_b, t) &= f(\vec{x}, t) - f(\vec{x} - H\vec{s}_b, t - dt) \\ &= f(\vec{x}, t) - f(\vec{x} - \vec{d}_b, t - dt) \end{aligned} \tag{27}$$

Let the covariance matrix be restricted to a diagonal matrix with σ_d^2 as diagonal elements. In that case the update for the first two elements of the state is given by:

$$\vec{u}_d = [\|\nabla^T f\|^2 + \mu]^{-1} dfd(\vec{x}, \vec{d}_b, t) \nabla f, \quad \mu = (\frac{\sigma_v}{\sigma_d})^2. \tag{28}$$

This equation is equal to the Wiener-based update for an observation window containing one element only.

4 Initial Experimental Results

Initial experimental results show that to obtain reliable smooth motion field realization from the 2-D AR vector model, the correlation parameter has to be fairly large (> 0.9) and the driving noise variance has to be fairly small ($< 10^{-3}$). This is in contrast with [9] who used much lower values for their estimator. In our experiments, the values according to the above specified ranges where used. Furthermore, small variances for the observation noise component where used (< 2.0).

Next, due to the large correlation, the small driving noise variance, the Kalman gain cannot be assumed to be relatively close to zero outside the local state. Therefore, in order to keep the modeling error in the ROMKF small, the state has to be

extended on the left hand side. If the ROM-state is extended with L vector elements, its state becomes:

$$\begin{aligned}\vec{s}_R(x,y) &= [\vec{d}^T(x,y),\cdots,\vec{d}^T(x-L-1,y),\\ &\quad \vec{d}^T(x+2,y-1),\vec{d}^T(x+1,y-1),\\ &\quad \vec{d}^T(x,y-1)]^T. \end{aligned} \tag{29}$$

The dimension of this state is increased with $2L$ since every vector has two components. Extension of the state at the right hand side has no effect which is due to the reduced-order modeling assumptions.

To more or less solve the boundary value problem, which we feel to be of even more significance in the nonlinear case compared with the linear case, each first row has been processed by a 1-D Kalman filter. Such an strategy has been used since the only reasonable boundary value seems to be zero, which will be very disturbing when this is not the true value. This is in large contrast with the boundary value problem im image restoration, where the available blurred and noisy data can be used to derive boundary values [10,11].

The final set of experiments were performed to investigate the errors in the filter due to linearization errors. In [6], it has been analyzed that the 1-D Kalman filter may become instable due to linearization errors and a very accurate linearization has been suggested. However, in their experiments linearization based on a least squares plane fit were used. It may be better to make a distinction between two operation modes of the extended Kalman filter used here: a convergence mode and a tracking mode. In the convergence mode a global linearization and in the tracking mode a local and accurate linearization has to be performed.

5 Discussion and Future Research

In this paper, a 2-D Kalman filter solution to the problem of estimating motion fields from consecutive frames has been proposed and analyzed theoretically. Existing pel-recursive techniques were shown to be degenerated cases of the Kalman filter. Sofar, experimental results concentrated on the convergence capabilities.

Future research will concentrate on performance comparison between the Kalman filter based algorithm and the existing pel-recursive algorithms. Major attention is focussed on elegant techniques to circumvent instabilities due to linearization errors and the inclusion of discontinuities in the motion field model.

A The Extended Kalman Filter

In this appendix the extended Kalman filter for a nonlinear observation equation is outlined briefly. Let the state-space equations be given by:

$$\begin{aligned}\vec{s}(x+1,y) &= A\vec{s}(x,y)+B\vec{u}+C\vec{v},\\ \vec{z}(x,y) &= \vec{f}(\vec{s}(x,y))+\vec{w}(x,y),\end{aligned} \tag{30}$$

where $\vec{h}(\vec{s}(x,y))$ is a nonlinear vector function of the state vector $\vec{s}(x,y)$. The extended Kalman filter is divided into a prediction and an update part. The prediction part follows the normal linear prediction equations:

$$\begin{aligned}\vec{s}_b(x+1,y) &= A\vec{s}_a(x,y)+B\vec{u},\\ P_b(x+1,y) &= AP_a(x,y)A^T+CP_vC^T,\end{aligned} \tag{31}$$

where the b means *before* updating and where the a means *after* updating. P_b and P_a are the state error covariance matrices before and after updating, respectively, or in other words they are the prediction and estimation error covariance matrices.

The update part of the extended Kalman filter differs from the linear Kalman filter and is given by:

$$\vec{s}_a(x,y) = \vec{s}_b(x,y)+K(x,y)[\vec{z}-\vec{h}(\vec{s}_b(x,y))], \tag{32}$$

$$K(x,y) = P_b(x,y)F^T[FP_b(x,y)F^T+P_v]^{-1}, \tag{33}$$

$$P_a(x,y) = [I-K(x,y)F^T]P_b(x,y),$$

with $K(x,y)$ the so-called Kalman gain, I the identity matrix with dimension equal to the state dimension and with F the *Jacobian* of the nonlinear vector function $\vec{f}(.)$ with respect to the state vector:

$$F=\begin{pmatrix} \frac{\partial f_1(\vec{s})}{\partial s_1} & \cdot & \frac{\partial f_1(\vec{s})}{\partial s_{N_s}} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \frac{\partial f_{N_f}(\vec{s})}{\partial s_1} & \cdot & \frac{\partial f_{N_f}(\vec{s})}{\partial s_{N_s}} \end{pmatrix}, \tag{34}$$

with N_s and N_f the dimension of the state and the vector function $\vec{f}(.)$, respectively. A familiar choices for the location to evaluate the Jacobian is the the most recent estimate for the state: $\vec{s}_b(x,y)$.

References

[1] Murray, D.W. and B.F. Buxton, "Scene Segmentation from Visual Motion using Global Optimization", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 2, March 1987, pp. 220-228.

[2] Hötter, M. and R.Thoma, "Image Segmentation based on Object Oriented Mapping Parameter Estimation", *Signal Processing* 15 (1988), pp. 315-334.

[3] Nagel, H.H. and W. Enkelmann, "An Investigation of Smoothness Constraints for the Estimation of Displacement Vector Fields from Image Sequenecs", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol PAMI-8, no. 5, September 1986.

[4] Anandan, P, "A Computational Framework and an Algorithm for the Measurement of Visual Motion", *International Journal of Computer Vision*, vol. 2, no. 3, 1989, pp. 283-310.

[5] Netravali, A.N. and J.D. Robbins, "Motion-Compensated Television Coding: Part I", *Bell System Technical Journal*, BSTJ-58, no. 3, March 1979, pp. 631-670.

[6] Stuller, J.A. and G. Krishanamurthy, "Kalman Filter Formulation of Low-Level Television Image Motion Estimation", *Computer Vision, Graphics and Image Processing*, vol. CVGIP-21 (1983), pp. 169-204.

[7] Biemond, J., L. Looijenga, D.E. Boekee and R.H.J.M. Plompen, "A Pel-Recursive Wiener-based Displacement Estimation Algorithm", *Signal Processing* 13 (1987), pp. 399-412.

[8] Tziritas, G., "Displacement Estimation for Image Predictive Coding and Frame Motion-Adaptive Interpolation", *Visual Communications and Image Processing '88*, T. Russell Hsing Ed., Proc. SPIE 1001, pp. 936-941 (1988).

[9] Namazi, N.M. and C.H. Lee, "Nonuniform Image Motion Estimation from Noisy Data", *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-38, no. 2, February 1990, pp. 364-366.

[10] Woods, J.W. and C.H. Radewan, "Kalman Filtering in Two Dimensions", *IEEE Transactions on Information Theory*, vol. IT-23, no. 4, July 1977, pp. 473-482.

[11] Angwin, D.L. and H. Kaufman, "Image Restoration using Reduced-Order Models", *Signal Processing* 16 (1989), pp. 21-28.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

EFFECTS OF MOTION ESTIMATION ERRORS ON VOLUMETRIC AND PICTORIAL RECONSTRUCTION

Aldo GRATTAROLA, Sandro ZAPPATORE

DIST - Università di Genova
Via Opera Pia 11 A - 16145 - GENOVA - ITALY

The paper deals with an integrated system devoted to volumetric and pictorial object reconstruction from a set of bidimensional perspective views, acquired with a standard TV camera. The critical aspect issue of the system is the calibration of each view, i.e. the accuracy in determining position and orientation of the viewpoint. In the proposed system such a calibration is obtained via a motion recovery technique from corresponding points. The effects of calibration errors, produced by imperfect data, on the quality of the volumetric model built from occluding contours are analyzed by means of simulation. We finally propose a new regularization technique, based on physical constraints concerning the reconstruction geometry, significantly improving the results in very preliminary simulations.

1. INTRODUCTION

The recovery of 3D information from a sequence of 2D views is a basic issue in many applications, like computer vision, robotics, object recognition, image processing and computer graphics [1-3]. According to the application, the sequence of views is obtained by a set of static sensors, or by means of a relative motion between the objects and the sensors.

This paper deals with an integrated system devoted to volumetric and pictorial object reconstruction from a set of bidimensional perspective views, acquired with a standard TV camera and calibrated by means of a motion recovery algorithm.

The basic modules of the proposed system are:

- an intrinsic camera parameters estimator (off line) based on one of the available methods [4];
- an extractor of corresponding point features in the different views. The used approach selects the candidate points in each frame and finds the correspondences by a relaxation labeling procedure (modified from [5]). In this step the object-background segmentation is also performed;
- a calibration module for the extrinsic camera parameters based on a motion estimation algorithm providing, for each 2D view, the spatial position of the viewpoint and the camera orientation. The motion estimation is performed according to a modified linear approach [6], that improves the robustness of the results with respect to errors in the input data by imposing suitable regularization constraints;
- a volumetric reconstructor of the object under analysis by means of an occluding contour technique: the volumetric reconstruction is based on an efficient 3D model generated by intersecting the infinite conic-like volumes obtained from the silhouettes of the perspective views [7];
- a module that analyzes the pictorial information and integrates it with the 3D model with a resolution that is independent of the volumetric resolution, and can reach the detail level of the original images [7].

Relevant features of the proposed system are: the use of motion estimation to calibrate the perspective views, so that the 3D reconstruction can be performed also if the environment is not enough structured or under control of the observer; an efficient volumetric representation, exploiting a run-length coding technique to minimize both the storage requirement and the rendering computational burden; the separate representation of the pictorial information with resolution independent from the volumetric one; the capability of iteratively refining the 3D model when new views become available without restarting the whole procedure.

This paper does not describe the volumetric and pictorial reconstruction modules, that have already been discussed elsewhere (see, e.g., [7]) and focuses on:

i) effects of errors in the estimates of relative positions and orientations of the viewpoints on the quality of the volumetric and pictorial reconstruction;
ii) requirements on the original data to constrain such errors within a range providing an acceptable reconstruction.
iii) a new proposed regularization technique based on geometrical constraints devoted to improve the reconstruction quality.

In the following we first describe (sec. 2) the technique used to estimate the extrinsic camera parameters. Then we present (sec. 3) an analysis of the reconstruction sensitivity to calibration errors and we illustrate some silmulation results. In the fourth section the basic steps of the proposed regularization technique are presented.

Work supported by the Italian Ministry of Education (MPI40%) and the National Research Council.

2. CAMERA CALIBRATION BASED ON MOTION ESTIMATION

The method used to calibrate the spatial position and orientation of the acquisition system consists of estimating the 3D motion leading the TV camera from a viewpoint to the next one. The motion parameters are estimated by means of an algorithm that exploits eight or more corresponding points, on the object under analysis, to build a set of linear equations whose solutions are the so called essential parameters [8] e_i, i=1, ..., 9:

$$A\,\underline{e} = 0 \qquad (1)$$

In eq. (1) A is a matrix of elements derived from the coordinates of the corresponding points and $\underline{e}$ is the vector of the unknowns. From the 3x3 matrix E obtained rearranging the elements of $\underline{e}$, the actual motion parameters can be computed by a singular value decomposition technique or other equivalent methods (see, e.g., [9]).

The matrix E describing the motion of a rigid body is characterized by the following specific algebraic structure [10]: 2 singular values coincide and the third one is zero. If A is affected by noise deriving from errors in the acquired data, the solution for E does not exhibit in general the previous properties. These latter cannot be imposed by explicitely constraining the image coordinates of the corresponding points, since their displacements induced by the varying perspective allow computing the motion parameters. On the other hand, the solutions of Eq. (1) can be constrained by means of two equations expressing the rigidity constraints: letting b_i's, i=1,2,3, be the rows of E, such constraints are

$$F_1(e_1,...e_9) = abs(det(E)) = 0$$

$$F_2(e_1,..e_9) = abs(4[\,||b_1||^2\,||b_2||^2 + ||b_1||^2\,||b_3||^2 + ||b_3||^2\,||b_2||^2 - (b_1\cdot b_2)^2 - (b_1\cdot b_3)^2 - (b_3\cdot b_2)^2] - [||b_1||^2 + ||b_2||^2 + ||b_3||^2]^2) = 0$$

in terms of the norms of b_i and of their scalar products.

The augmented system

$$\begin{cases} A\,\underline{e} = 0 \\ F_1^u(e_1,...e_9)=0 \\ F_2^v(e_1,..e_9)=0 \end{cases}$$

where the parameters u and v allow optimizing the weights of the constraints with respect the basic equations, becomes globally nonlinear and can be solved by an iterative procedure (of the Gauss - Newton kind) starting from a guess solution provided by Eq. (1). A number of simulations have proven that the augmented system yields estimates closer to the real motion parameters than in absence of regularization constraints, especially when few corresponding points are available and at poor resolution.

3. ERROR ANALYSIS

As previously stated, the object reconstruction from multiple views requires that these latter be exactly calibrated with respect to each other. Of course, in the proposed procedure, any error produced in the extrinsic calibration parameters by approximations in the relative motion estimate affects the result of the volumetric and pictorial reconstruction. We have started the investigation of this basic issue from the two following viewpoints:

i) effects of errors in the estimates of relative positions and orientations of the viewpoints on the quality of the volumetric reconstruction;
ii) requirements on the original data to constrain such errors within a range providing an acceptable reconstruction.

The first kind of effects has been analyzed by simulating various error levels in the relative position estimates within the reconstruction procedure. To keep the results fully under control, a synthetic "original" object has been generated, namely a pyramid with a squared basis, that is both simple to manipulate and characterized by features (e.g. vertices) enhancing the reconstruction errors. On the basis of the results obtained in analogous real cases, it has been decided to base the reconstruction on five suitable (synthetic) views: specifically, we have generated the top view and four evenly spaced lateral views of the pyramid. These data have been used first in association with the exact calibration data, to produce a reference reconstructed object, at a resolution of 40x40x100 voxels (along the length, depth and height axes, respectively). The calibration data have been subsequently perturbated by simulated random errors affecting both the coordinates of the optical center and the direction of the optical axis in each view, with respect to a unique reference system. A set of reconstructions have been performed using the original five views in association with the wrong calibration data, and the reconstruction errors have been evaluated in various ways. A simple global measure of the reconstruction defects is represented by the exceeding and missing volumes in the reconstructed object with respect to the reference. The percentage ratio of their sum over the correct figure is shown in Tab. 1 versus the errors in the viewpoint position and in the direction of the optical axis. For every value of ε_t and ε_θ (average percentage errors of the viewpoints coordinates and of the optical axis angles, respectively), the reported

ε_θ \ ε_t	0	1	2	3	5	7
0	-	0.121	0.127	0.142	0.192	0.254
1	0.198	0.201	0.213	0.232	0.285	0.338
2	0.318	0.314	0.317	0.326	0.356	0.399

Table 1

Average percentage volume error versus percentage errors of the viewpoint coordinates, ε_t, and of the angle the fining the camera orientation, ε_θ.

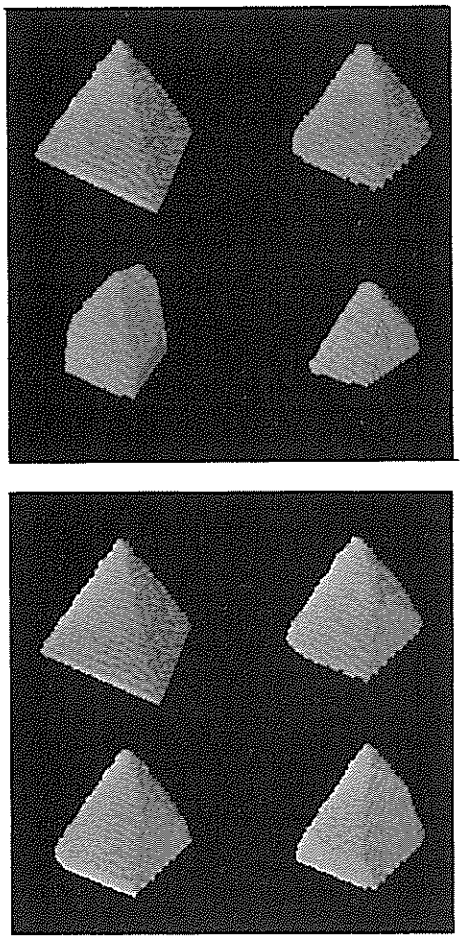


Figure 1

Reconstruction results from simulated imperfectly calibrated views. In each figure, the object reconsructed from perfectly calibrated data is shown at top left. The other reconstructions are obtained from data affected by the following percentage calibration errors: TOP) t.op right ε_t=0, ε_θ=1; bottom left ε_t=0, ε_θ=2; bottom right ε_t=0, ε_θ=3; BOTTOM) t.r. ε_t=1, ε_θ=1; b. l. ε_t=3, ε_θ=1; b.r. ε_t=7, ε_θ=1. The resolution is in all cases of 40x40x100 voxel.

volume error is the average over many simulations, performed by randomly varying the components of the calibration errors. It should be noticed that the actual values of the volumetric error in single simulations may differ significantly (even more than 50%) from the displayed averages, especially for large errors.

A less quantitative but perhaps more immediate evaluation of the reconstruction errors can be visually performed by examining the Fig. 1, that show some reconstruction results obtained from imperfectly calibrated data. These examples are chosen from the simulation runs characterized by error values close to the average reported in Tab. 1

It can be noticed that, in general, the reconstruction results are more haevily affected by the miscalibration of the direction of the optical axis than by the error in the optical center, due to the higher sensitivity of the procedure to the former parameter. Moreover, the results are not acceptable for calibration errors outside the range of 1%, especially on the direction. We have therefore investigated (by extending the study and the simulations reported in [6]) which requirements on the input data to the motion estimation procedure constrain the calibration errors within this range. The relevant parameters are the number of corresponding points, their distribution in space, the kind and amount of 3D motion and the resolution of the original 2D data. We have considered in particular the dependence on the number of correspondences, by choosing an image resolution of 512x512 pixel and amounts of relative motion corresponding to those leading from one view to another in the previously described simulation, and by averaging out the effects of the spatial distribution of the points over many simulations with randomly generated coordinates of the corresponding points. The results are summarized in Fig. 2, where the average errors in the location of the optical center and in the direction of the optical axis are plotted against the number of corresponding points n, in the two cases of linear and nonlinearly constrained estimation. It is apparent from the general error behaviour that, for the chosen values of the other parameters, it is not difficult to constrain ε_t and, more important, ε_θ within the 1% range, by choosing n suitably higher than the minimum value of 8. The results confirm once more that better performances, i.e. lower errors for a given value of n, are usually provided by the technique exploiting the rigidity constraint.

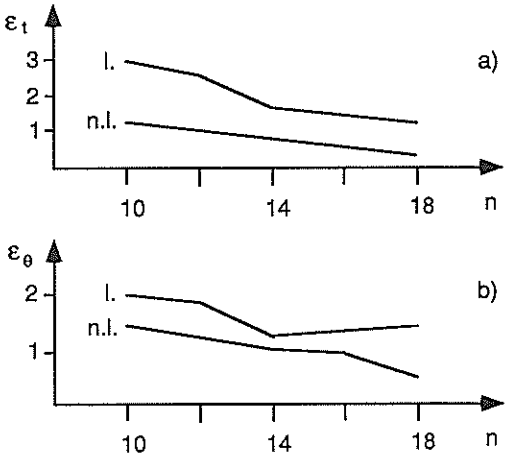


Figure 2

Average percentage errors of the viewpoint coordinates ε_t (a) and of the camera orientation angles ε_θ(b) versus the number n of corresponding points used in the linear (top curves) and nonlinearly constrained (bottom curvers) motion estimation algorithm.

4. RECONSTRUCTION IMPROVEMENT

As we have seen, the basis of the volumetric reconstruction procedure is determining the intersection volume of the generalized cones defined by the projection of the object in the available perspective views. Consider the i-th generalized cone C_i consisting of the infinite straight lines connecting the vertex (the viewpoint) with all the points belonging to the object projection in the corresponding view. The system geometry implies that, for any other cone C_k, $k \neq i$, every line of C_i must have at least a point in common with C_k, i.e. it must intersect at least one of the lines of C_k. In other words, no path exists internal to C_i, leading from its vertex to any point internal to C_i on the other side of the object, and not touching at least one point belonging also to the volume defined by C_k.

Due to the unavoidable errors, in particular on the calibration parameters, such a constraint is not satisfied, in general, for the available data. The most visually noticeable effect of cone "misregistration" is the increasing "erosion" of the reconstructed object when the error increases, as shown in Fig. 1. Our hypothesis is that, if the errors are sufficiently small, imposing the above constraint on the available data reduces the reconstruction error while improving the calibration parameters.

To reduce the problem dimensionality we notice that the errors on the extrinsic parameters are in general greater than those connected with image quantization, object-background segmentation and approximated intrinsic parameter estimation. Assuming then that these latter can be neglected, at least as a first approximation, we seek a regularization procedure to improve the estimate of the six extrinsic parameters of each view according to the previously mentioned constraint. The dimensionality of the problem is still high, involving 6N variables if N is the number of available views, and the equations representative of the 3D geometrical model are not simple, so that we have discarded the analytical formulation of the proposed system, in favor of a numerical algorithm, presently under test, based on the iteration of simple steps to search the minimum of a cost function along the steepest descent direction.

The chosen error function is obtained as follows. Let us consider the generic i-th view V_i representing the object projection in binary form separated from the background, and let us project the viewpoint corresponding to a different view V_k, $k \neq i$, onto the projection plane of V_i. If the projection point falls within (or very close to) V_i, V_k cannot be used in conjunction with V_i, otherwise the cone defined by V_k is projected onto the projection plane of V_i, where an angle is obtained. Two distances, one for each edge of the angle, are defined: if the edge does not intersect V_i, d is the minimum distance between the two, otherwise it is the maximum distance between the edge and the points of V_i outside the angle. The relative error between V_i and V_k is defined as $E_{ik} = d_{ik1} + d_{ik2}$ and the total error with respect to V_i is $E_i' = \sum_k^N E_{ik}$, for $k \neq i$ and not including the views that cannot be used for the condition on the viewpoints. The total error for the complete set of views is $E_T = \sum_{i=1}^{N} E_i'$. The proposed procedure to minimize E_T is the following:

i) choose suitable values for the steps Δ by which each calibration parameter is varied during the search (e.g. as a function of the error range);

ii) compute E_T for the $3^6 \cdot N$ combinations of parameter values obtained by varying the 6 parameters of each view by 0, $+\Delta$ and $-\Delta$, while those of the other views are unchanged (for N=5 this means computing 3645 values of E_T);

iii) modify the parameters corresponding to the minimum value of the E_T so computed and go to ii), provided the error reduction is above a fixed threshold.

The whole procedure can be iterated with decreasing step sizes Δ, until a stop condition (e.g. on the error decrement) is met.

REFERENCES

[1] D.H. Ballard, C.M. Brown, Computer Vision, Prentice-Hall, Englewood Cliffs, New Jersey, 1982.

[2] P.J. Besl, R.C. Jain, "Three-Dimensional Object Recognition", Computing Surveys, vol. 17, pp. 75-145, 1985.

[3] V. Cappellini, R. Casini, M.T. Pareschi, C. Raspollini, "From Multiple Views to Object Recognition", IEEE Trans. Circuits and Systems, vol. CAS-34, pp. 1344-1350, 1987.

[4] R.Y. Tsai, "An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision", Proc. of the IEEE Comp. Soc. Conf. on CVPR, Miami Beach, Florida, pp. 364-374, 1986.

[5] S.T. Barnard, W.P. Thompson, "Disparity Analysis of Images", IEEE Trans. PAMI, vol. PAMI-2, pp. 333-340, 1980.

[6] C. Braccini, G. Gambardella, A. Grattarola, S. Zappatore, "Motion estimation of rigid bodies: effects of the rigidity constraints", in Signal Processing III: Theories and Applications, I. T. Young et al., Eds. New York, NY: North -Holland, pag. 645-648, 1986.

[7] C. Braccini, G. Gambardella, A. Grattarola, M. Milanta, M. Sassoli, "Pictorial Reconstruction of Three-Dimensional Objects through Multiple Views", in J.L. Lacoume et al., Eds., Signal Processing IV: Theories and Applications, North-Holland, Amsterdam, pp. 1461-1464, 1988.

[8] R.Y. Tsai, T.S. Huang, "Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces", IEEE trans. PAMI, Vol. PAMI-6, N° 1, pag. 13-26, 1984.

[9] J. Weng, T.S. Huang, N. Ahuja, "Error analysis of motion parameter estimation from image sequences", in Proc. of First Int. Conf. Comp. Vision, London, pag. 703-707, 1987.

[10] T.S. Huang, O.D. Faugeras, "Some Properties of the E Matrix in Two-View Motion Estimation", IEEE Trans. PAMI, Vol. 11, N° 12, pp. 1310-1312, 1989.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

On a statistical model for moving pictures

Peter Vogel

Philips Kommunikations Industrie AG, Thurn-und-Taxis-Str. 14,
8500 Nürnberg 10, West Germany

Typical Moving pictures for visual telephony are analyzed. The Concept of a spherically invariant random process (SIRP) proves to be very helpful for modelling image signals. This concept, which has already been used for speech coding, is promising for image coding as well.

1. Introduction

In image processing, signal models are mostly restricted to second order moments, e.g. autoregressive random fields [1]. They do not consider multivariate probability density functions (pdfs) of the signal. However, high dimensional pdfs are indispensable for a precise source model, e.g. for coding purposes. Such pdfs arise from a SIRP model, for example, which has already been used for description [2] and coding [3] of speech signals. As will be shown, they are also expedient in the description of the local behaviour of image signals.
The main intention is for improvements to a hybrid DPCM/transform coder for low bit rate video telephony (64kbit/s) standardized by CCITT (Fig. 1). Since this scheme incorporates motion compensated prediction, statistical dependencies in temporal direction are utilized by this scheme. Statistical properties of the remaining prediction error in spatial direction are investigated in the following. Rate distortion functions based on this properties have already been evaluated [4].

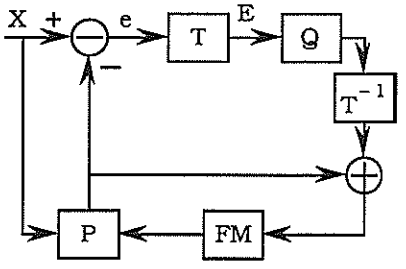
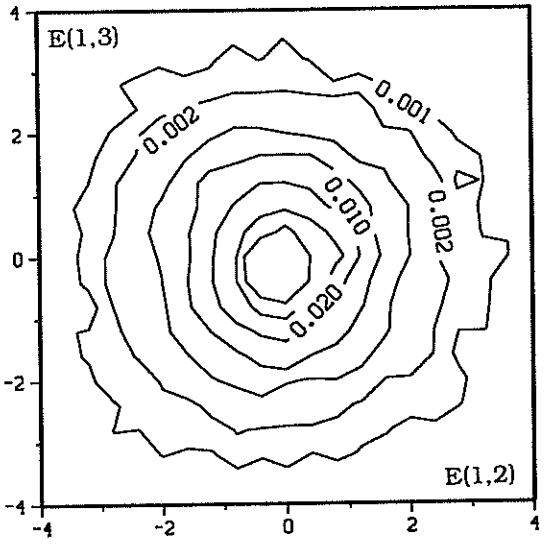


Fig. 1 Hybrid DPCM/transform coder for moving pictures

X Actual frame to be encoded (10 frames per sec, 288x352 luminance pixels per frame)
e Prediction error
E Prediction error after transform
FM: Frame memory for previous reconstructed frame
P: Linear predictor
T: 2-dimensional discrete cosine transform (block size 8x8)
Q: Quantizer

2. Modelling of the prediction error

The prediction error is partitioned into spatial square blocks e of size n=mxm, m=8. These blocks are subjected to the DCT for blockwise decorrelation (Fig. 1). The resulting block of coefficients is denoted by E. For the 2-dimensional pdf of closely neighbouring coefficients concentric ellipses are found for the contour lines (Fig. 2). Due to the normalization of the coefficients to variance 1 they appear as circles. Since the DCT is orthogonal, i.e. performes a rotation of the coordinate axis, the contour lines before the DCT are also ellipses.



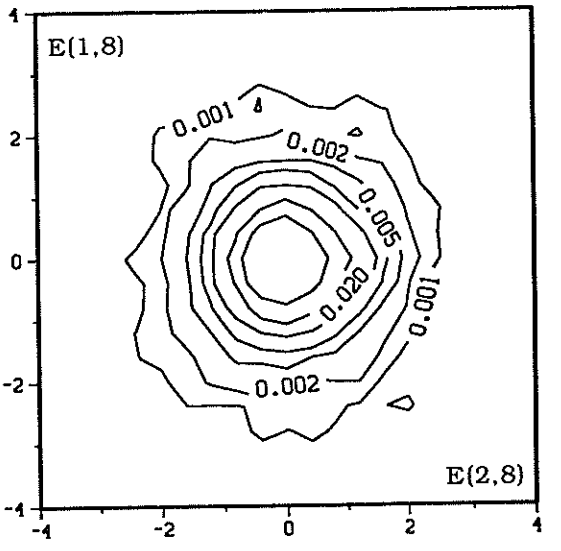


Fig. 2 Contour lines for 2-dimensional pdf of DCT coefficients E(1,2),E(1,3) and E(2,8),E(1,8) normalized to variance 1.

The spherical characteristic indicates vanishing correlation between DCT coefficients. In fact, the decorrelation property was found to be true for all DCT coefficients.
The spherical characteristic holds for a Gaussian pdf, for example. The Gaussian assumption would imply statistical independence between coefficients due to the decorrelation property. Statistical independence between the DCT coefficients is often asserted but is false, as will be outlined.
A contradiction can be obtained by the pdf of

$$(1) \qquad \sigma_n^2(e) \triangleq \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{m} e_{i,j}^2$$

which is the block energy per sample (Fig. 3).

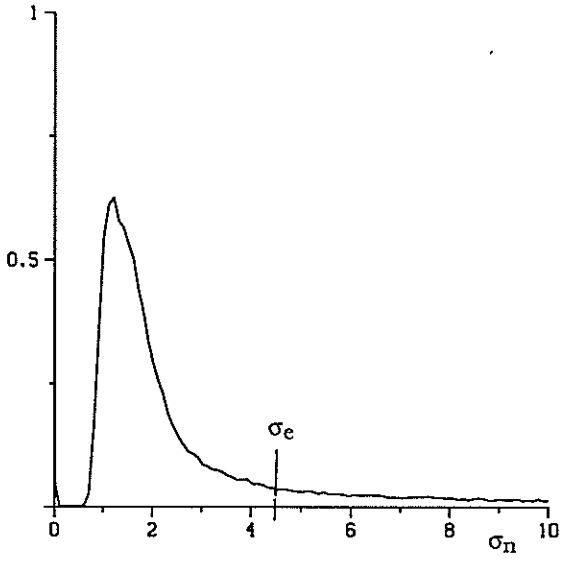


Fig. 3 pdf of block energy per sample
σ_e is the standard deviation of the prediction error process.

In case of statistical independence, most of the probability would lie on the expected value σ_e (Fig. 3). Although σ_e is approximated by σ_n when n tends to infinite, the probability is not concentrated on σ_e. Thus statistical dependence is proven in spite of nearly decorrelated DCT coefficients. Fig. 3 reveals that σ_n is concentrated at a smaller value than σ_e. This shows that large and small coefficients are not mixed in an arbitrary way, but that coefficients with small values often appear together with small coefficients. A physical interpretation of this phenomenon becomes obvious considering e a prediction error block. Since the main part of a picture for visual telephony consists of background and slow motion, many prediction error blocks possess only small energy.

The spherical characteristic of contour lines in Fig. 2 suggests a SIRP model. This model is expedient for statistical dependency expressed by the pdf of the block energy. A SIRP can be characterized by a composite source as follows [2]. Depending on the value of a positive random variable σ_{SIRP}, a Gaussian process with standard deviation σ_{SIRP} and autocorrelation function (ACF) $\sigma_{SIRP}^2\, C(k)$ is switched towards the output of the source. Here C(k) denote the correlation coefficients of the prediction error process. This SIRP characterization proves to be very helpful for generation of SIRPs [2] as well as for coding SIRPs [3]. It implies that the n-dimensional conditional pdfs are Gaussian, i.e.

$$(2) \qquad f_n(\underline{e}\,|\,\sigma_{SIRP}) \overset{SIRP}{=} f_n^G\{\sigma_{SIRP}^2\, C(k)\}(\underline{e}) \quad .$$

In the following determination of σ_{SIRP} is discussed.

For a given value of σ_{SIRP}, e is Gaussian distributed with standard deviation σ_{SIRP} by (2). With respect to this pdf,

$$(3) \qquad \sigma_n(\underline{e})\,\big|\, f_n^G\{\sigma^2\, C(k)\} \approx \sigma$$

holds for n=64 where σ stands for σ_{SIRP}. Here it is assumed that the correlation coefficients C(k) are vanishing for k=64. Consequently, σ_n defined in (1) can be used as an estimate of σ_{SIRP} for n=64 (good approximation of the σ_{SIRP}-pdf by the σ_n-pdf for n=40 is outlined in [3], pp. 134).

For a SIRP (2) holds for all integer n. This yields

$$(4) \qquad \sigma_{SIRP} \overset{SIRP}{=} \lim_{n\to\infty} \sigma_n(\underline{e}) \quad .$$

An experimental evaluation of (4) would be senseless, because this would result in a deterministic value which is σ_e, the standard deviation of the prediction error process. This follows from the ergodicity of the prediction error process, which is assumed during all measurements carried out in this paper. Contrarily, a SIRP is not ergodic because σ_{SIRP} is a random variable (unless the SIRP is a Gaussian process). Thus a SIRP is only used here for modelling the local behaviour of the prediction error process up to n=64 samples.

Replacing σ_{SIRP} in (2) by σ_n yields the estimate

$$(5) \qquad \hat{f}_n(\underline{e}|\sigma_n(\underline{e}) = \sigma) \triangleq f_n^G\{\sigma^2 C(k)\}(\underline{e}) .$$

This estimate is not exact, since (5) does not vanish outside the sphere

$$\sigma_n(\underline{e}) = \sigma$$

as

$$f_n(\underline{e}|\sigma_n)$$

does. However, (5) is concentrated next to this sphere due to (3).

From (5) an estimate

$$(6) \qquad \hat{f}_n(\underline{e}) \triangleq \int f_{\sigma_n}(\sigma) f_n^G\{\sigma^2 C(k)\}(\underline{e})\, d\sigma$$

is also obtained for unconditioned multivariate pdfs. Statistical dependency expressed by the pdf of the block energy is directly incorporated in (6) in terms of this pdf. Both (5) and (6) will be substantiated by experimental results.

By (5), the ACFs under σ_n- condition differ only in a constant factor which agrees with experimental results.

Since the DCT maintains the Gaussian property, by (6), a univariate coefficient pdf is composed of Gaussian pdfs. In particular, the value at position 0 is obtained by integrating

$$f_{\sigma_n}(\sigma) / \sigma.$$

Since small σ_n-values occur (c.m.p. Fig.3), the value at position 0 is high. This agrees with the spiking characteristic of univariate coefficient pdfs (Fig. 4.1).

From (5), it follows that the conditional univariate coefficient pdfs conditioned by σ_n are Gaussian. This agrees with Figs. 4.2-4.

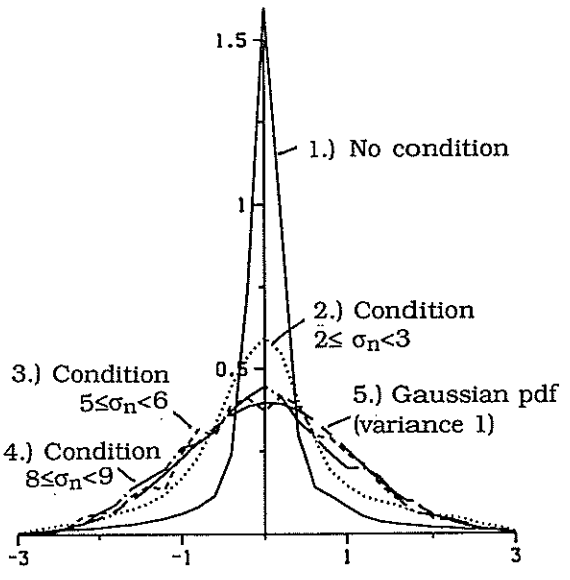


Fig. 4 Pdfs of DCT coefficient E(1,3) normalized to variance 1.

3. Conclusions

The experimental results given in Figs. 2 and 4 suggest modelling of moving pictures with the aid of SIRPs. By (5) and (6), multivariate pdfs of the prediction error are composed of Gaussian pdfs with an identical correlation coefficient structure. As a result, these pdfs are constant on ellipsoids.
In [4] it has been shown that purely horizontal and vertical frequency coefficients occur simultaneously only rarely. This kind of statistical dependency excludes a SIRP model. However, a SIRP model is expedient to express statistical dependency discovered by the pdf of block energy (Fig. 3).

[1] Chellappa,R., Kashyap, R.L., Texture synthesis using 2-D noncausal autoregressive models, IEEE Trans. ASSP (1985), 194-203.

[2] Brehm,H., Stammler,W., Description and generation of spherically invariant speechmodel signals, Signal Processing (1987), 119-141.

[3] Trottler,K., Informationstheoretische Untersuchungen zur Vektorquantisierung von sphärisch invarianten Sprachmodellprozessen, Diss., tech. Fak. der Univ. Erlangen-Nürnberg, 1987.

[4] Vogel,P., On the determination of rate distortion functions for moving pictures,ITG-Fachbericht 107, VDE-Verlag Berlin (1989), 35-41.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Multi-Resolution Image Segmentation In Higher Dimensional Feature Spaces Using Local Transforms

Caspar Horne*
Signal Processing Laboratory
Swiss Federal Institute of Technology
EPFL-Ecublens, CH-1015 Lausanne, Switzerland

Abstract

The problem of using multiple features for unsupervised image segmentation is addressed. The high dimensional feature space that results from many feature extraction processes is reduced in dimensionality in a local manner, resulting in a lower complexity in decision making. The locality is provided by the segmentation method that uses a combined region analysis and region estimation approach. Such an approach allows well developed supervised dimensionality reduction techniques to be used in an unsupervised manner. In this paper the approach is illustrated by using discriminant analysis, which reduces the dimensions of the feature space to a single one, so that the region properties are represented in a local subspace with maximum discriminatory power.

1 Introduction

The partitioning of an image into regions of homogeneous property is an important step in machine vision and consequently has received considerable attention of many researchers over the past two decades. The many different approaches to this problem that have been developed during this time can be divided into two broad categories, namely edge based and region based techniques.

In the edge based techniques the regions are defined by discontinuities established by applying local operators to feature images. The main problem in these techniques is the selection of those discontinuities that correspond to regions boundaries and the rejection of those that are inside regions. To get to such a robustness usually the data is processed over several scales. The problem arises here how to combine the output of the different operators working at different scales. A second problem, one relevant to the work described here, is the design of operators that work not only over a range of different scales, but that work also with high dimensional signals, such as those that are obtained by texture feature extractors. A set of scalar operators can be used working on projections onto one dimensional signals, but the problem of how to combine them arises.

Region based approaches directly work on homogeneity properties to establish the regions. Here the main problem is one of scale, that is, the establishment of the homogeneity of both small and big regions. Recently a set of techniques has evolved that process the data over a range of scales. The most well known techniques are the linked pyramid algorithm of Burt et al. [1], and the quadtree segmentation algorithm of Spann and Wilson [2]. These techniques showed the power of multi resolution processing, but suffer from a lack of generality, the former algorithm requiring knowledge of the number of regions present, the latter requiring assumptions about the minimum region size.

To surmount these difficulties a technique was developed by Spann and Horne [3,4] which is able to surmount these difficulties. Here a strategy is adopted where the regions are simultaneously analyzed and estimated going from a high resolution to a low resolution. The technique uses a local clustering algorithm where the amount of clustering is controlled by the local information present in the analysis window. Such an approach allows the automatic spawning of seeds (compact descriptions of regions) at resolution levels

*The support of Thomson CSF is gratefully acknowledged

corresponding to their natural size, and has been shown to give accurate segmentations, even in extremely low inter-region signal to noise ratios.

The work described in this paper extends this work to images that can only be accurately segmented using high dimensional feature spaces. Examples of such images are satellite images where the features are the spectral bands provided by the satellite, and textured images, where the features are computed by a texture feature extraction process. In such high dimensional spaces the problem arises which features to use. To discriminate neighboring regions it is often advantageous to discard noisy features and take only the representative ones. Clearly this feature set will be different from region to region. Therefor the feature selection is embedded in a segmentation algorithm that provides locality with respect to the regions.

The organization of the paper is as follows. Section 2 describes the segmentation algorithm. Section 3 and section 4 describe the local transforms of the feature space and section 5 shows some results on a three dimensional feature space. Finally section 6 presents conclusions and a discussion.

2 Multiresolution segmentation

The segmentation algorithm is based on a linked pyramid structure [1], which allows regions of any shape and size to be represented and processed in a multiresolution representation. The representations at each level of this pyramid are computed in a non linear manner. Instead of convolving the image with a set of filters that are (band)limited in the frequency domain, and then subsampling the image, the representations at different resolutions are obtained by a simultaneous region analysis and region estimation procedure. The algorithm proceeds from the base of the pyramid, which corresponds to the highest resolution, to the top of the pyramid, thus constructing the pyramid level by level in bottom up fashion.

The algorithm is based on the idea that a region can be represented in a hierarchical structure in such a way that at each resolution in the hierarchy the dimensions of the region are reduced by approximately the same amount as the total image dimensions. Allowing only integer dimensions, regions then disappear at resolutions in the hierarchy corresponding to their original dimensions. In Figure 1 a one dimensional example of such a structure representing two regions, containing respectively three and five pixels, is shown. From this figure it can be seen that a simple structure as the quadtree (which corresponds to a binary tree in 1-D) is not sufficient for representational purposes, and that a more complicated structure as the linked pyramid is needed.

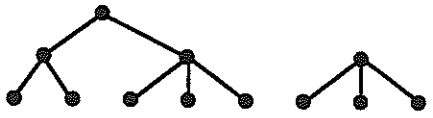


Figure 1: Pyramid structure

Now if the regions are known beforehand, then the complete pyramid can easily be constructed. Once constructed, the complete pyramid representing the whole image consists of a set of sub pyramid structures, each distinct region being represented by its own pyramid structure. The top of each pyramid will then be the node containing all descriptive region information, such as gray level or textural information, while the structure of the pyramid, together with the location of its top node describes all spatial information. If we describe segmentation as "detecting what is where", then the top node contains the "what", and the structure of the pyramid indicates the "where".

In the general segmentation problem, the regions of an image are not known beforehand, and have to be estimated by the segmentation procedure. This is done by a local clustering method. Here the nodes at each level of the pyramid are grouped together, the grouping being controlled by a dynamically estimated local threshold. This procedure makes it possible that nodes that do not belong to a bigger group of nodes, become automatically the top of the pyramid, and thus the "seed" node that represents a region.

3 Multiple features

The use of multiple features as input for the segmentation process results in a much greater complexity in decision making, as compared to the use of only one feature, such as only gray level information. This "curse of dimensionality" often makes it advantageous to work in a lower dimensional feature space. The dimensionality of the feature space can be reduced by using a projection in a lower order space. This lower order subspace has to contain all the relevant information that is necessary for the segmentation process, so that all the regions can be discriminated sufficiently.

A key observation here is that in order to discriminate between the regions they have to differ significantly in properties. Such an observation seems obvious in a one dimensional space where all regions properties can be described in the same feature space. In a

high dimensional feature space, however, region properties can be described in different feature subspaces. Now looking at the image at the region level we can see that each region is best represented in its own subspace, and that it is not only the value of the properties, but also the selection of particular properties that characterizes a region best.

In literature various techniques have been developed for reducing the dimensionality of the feature space to obtain a more manageable problem. Here we propose to use these transforms in a local way, without using a priori knowledge about classes or regions present in the image. Once these transforms are computed the output can be used for a stage of feature reduction or for a local weighting of a distance function, distance being the distance between classes in feature space.

4 Discriminant analysis

The most often used feature reduction method for unsupervised segmentation is the Karhunen-Loeve transform. Although this transform is optimal for minimum error representation, this is not necessarily the transform that gives the highest discriminating power. A more appropriate transform for segmentation purposes can be obtained by using Fisher's linear discriminant functions. These functions are optimal for a large variety of separability criteria. The Fisher's linear discriminant function maximizes the between-class scatter as compared to the within-class scatter.

Following [5] we can denote the samples feature vector by $\mathbf{x}$, then

$$y = \mathbf{w}^t\mathbf{x} \quad \text{with } ||\mathbf{w}|| = 1 \tag{1}$$

is the projection of $\mathbf{x}$ onto one dimension. Suppose now that in the two-class case the two classes $\mathcal{X}_1$ and $\mathcal{X}_2$ are known. In this case a within-class scatter matrix $\mathbf{S}_W$ and a between-class scatter matrix $\mathbf{S}_B$ can be defined as

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 \tag{2}$$

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t \tag{3}$$

where

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\mathcal{X}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \tag{4}$$

The projection vector $\mathbf{w}$ can be computed such that the criterion function

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}} \tag{5}$$

is maximized, resulting in

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{6}$$

In a general segmentation problem the number of classes are not known, and neither are the classes themselves. Fisher's linear discriminant functions are in that case not computable, because the area over which to compute the scatter matrices is unknown. The main problem is thus to estimate these regions, which is the same problem as the segmentation problem.

In the segmentation algorithm described previously in section 2, the region analysis is performed simultaneously with the region estimation procedure. At each moment of the algorithm each node represents one region. This means that at each moment and at each node of the pyramid the information present is restricted to one class. Thus at each pair of nodes the problem is reduced from a general n-class problem, with the classes and their number unknown, to a two-class problem, with the two classes known. The scatter matrices $\mathbf{S}_W$ and $\mathbf{S}_B$ from Equations 2 and 3 can then easily be computed.

5 Results

To illustrate the method we constructed a testimage of size 256 * 256 pixels consisting of two natural textures, shown in Figure 2 on the left. The textures were taken from the Brodatz album, D29 "beach sand" and D9 "grass lawn". The textures were arranged according to the maskimage shown in the same figure, on the right.

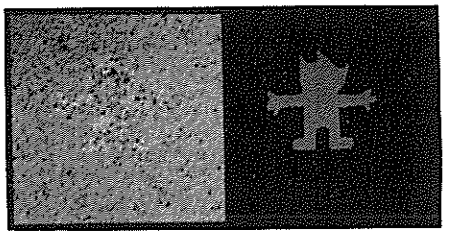


Figure 2: Testimage (left) and maskimage (right)

From this textured image three texture features were computed. The features are computed by convolving local masks with the image. The masks are three masks of the Unser feature set [6] and are approximations of horizontal, diagonal, and vertical edge detectors. In Figure 3 the three feature images are shown, together with the segmentation result. It can

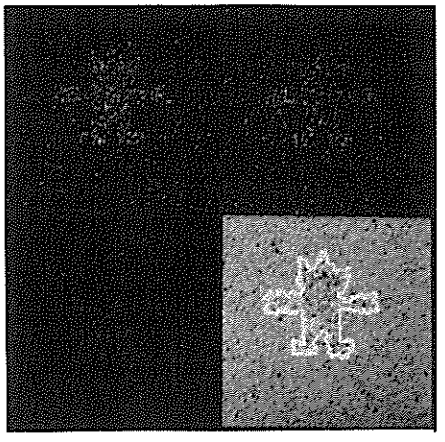


Figure 3: Texture features, and segmentation result

be seen that all three feature images contain information, so that the problem is really three dimensional, and that the information is very noisy.

The segmentation result shows the pixels having one or more neighbors belonging to a different region in white, superimposed on the original textured image. Due to the noise in the features, the boundaries are somewhat irregular, and have a somewhat blocky appearance. This blocky appearance is mainly caused by the way the boundary regions are displayed, the white boundary region being more than one pixel large.

The segmentation algorithm has found the two regions, corresponding to the two regions of the maskimage. Considering the low inter region signal to noise ratio, the boundaries are very well placed, and the overall result is excellent.

6 Conclusion

A segmentation algorithm has been presented that partitions images into homogeneous regions using high dimensional feature spaces. Due to the locality with respect to the regions local transforms of the feature space can be used. This locality is provided by the segmentation algorithm, which uses a simultaneous region analysis and region estimation process. Using local transforms of the feature space gives the advantage that features can be selected that are optimal with respect to the regions, and that can differ from one region to another. The method has been tested using discriminant analysis techniques and results have been presented.

It is anticipated that the method can used on the basis of other local transforms as well, such as the Karhunen Loeve transform, which is optimal for representational purposes. A further possibility will be the use of features that are computed on a multiresolution basis. In that case a different number of features can be used at each level of the pyramid, and each feature can be used at a level corresponding to the resolution of the feature extraction process. It is intended to report on progress of these topics in the future.

References

[1] P.J. Burt, T.H. Hong, and A. Rosenfeld. Segmentation and estimation of region properties through co-operative hierarchical compution. *IEEE Trans. Sys. Man. and Cyb.*, SMC-11:802–809, 1981.

[2] M. Spann and R. Wilson. A quad-tree approach to image segmentation. *Pattern Recognition*, 18:257–269, 1985.

[3] C. Horne and M. Spann. Region extraction using a dynamic thresholding pyramid. In *Proc. of the SPIE Conf. on Visual Comm. and Image Processing '88, Vol. SPIE-1001*, Cambridge, MA, USA, November 1988.

[4] M. Spann and C. Horne. Image segmentation using a dynamic thresholding pyramid. *Pattern Recognition*, 22:719–732, 1989.

[5] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* Wiley-Interscience, New York, 1973.

[6] M. Unser. Local linear transforms for texture measurements. *Signal Processing*, 11:61–79, 1986.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Texture Boundary Detection Based on LVQ Method

Ari Visa
Helsinki University of Technology
Laboratory of Information and Computer Science
Rakentajanaukio 2 C, SF-02150 Espoo, Finland

This paper concerns images containing stochastic textures. A new image segmentation method is shortly described. It's behaviour is studied at region boundaries. The power of the method is demonstrated on realistic stochastic textures. The suggested method is based on multiresolution representation of co-occurrence matrices, feature maps and Learning Vector Quantization (LVQ). The edge detection is achieved by region recognition. Each region is assumed to consist of unique texture type. The reported results are promising.

1 Introduction

The importance of texture recognition is increasing within the image processing field. There are four main ares of image processing in which texture plays an important role: classification, image segmentation, realism in computer graphics and image coding. Classification and segmentation can also be interpreted as a part of image analysis. All these fields imply that texture has many practical applications. There are three main problems in the texture analysis: 1) How can a textured region be described? 2) Given a textured region, to which of a finite number of classes does the sample belong? 3) Given a scene, how can the boundaries between the major textured regions be established? A proposition is made to the third question but the other questions are also discussed. First a definition to texture is given. Texture could be defined as a structure composed of a large number of more or less ordered similar elements or patterns without one of these drawing special attention. So a global unitary impression is offered to the observer. A texture can be a strictly ordered array of identical subpatterns, for instance a chessboard [Br 68]. Such a texture is called deterministic and it can be described by the characteristics of one subpattern or primitive and by the placement rules defining the spatial distribution of the primitives. The impression of a pattern can also obey some statistical laws. The resulting structure might resemble noise on a television screen [Br 68]. Such a texture is said to be stochastic. One should notice that deterministic textures can be heavily disturbed in their repetition and their primitives might be similar, but not at all identical.
First there is the problem with texture description. Several methods to describe texture are known [Va85]. However, stochastic textures are different in character from other ones. For instance syntactical methods are not successful on stochastic textures. To classify stochastic textures the gray tone co-occurrence matrix method (GTCM) [Ha 73] has been found to be superior both theoretically [Co 80] and empirically [We 76]. The reported disadvantages of the GTCM method are discarding shape information in image and fixed resolution.The first one is not a drawback when stochastic textures are considered. The second one can be alleviated by using multiresolution representation.
The segmentation of a textured image is not a new problem. There have been several attempts to solve the problem. Zucker and Rosenfeld [Zu 75] made segmentation based on local picture properties and histograms. Deguchi [De 78] developed a two dimensional linear estimation technique to characterize and segment textured images. Conners [Co 84] and Raafat [Ra 88] have also studied the problem. Conners has used cooccurrence matrices and texture measures derived from co-occurrence matrices but he used a statistical segmentation. The segmentation procedure consideres three kind of regions at each level of the segmentation: uniform, boundary and unspecified. At every level the procedures differentiated uniform regions from boundary and unspecified regions. Raafat introduced the idea of texture distance measure for classifying textures and directing the region growing process. There is, however, the learning problem. Attempts to apply neural networks on image compression [Lu 89], edge detection [Fu 88] and texture recognition [La 89] have lately reported.

2 TEXTURE SEGMENTATION

The texture boundaries are detected by a segmentation procedure. A direct approach to edge detection is difficult because edges have a different meaning in texture context than in ordinary images. The idea of the method is to use co-occurrence matrices in a multiresolution way. The co-occurrence matrices are considered as feature vectors. The learning samples define a feature map. The teaching of the feature map is accomplished by neural networks methods. The classification is done in feature space. The classification is based on distances.
A GTCM is computed in given direction and distance. The computation is done in three distances. The GTCM is computed as follows : Let $f(x,y)$ be a picture with $x \in 1,2,...NX, y \in 1,2,...NY$, where NX, NY are the

dimension of the picture in the $x-$ and $y-$ directions, respectively. The picture is digitized to NG gray levels, $G = 0, 1, ...(NG-1)$. Let D be a set of displacement vectors,

$D = \{d | d = (dx, dy),$
$0 < dx < NX,$
$-NY < dy < NY\}.$

Thus, $M_d(i,j)$ is defined to be the matrix whose (i,j)th element is the number of times that gray levels i and j occur in the relative position d. The entries in the GTCM, $M_d(i,j)$ of f with respect to a particular $d = (a,b)$ are defined as follows:

$$M_d(i,j) = \#\{((x1,y1),(x2,y2)) : \quad (1)$$

$f(x1,y1) = i, f(x2,y2) = j$, and $d = (a,b), d \in D$, such that $x2 = x1 + a$ and $y2 = y1 + b\}$ where $\#$ denotes the number of elements in the set.

A triplet $(M_{d1}(i,j), M_{d2}(i,j), M_{d3}(i,j))$ is extracted and considered as a pattern vector. To reduce the memory demand the co-occurrence matrices are usually replaced by some measures. In this case so is not done. In stead the co-occurrence matrices $(M_{d1}(i,j), M_{d2}(i,j), M_{d3}(i,j))$ are requantized and used as feature vectors.
The texture segmentation is executed partly in classical way partly with neural netwoks methods. A Feature map of N reference vectors are chosen [Ko 83]. The topology of the feature map doesn't matter. The distance metric can be calculated in many ways. In this case Euclidean distance is employed. N should be large enough that each texture class contains sufficient number of reference vectors. Each texture class corresponds with a texture region. The reference vectors are taught to each class in supervised learning mode. Suitable areas of image are selected and taught to the segmentation procedure. More samples, variations of reference samples, are further selected and taught to the segmentation procedure. Some of extracted feature vectors are used directly as reference vectors and the others are used to fine tune the feature map. The learning process is realized by learning vector quantization (LVQ) [Ko 86]. The LVQ method consists of two parts: First the closest reference vector m_c is localized on the map. Secondly the reference vector is fine tuned. The closest reference vector m_c is located by nearest neighbour method. N distances $D_j = ||x - m_j||$ are calculated to localize m_c to a feature vector x.

The m_c is minimum of

$$D_j = ||x - m_j||. \quad (2)$$

During the fine tuning phase labelled samples and a feature map of N elements are needed. The exact form of the fine tuning part is:

$$m_c(t+1) = m_c(t) + \alpha(t) * (x(t) - m_c(t)) \quad (3)$$

if $x(t)$ and the closest unit $m_c(t)$ belong to the same class,

$$m_c(t+1) = m_c(t) - \alpha(t) * (x(t) - m_c(t)) \quad (4)$$

if $x(t)$ and the closest unit $m_c(t)$ belong to different class,

$$m_i(t+1) = m_i(t) \quad (5)$$

for $i \neq c$,
where function $alpha(t)$ is decreasing function with properties:

$\sum_{t=-\infty}^{\infty} \alpha(t) = \infty, \sum_{t=-\infty}^{\infty} \alpha(t)^2 < \infty.$

The feature map determines how many regions can be found in an image. Depending on the number of reference vectors in each class it is possible to choose some reference vectors to the inner regions and the others to the boundaries.
During segmentation process the LVQ method works in a sense of nearest neighbour classifier. The image is processed in raster scan manner from upper left to lower right corner. A $n * n$ window glides over the image. The feature vectors are calculated within the window. The extracted feature vector is compared with the feature map. The classification is given by the feature map. The classification results in reference vectors on the feature map. The reference vectors belong to a class. The classmembership is transferedinto the image. All the classmemberships create the segmentation. Besides the segmentation it is possible to get more detailed information of the regions because there are more reference vectors than classes.
Classification in feature space tends to result in small regions with uncertain and rugged boundaries. This depends on the feature selection and on the size of window. To improve segmentation and region boundaries either the features or the window size had to be adjusted. There is, however, another solution too. One can adjust decision surfaces in feature space by teaching new samples. The LVQ method is capable to fine tune the class boundaries or decision surfaces in feature space. This will result in smoother region boundaries in image.

3 EXPERIMENTS AND RESULTS

The method has been first tested on simulated images. The image size has been 512*480 pixels and the whole gray level range 0 .. 255 has been used. The number of reference vectors N on the map has been 16. The feature map has been taught further by 116 samples that have been extracted from the image. The stochastic textures have been generated by published methods [Cr 83].
One test image containing four textures is shown, Figure 1. The image consists of four regions that have uniform, logarithmic gaussian, gaussian and logarithmic gaussian distributions. Grain size is less than 5 pixels.
The image has been segmented by the given segmentation procedure. The window size has been $50 * 50$ pixels. The window size has been selected due to resolution and due to representativeness of the sample. The detailed segmented image is shown in Figure 2. It can be seen that the ex-

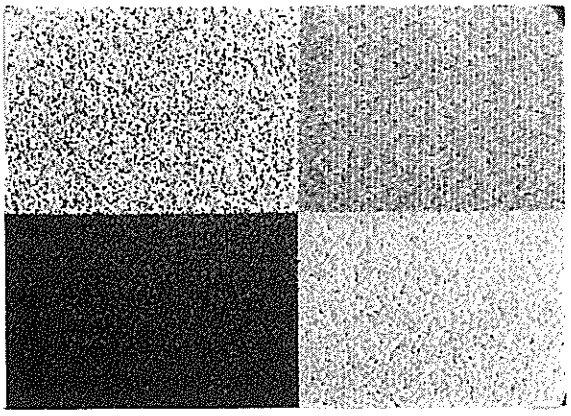


Figure 1: Test image consisting of four generated textures

pected regions have been detected relatively well. So has been the case with boundaries, too. The stochastic texture with uniform distribution has been hard. The place of the texture boundary is quite correct. There is an uncertain region of half window size around the correct border. This uncertain region can be decreased by smaller windows.

To test the suggested method on real textures some radiographs of paper formation have been captured. An image containing four captured stochastic textures have been composed, Figure 3. The four stochastic textures have been defined by a human observer.

The image containing real stochastic textures is more vivid. This can be noticed considering the segmented image, Figure 4. The regions have been detected. They are not perfect. The problem has been the inhomogeneity of real samples. The feature selection might also play a role. If regions are distinguishable considering certain features boundary detection works quite well.

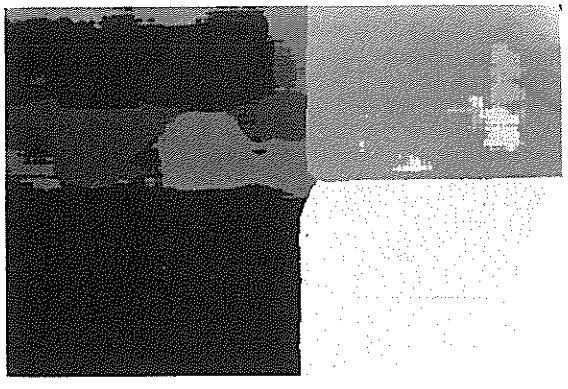


Figure 2: Generated test image after segmentation

4 DISCUSSION AND CONCLUSIONS

The reported results are mainly based on artificial images. The reason is that it is otherwise difficult to determine the exact position of region boundary. The real stochastic tex-

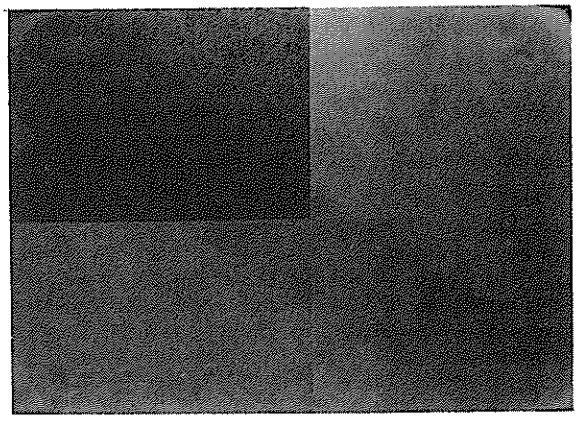


Figure 3: Test image containing four real stochastic textures

Figure 4: Test image containing four real textures after segmentation

tures are seldom homogeneous. The comparison between the detected and the known boundary is then difficult. The capability of the new method has been shown on realistic images, too. The problem with real stochastic textures is texture description. The problem is also known as feature selection problem. First the resolution of the actual texture should be chosen so that the recognition is meaningful. Then the size of the actual window should be chosen so that the extracted features differ from each other when it is necessary. This should be weighted against need of memory, computing power and uncertainty of boundary localization.

The selected window scans over the whole image which means a large amount of computations. The reported results are based on four regions. Images with eight and sixteen regions have been tested but they have needed many hours to run on a 386 based SUN workstation. This problem can be alleviated by parallel processing for which the suggested method is suitable. The problem how estimate the number of regions is not discussed here. It is, however, possible [Vi90]. The size of the feature map will determine the upper limit of detectable regions.
Several stochastic textures are taught before hand and a map is created. This map is used to classification later. The uncertainty at region boundaries can be partly controlled by the LVQ method and further teaching. The remaining problem is the size of the learning set and speed of the convergence.
To summarize a new method to segment textured images is represented. The definition and detection of region boundaries are done by segmentation. The quality of boundary detection can be controlled by a fine tuned learning procedure. The results are also promising when the method is applied to real stochastic textures.

Acknowledgement The author wishes to thank Professors T. Kohonen and O. Simula for support.

References

[Br 68] Brodatz, P., Textures: A Photographic Album for Artists and Designers, Reinhold, New York, 1968.

[Co 80] Conners, R., Harlow, C., A Theoretical Comparison of Texture Algorithms, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No 3, 1980, 204-222.

[Co 84] Conners, R., Trivedi, M., Harlow, C., Segmentation of a High-Resolution Urban Scene Using Texture Operators, Computer Vision, Graphics, and Image Processing, 25, 1984, 273-310.

[Cr 83] Cross, G., Jain, A., Markov Random Field Texture Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5 ,No 1, January, 1983, 25-39.

[De 78] Deguchi, K., Morishita, I., Texture Characterization and Texture-Based Image Partitioning Using Two-Dimensional Linear Estimation Technique, IEEE Transactions on Computers, Vol C-27, No. 8, 1978, 739-745.

[Fu 88] Fukushima, K., A neural network for visual pattern recognition, IEEE Computer, 31(3),March, 1988, 65-75.

[Ha 73] Haralick, R., Shanmugam, K., Dinstein, I., Textural Features for Image Classification, IEEE Trans. on System, Man, and Cybernetics, Vol. SMC-3, No.6, November, 73, 610-621.

[Ko 83] Kohonen, T., Self-Organization and Associative Memory, Springer-Verlag, Berlin, Heidelberg, New York, Tokio, 1983.

[Ko 86] Kohonen, T., Learning Vector Quantization for Pattern Recognition, Helsinki University of Technology, Report TKK-F-A601, 1986.

[La 89] Lampinen,J., Oja, E., Self-Organizing Maps for Spatial and temporal AR models,Proceedings of The 6th Scandinavian Conference on Image Analysis, Oulu, Finland,June 19-22, 1989, 120-127.

[Lu 89] Luttrell, S.P.,Image compression using a multilayer neural network, Pattern Recognition Letters, 10, July, 1989, 1-7.

[Ra 88] Raafat, H., A Texture Information-Directed Region Growing Algorithm for Image Segmentation and Region Classification, Computer Vision, Graphics, and Image Processing, 43, 1988, 1-27.

[We 76] Weszka, J., Dryer, C., Rosenfeld, A., A Comparative Study of Texture Measures for Terrain Classification, IEEE Trans. on System, Man, and Cybernetics, Vol. SMC-6, No. 4, April, 1976, 269-285.

[Zu 75] Zucker, S., Rosenfeld, A., Picture Segmentation by Texture Discrimination,IEEE Trans. on Computers, C-24, No. 12, 1975, 1228-1233.

[Va85] Van Gool, L., Dewaele, P., Ooterlink, A., Survey Texture Analysis Anno 1983, Computer Vision, Graphics, and Image Processing, Vol. 29, 1985, 336-357.

[Vi90] Visa, A., Identification of Stochastic Textures with Multiresolution Features and Self-Organizing Maps, in 10th International Conference on Pattern Recognition, IEEE Computer Society Press, 1990,in print.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A MODEL BASED IMAGE SEGMENTATION METHOD

Anu Langinmaa

Technical Research Centre of Finland, Graphic Arts Laboratory
Tekniikant. 3, SF-02150 Espoo
tel. Int + 358 0 4561, telefax Int. + 358 0 463848

A model based image segmentation method has been developed to segment a contact image of paper. When the segmented image is further processed the contact area size distribution of the even areas of the paper is obtained. This distribution can be used to estimate the printability of the paper.

1. Introduction

Paper roughness is defined as small scale variation of paper surface (Fig. 2). It is a structure property of paper which is of importance especially when paper is printed using the gravure printing method. In the printing press the paper is pressed against the cylinder in the print nip and it is important that the pressing surface reaches the paper surface, otherwise the print quality suffers. From the practical point of view it is thus essential to measure the paper roughness under pressure which simulates the real print circumstances, i.e., what happens in the print nip. Various measurement devices for this purpose have been developed [2]. One lack of the existing measurement devices is that the results are not processed in the signal processing sense. It is, e.g., not possible to get information about the size distribution of the even areas which is an interesting issue in the printing point of view.

2. The problem

The problem was to segment automatically a contact area image of paper surface. The image was acquired using a FOGRA/KAM surface roughness measuring device [3] shown in Fig. 1. The operation principle of the device is as follows. The specimen is pressed against a prism using a pressure which corresponds with the pressures used in printing presses. The specimen is illuminated at an angle which is greater than the angle of total reflection. When the prism is in optical contact with the paper (paper is even) diffuse reflection occurs. When there is no contact total reflection occurs. Light cell measures the amount of optical contact between the prism and the specimen and gives a reading which tells the contact percentage. The evener the paper the greater the reading. When the prism is looked from above a contact area image is seen where the even areas are light and rough areas dark. In this work a contact area image was acquired from above using a video camera and this image was further processed.

The image was digitized into 512x512 pixels. In Fig. 3 a pressure image of a newspaper specimen is shown. The area is 0.9x0.9 cm2. So one pixel corresponds to about 20 um. The dark areas which represent the rough areas are referred to as background areas and white areas as contact areas. The con-

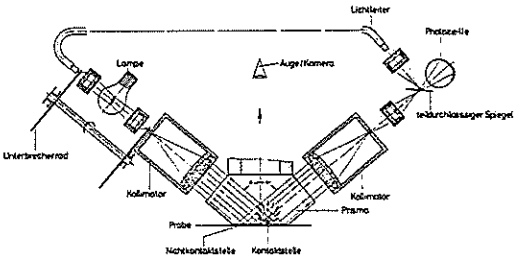


Fig. 1 FOGRA/KAM roughness measurement device

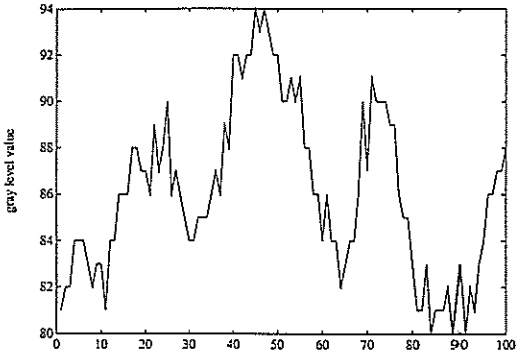


Fig. 2 Cross-section of newspaper, pressure 10 kP/cm2

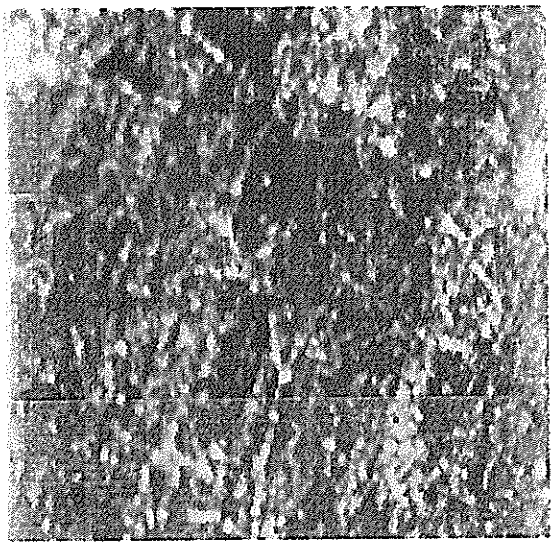


Fig. 3 Newspaper I, pressure 10 kP/cm^2

trast of the image is not especially good, at most 30 gray level values as can be seen from the cross section shown in Fig. 2. By higher pressure and/or more even paper type the percentage of white areas as well as the overall gray level value of the image increases.

3. The suggested method

In Fig. 3 and 7 two typical examples of an original image are shown, in Fig. 3 newspaper and in Fig. 7 super calendered (sc) gravure paper. The applied pressure is 10 kP/cm^2. The images have been equalized to improve hardcopy quality.

The starting-point of the work was the histogram of the image. In Fig. 5 and 6 we see histograms computed for newspaper and sc paper samples. In each image several histograms are shown. In Table 1 we see four characteristic values computed from the histogram of a typical sample. The chosen characteristic values are mode, average, median and standard deviation.

Characteristic value	Newspaper I pressure, kP/cm^2 10	50	SC paper I pressure, kP/cm^2 10	50
average	84,7	102,7	81,2	106.0
median	84	84	80	104
mode	82	98	78	98
std deviation	5.2	8.8	6.0	17.1

Table 1. Characteristic values of Newspaper I and SC paper I

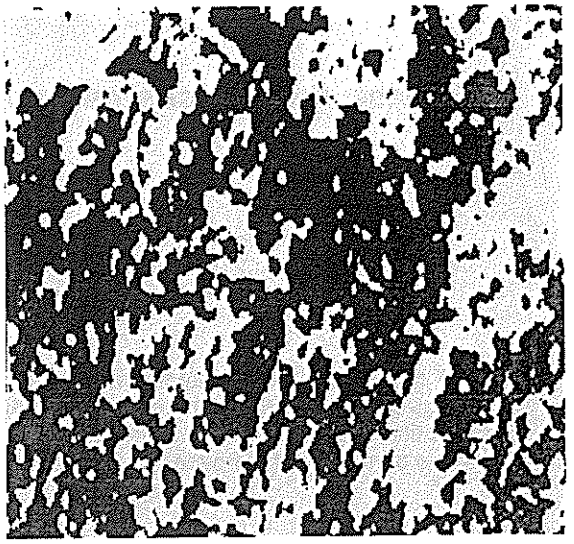


Fig. 4 Segmented newspaper I, pressure 10 kP/cm^2

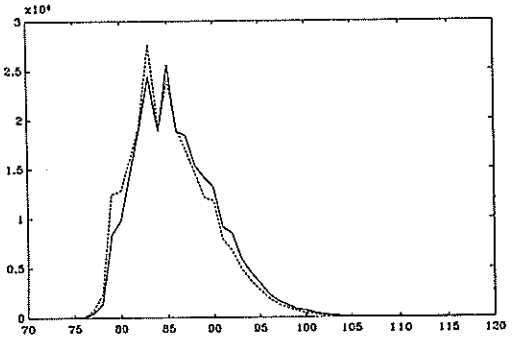


Fig. 5 Histogram of newspaper, pressure 10 kP/cm^2, 3 samples

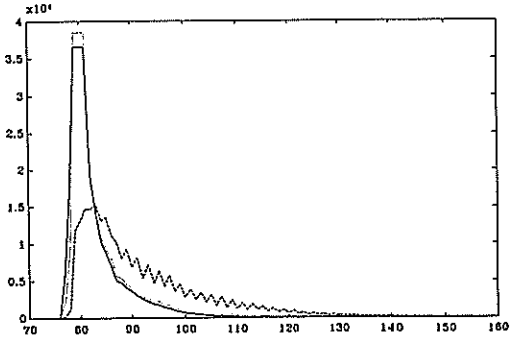


Fig. 6 Histogram of SC-paper, pressure 10 kP/cm^2, 3 samples

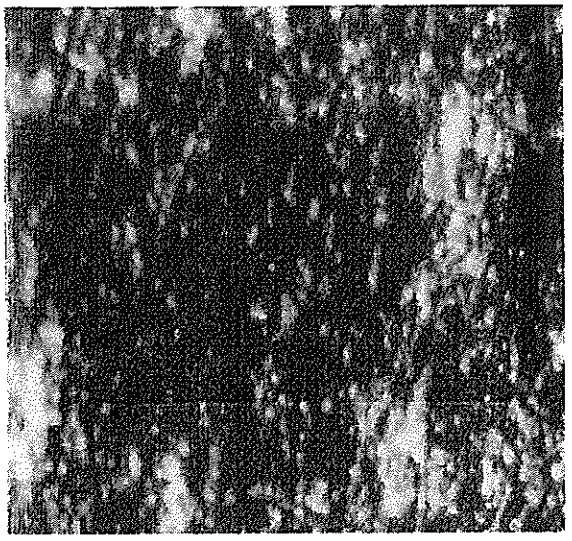


Fig. 7 SC paper I, pressure 10 kP/cm^2

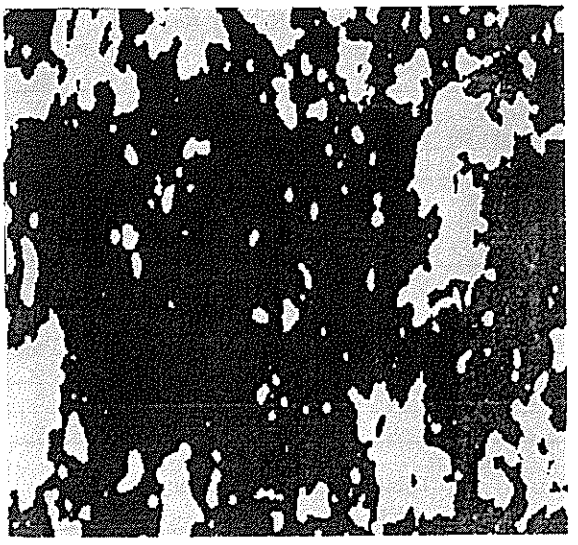


Fig. 8 Segmented SC paper I, pressure 10 kP/cm^2

Several paper types were tested and it was noticed that every paper type has a very characteristic histogram which is approximately the same for all samples. The characteristic values describe the shape of the histogram. If average, median and mode are about the same the distribution is about gaussian. If mode is much greater (smaller) than the average value the distribution is a on the right (left) tailing one. Standard deviation characterizes the size and dynamics of the contact areas. Standard deviation increases when the paper gets more even or the applied pressure is higher.

A histogram tailing on the right (Fig. 5) can be modelled as a sum of a gaussian distributed background area and contact area. In the case of a very even paper (e.g. art paper) under great pressure the tail is on the left and represents the background of the paper.

The aim of the work was to extract the contact areas from the image. Because of speed requirements thresholding on the basis of the image histogram was an appealing alternative. Unfortunately the image histograms are not bimodal. Thus standard threshold selection methods [4,5] which select a valley between two peaks in the histogram could not be applied. Some experiments to estimate the underlying background and contact area distributions to determine the threshold value using their characteristic values were carried out. However, it turned out that it was impossible to estimate the distributions reliably enough. Thus some other method had to be applied.

The solution was to suppose that the background distribution is about gaussian and estimate its' mean by the mode of the histogram. The standard deviation of the background area was estimated using the whole histogram. Thus the threshold value is

$$\text{threshold} = \text{mode} + x^*(\text{standard deviation})$$

where x is a confidence interval coefficient which depends on paper type and pressure.

The hypothesis was tested by first thresholding a small amount of samples of each specimen by hand and determining the appropriate confidence interval coefficients. It turned out that in practice x was 0.0-1.0. The value of x for a given paper type and pressure was about constant. Then totally 158 samples were thresholded using the acquired confidence interval values and the results were compared to the result of thresholding by hand. The samples consisted of newspaper (two types), sc gravure paper (two types) and light weight coated paper (LWC paper). Each automatic thresholding was a given a mark which corresponds to its' quality. The results are shown in Tables 2, 3, 4 and 5.

In Fig. 4 and 8 we see the newspaper and sc gravure paper images of Fig. 3 and 7 thresholded using the suggested method. The thresholded images have been lowpass filtered using a 3x3 median filter to get rid of insignificant details. The region size distribution of the even areas in Fig. 4 and 8 are shown in Fig. 9 and 10. The region size distribution has been obtained by labelling the thresholded and median filtered image and then computing the size of each region in "square pixels".

quality	Newspaper I, N=39 pressure, kP/cm^2			Newspaper II, N=24 pressure, kP/cm^2		
	10	15	50	10	15	50
good	9	7	6	11	9	0
acceptable	4	6	6	1	3	0
bad	0	0	1	0	0	0

Table 2. Newspapers I and II

quality	SC paper I, N=26 pressure, kP/cm^2			SC paper II, N=36 pressure, kP/cm^2		
	10	15	50	10	15	50
good	10	10	0	15	10	0
acceptable	3	1	0	3	8	0
bad	0	2	0	0	0	0

Table 3. SC papers I and II

quality	pressure, kP/cm^2		
	10	15	50
good	8	7	5
acceptable	3	3	4
bad	0	1	2

Table 4. LWC paper, N = 33

quality	pressure, kP/cm^2		
	10	15	50
good	53	43	11
acceptable	14	21	10
bad	0	3	3

Table 5. Total, N = 158

4. Discussion and conclusions

Model based image segmentation techniques have been applied to contact area extraction of paper.

The image acquiring and digitizing was performed using a DT2851 image processing card. The computations were carried out using a ALR/386 computer equipped with mathematical coprocessor.

The method has been applied to characterize newspaper, sc gravure paper and LWC paper. The applied pressures have been 10, 15 and 50 kP/cm^2.

The method turned out to be fast. It is also reliable enough for newspaper . It does, however, not always work correctly when applied to sc gravure and LWC paper. The obtained contact area size distribution is of importance when the printability of the paper is concerned.

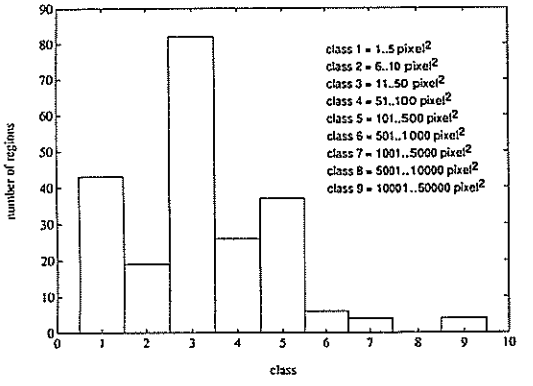


Fig. 9 Contact area size distribution of Fig. 4

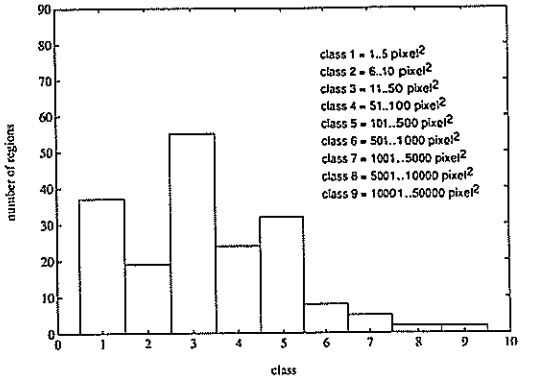


Fig. 10 Contact area size distribution of Fig. 8

5. Acknowledgments

This work has been funded by the Technology Development Centre of Finland.

6. References

[1] J.A. Bristow and P. Kolseth, Paper Structure and Properties, Marcel Dekker Inc. 1986.

[2] W.I.Wilt, Paper testers for instaneous measurement of smoothness and porosity. Tappi 39, 1956.

[3] FOGRA/KAM, Kontaktanteilmessgerät. Operating instructions.

[4] M. Haralick and L.G. Shapiro, Image Segmentation Techniques, Computer Vision, Graphics and Image Processing 29, 1985.

[5] J.S. Weszka, A Survey of Threshold Techniques. Computer Graphics and Image Processing 7, 1978.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

STUDY OF STONES BY IMAGE PROCESSING

R. Harba*, G. Jacquet* and M. Rautureau**

* Groupe de Recherche sur l'Etude des Milieux Ionisés
** Laboratoire de cristallographie

U.F.R. faculté des sciences
rue de Chartes BP 6759
45067 Orléans CEDEX 2 FRANCE

ABSTRACT :

This paper describes a method to extract two basic components of a stone from three images of the same scene made by scanning electron microscopy. This is realized by taking into account prior information that is to say that these three images would be superposable as a jigsaw puzzle for an ideal case. So, a discriminating fonction D(T) which represents the reliability of the superposition is defined versus a vector T whose components are the thresholds of the three images. It is planned to minimize D(T) in order to yield the best vector T. Our results are compared with classical ones and also with extraction made by an expert.

INTRODUCTION :

Up to now, the primary expertise of historical monuments is essentially visual, and it cannot be either reproduced or even quantified. So, the failure to recognize some parameters related to composition, porosity and anisotropy of stones does not permit a correct evaluation of their growing old in order to decide when, why and wherefore to restaure it.

A very precise knowledge of stones is necessary to evaluate some important parameters. For example, silice particles agglomerate to calcite ones and compose the cement. Therefore, silice proportion reflects the strength of the stone. Chemical and physical reactions of the stone to water depend on the quantity and the shape of porousness.

Quantitative studies are made. For example, a chemical analysis gives the proportions of the various components. Another solution is to observe a thin section of stone by electronic microscopy [1] : a spectral analysis also gives these proportions. But, only an image processing leads to information on the shape and on the relative position of the components. The present study reports preliminary investigations of such a processing on very porous stones (about fifty per cent of porousness) called tuffeau often used in the past for building monuments which are mostly composed of siliceous and calcareous particules.

MICROSCOPIC STONE IMAGES :

Samples are made of thin sections of stones of a few square centimeters large and thirty micrometers thick, where porousness is filled up by synthetic resin : induration is necessary because the tuffeau is a polyphased material with little coherence. Such petrographic sections are observed by scanning electron microscopy, yielding a polaroïd image (fig. 1a) in which the secondary and backscattered electrons are analysed. This good quality image is called porousness image and is composed of a lot of particles :

i) the very black ones are calcite, little calcareous fossils and siliceous particles which are quartz.

ii) the medium grey ones are quartz or mica principaly made of silicium.

iii) the small and clearer ones are silice or clay (siliceous particle).

iv) A very few of them (about one per cent) are iron oxydes or iron sulfurs.

v) porousness is the white parts of the image and is easilly separable from the particles.

It seems difficult to separate all these particles by only taking into account the porousness image. In spite of this, it can be expected from the histogram to be composed of four modes. But acquisition problems do not allow such a perfect case (fig. 1b). As seen, only the white distribution which is porousness is separated from the three others. A correct

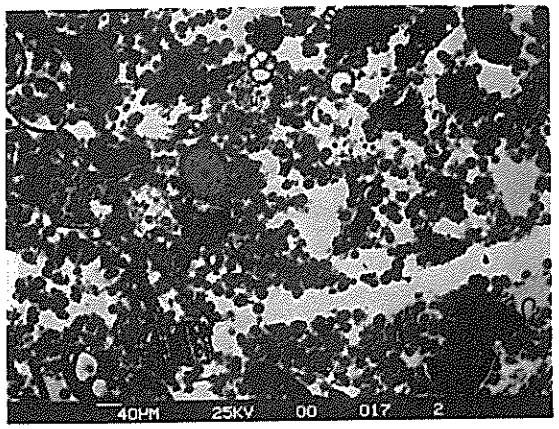


figure 1a : Porousness image and ...

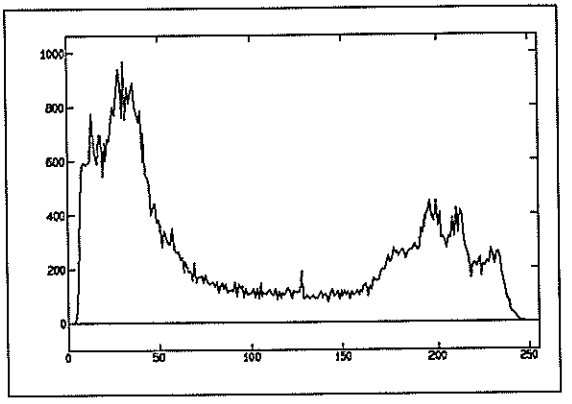


figure 1b : Histogram of the porousness image.

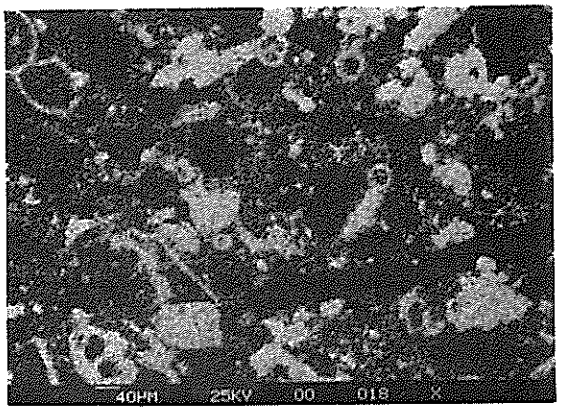


figure 2a : X-ray calcium image and ...

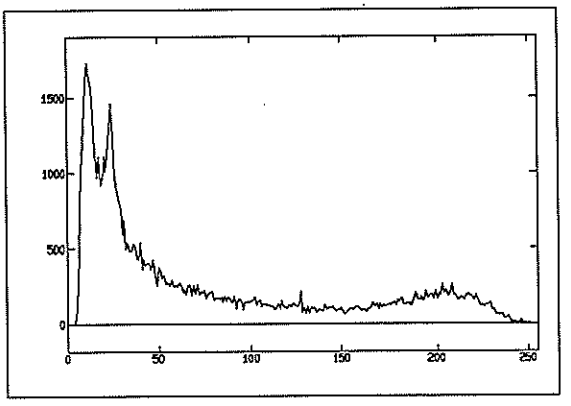


figure 2b : Calcium histogram.

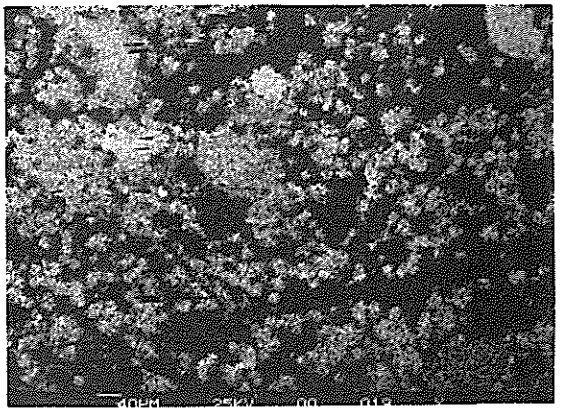


figure 3a : X-ray silicon image and ...

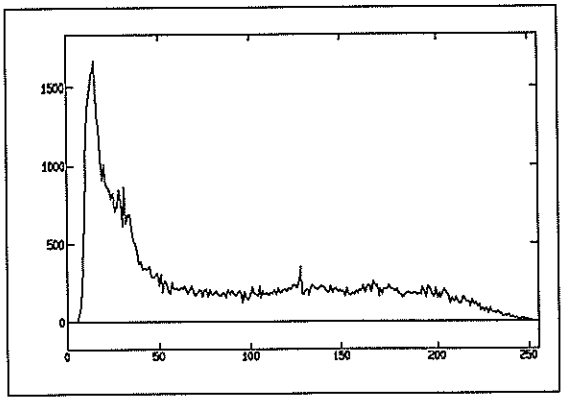


figure 3b : Silicon histogram.

extraction must rely on different information.

Scanning electron microscopy gives extra indications by analysing photons. From the same section, two other noisy X-ray images are available :

i) an X-ray calcium image (fig. 2a) where the Kα ray is taken into account. Particle with calcium are in white, everything else in black.

ii) an X-ray silicon image (fig. 3a) where the siliceous particles are in white.

The interesting thing is that mostly all the particles of the stone are composed either of silicon or of calcium with different concentration. So, they appear only on one image at a time and permit to separate the siliceous particles from the calcareous ones.

As the information of the electron microscope are not directly converted in digital images, the polaroid photos must be well adjusted one to the others during acquisition. In addition, technical difficulties for lighting and for the adjustement of the camera induce a lost of information.

CLASSICAL THRESHOLD VERSUS EXPERT THRESHOLD:

As seen in the three histogramms, the average grey level of particles and background are different enough. So, a global threshold technique can be used [2][3] for each image. The method used is explained in [4]. Each component of the mixture is supposed to be normally distributed by the mean of μ, a standard deviation σ and a *a priori* probability P. First, an initial threshold vector is crudely obtained from smooothed histograms. Each component is given by the position of the minimum between the two maxima of the related histogram. Then, an updated threshold is computed according to the Bayes minimum error rule by an iterative method. This is done separately on the three images and this gives a T vector with three components : Tp for porousness image, Tc for calcium image and Ts for silicon one. The T vector is (88,58,53).

For comparison, partition of the three images is made by an expert and the T vector is equal to (104,144,130).

Tp is quite well but Tc and Ts are not. This is due to the gaussian hypothesis on the histograms and spread of particles in X-ray images : they appear bigger than they are [5]. This important experimental artefact disturbs the method and another approach seems necessary.

IMAGES SUPERPOSITION :

As told before, the three images are issued from the same scene and each image is only related to a particular component : porousness, calcium and silicon. If the extraction was perfect, these images would be superposable as a jigsaw puzzle. This does not suppose any sort of distribution of the histograms. Then, a criterion related to the reliability of this complementarity will be explained. In this purpose different levels are affected to the two components of binarized images :

i) porousness image : threshold Tp, level 1 for porousness and level 0 for the background.

ii) calcium image : threshold Tc, level 2 for calcium, level 0 elsewhere.

iii) silicon image : threshold Ts, level 4 for silicon, 0 for the background.

The quality of the superposition is tested by making an addition of the three binary ones. It is planned to find the T vector leading to the best attainable complementarity.

In an ideal case, the result will be an image with only three levels corresponding to well classified pixels :

i) Np pixels for porousness at level 1.

ii) Nc for calcium (level 2).

iii) Ns for silicon (level 4).

In the actual case, in addition to the three well classified pixels, five badly classified ones occur at intermediate levels. These are:

i) No at level 0 : pixels belonging to no component.

ii) Npc at level 3 : those belonging to porousness and calcite.

iii) Nps at level 5 : porousness plus silicon.

iv) Ncs at level 6 : calcium plus silicon. It should be noticed that this class includes a very few calcium silicate particles which are well classified. It will be easy to display them once the extraction will be performed.

v) Npcs at level 7 : porousness plus calcite plus silicon.

Most of these badly classified pixels appear at the boundaries of the different particles but silicate of calcium and particles whithout the analysed components can be seen.

PROCESSING METHOD :

The noise of the grey level X-ray images is lowered by a median filter of size 3 by 3. To find the optimal threshold T = (Tg,Tc,Ts), use is made of a discriminating function D(T) which represents the reliability of the superposition. It may be the sum of the badly classified pixels :

$$D(T) = No + Npc + Nps + Ncs + Npcs$$

To place this optimization problem without constraint[4], D(T) is represented figure 4a and 4b for a given Tg and for Tc and Ts varying from 56 to 251 step 13. As seen, the shape at this scale seems to present only one minimum. But with a step one and close to the minimum, local minima appear. The chosen method must take this into account and in our case, gradient and simplex methods have been tested [6].

PRACTICAL RESULTS AND COMPARISONS :

The minimum of D(T) has been found equal to 8754 and in an automatic implementation, D(T) is a good evaluation of the partition quality. The T vector equal to 148, 140 and 64. Expert, classical and optimization results are compared

in figure 5a and 5b. As seen, badly classified pixels are minimum for our method and the T vector is close to the expert appreciation for X-ray images but not for the porousness one. This is due to particles which are not siliceous neither calcareous. The expert recognizes these particles and choses the good threshold. A similar approach can be taken. The first one is to make iron and sulphur X-ray images and every particle will be analysed. An other way is to extract these particles using mathematical morphology ; this method is under active investigation.

Afterwards, the final work is to assign badly classified pixels to one of the three components. For this purpose, the neighbourhood of each badly classified pixel is watched in each of the three binary images and this pixel is classified.

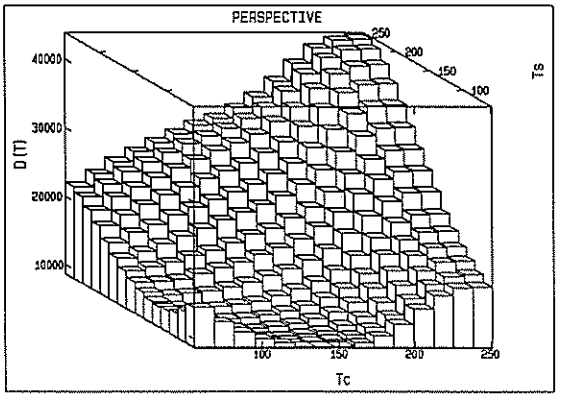


figure 4a : Tridimentional shape of D(T) versus Ts and Tc for a given Tp (147).

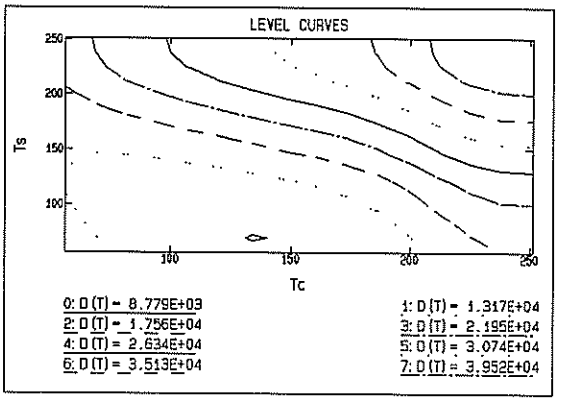


figure 4b : Level curves of D(T) versus Ts and Tc for a given Tp (147).

	Tp	Tc	Ts	D(T)
Expert	104	144	130	9142
Baye. Min.	88	58	53	21166
Our method	148	140	64	8754

figure 5a : Comparison of the three methods. D(T) is the sum of the bad classified pixels.

	porousness	calcium	silicon
Expert	25482	12120	18892
Baye. Min.	18447	11907	14006
Our method	19337	11936	25509

figure 5b : Comparison of the three methods. The number of pixels of each class is given.

CONCLUSION :

This paper describes a method to extract particles of a stone using complementary images of a same scene by taking into account the superposition of these images. This method gives the best result in the term of superposition and does not suppose any kind of distribution of the histograms. This method can be improved taking into account the few particles whithout calcium or silicon.

REFERENCES :

[1] D. Jeulin " Mathematical morphology and material image analysis " ed Scanning Microscopy International vol 2 p 165 184 (1988).

[2] Y. Nakagawa and A. Rosenfeld " Some experiment on variable thresholding " Pattern recognition vol 11 pp 191 204 (1979).

[3] P.K. Sahoo, S. Soltani A.K.C. Wong and Y.C. Chen " A survey of thresholding techniques " Computer vision, graphics and image processing 41 pp 233 260 (1988).

[4] J. Kittler and J. Illingworth " Minimum error thresholding " Pattern Recognition Vol. 19 pp 41 47 (1986).

[5] J.P. Eberhart " Méthodes physiques d'étude des mineraux et des matériaux solides " Doin éditeurs (1976).

[6] R. W. Daniels " An introduction to numerical methods and optimization techniques " North-Holland editor (1978).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Texture Synthesis Using Nonhomogeneous Gaussian Markov Random Fields Model

Zou Cairong Wang Taijun He Zhenya

Dept. of Radio Engineering, Southeast University, Nanjing , P.R.China

This paper is mainly concerned with the texture synthesis using nonhomogeneous Gaussian Markov Random Fields (GMRF) Model. The existing texture synthesis method using MRF model is discussed. Two aspects of MRF are considered to synthesize the texture. Suppose the autocorrelation function of the texture image is exponentialy distributed, we get simple recursive form of the algorithm.A more general texture synthesis method using Gibbs distribution is also described.

1. INTRODUCTION

Texture synthesis and texture analysis are important contents of computer vision, remote sensing and image analysis [1-3]. We can segment the natural scenes using texture measure. For example , we can discriminate among rivers, grass lands and streets from an aeronautical image.The surface orientation and depth can also be computed according to the texture change. For texture,there is no precise definition. Generally, we think of texture as something that is random, periodic or has some definite structure and placement. There are mainly two methods in the study of texture analysis, they are statistical estimation and structure description. Because texture covers the entire image, the statistical models become a powerful method in both texture synthesis and texture analysis. Markov Random Field model has found wide application in the field of image processing and image analysis, such as image restoration, image segmentation, image smoothing and texture analysis [4-8]. Among them ,texture synthesis and texture analysis using MRF model attracted many researchers. There are mainly two methods in the study of texture analysis using MRF. One is proposed by R.L. Kashyap et al.[9], they use Gaussian MRF model because it results in a diffirence equation in spatial domain. They derived a texture synthesis algorithm which use the Fast Fourier Transform , we must choose the optimal neighbors when we analyze the texture. Another is used by A.K. Jain and G.C.Cross[10]. They compute a local conditional probablity according to the binomial distribution. A neighborhood system model is used. Using Metropolis algorithm they obtain many different texture. D.Geman and S. Geman point out the method used by A.K. Jain and G.C. Cross is really one example of Gibbs Distribution[4]. But untill now, there is no report that apply Gibbs Distribution to the synthesis of texture. In this paper we describe a fast texture syhthesis method . We fist discusse the one dimensional case, then the two dimensional case.The homogeneous and nonhomogeneous case are applied to produce the texture. We synthesize texture using nonhomogeneous random fields because there exist a lot of nonhomogeneous natural texture. We also describe a texture synthesis algorithm which use Gibbs Distribution, three kinds of optimization methodws (ICM, SA, DP) can be used.

2. TEXTURE MODEL

We first discuss one dimesional case. The local probablity expression for 1th Markov property is

$$Pr\{y(n)|\text{all } y(m),m=n\}=Pr\{y(n)|y(n-l)\}$$

Where y(n) is one dimensional random signal. If y(n) is also Gaussian distributed we can obtain the equation as follows

$$R(t_1,t_3)*R(t_2,t_2)=R(t_1,t_2)*R(t_2,t_3)$$

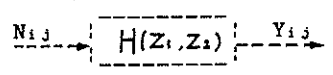
Where $R(t_i,t_j)$ is autocorrelation functon of y(t). Assuming y(t) is stationary, that is to say , $R(t_i,t_j$ depends on the difference (t_i-t_j) only. It is easily shown that

$$R(t_i,t_j)=K*\exp\{a(t_i-t_j)\}$$

This equation is an equivalent expression of 1th order Gaussian Markov process .Extending this one-dimensional case to two-dimensional case, we have

$$R(t_{mn},t_{kl})=A_mA_nB_kB_l*\exp\{a(m-k)+b(n-l)\}$$

If y_{ij} is homogeneous we can use Z-transform to get the sample function of a linear system which outputs the signal y_{ij} while the input sinal is i.i.d. Gaussian white noise N_{ij}.



For stationary homogeneous case we have

$$Z\{R(x,y)\}=H(z_1,z_2)*H(z^1,z^2)*Var$$

Where Var is the variance of input white noise, $H(z_1,z_2)$ is Z-transform of sample function.It is easily shown that

$$Z\{(R(x,y)\}=\frac{Var(1-M^2)(1-N^2)}{(1-MZ_1)(1-NZ_2)(1-MZ_1^{-1})(1-NZ_2^{-1})}$$

Where M=exp(-a),N=exp(-b),and we select

$$H(z_1,z_2)=\frac{1}{(1-MZ_1^{-1})(1-NZ_2^{-1})}$$

and

$$Var=K(1-M^2)(1-N^2)$$

Thus we the random value y_{ij} is the combination of neighboring elements and i.i.d. white noise. It is a unilateral Gaussian Markov random filed. The bounded recursive procedure is

$$Y(1,1)=W(1,1)$$

$$Y(i,1)=MY(i-1,1)+W(i,1) \qquad 1<i<NL+1$$

$$Y(1,i)=NY(1,i-1)+W(1,i) \qquad 1<i<ML+1$$

$$Y(i,j)=MY(i-1,1)+NY(i,j-1)-MNX(i-1,j-1)+W(i,j) \qquad 1<i<NL+1,\ 1<j<ML+1$$

Where NL and ML are image lattice size. It is obvious that the resulted bounded texture image is both Gaussian and Markovian.The texture synthesis results using homogeneous GMRF model are illustrated in Fig.1. For nonhomogeneous case we first write $Y(j)=\{y_{1j},y_{2j},y_{3j},\ldots,y_{nj}\}$.Since the random field is Markovian, we suppose the two dimensional signal y_{ij} is computed according to a recursive procedure. So it is an unilateral Markov random field model .That is to say Y(j) can be calculted form the combination of Y(j-1) and W(j), where $W(j)=\{w_{1j},w_{2j},w_{3j},\ldots,w_{nj}\}$. A recursive method that extends the one dimensional case to two dimensional case is used. Fortunately, there exists algothm for the generation of two- dimensional random signals. E.Hryniewicz proposed the numerical generation method for two dimensional Gaussian random fileds[14]. The recursive procedure is stated as follows

$$Y(1,1)=a_1b_1W(1,1)$$

$$Y(i,1)=a_i/a_{i-1}MY(i-1,1)+a_ib_1(1-M)W(i,1) \qquad 1<i<NL+1$$

$$Y(1,i)=b_i/b_{i-1}NY(1,i-1)+a_1b_iW(1,i) \qquad 1<i<ML+1$$

$$Y(i,j)=a_i/a_{i-1}MY(i-1,1)+b_j/b_{j-1}NY(i,j-1)-a_ib_j/(a_{i-1}b_{j-1})MNY(i-1,j-1)+a_ib_j\ 1-M\ 1-N\ W(i,j) \qquad 1<i<NL+1,\ 1<j<ML+1$$

The simulation result using nonhonogeneous GMRF model is illustrated in Fig.2. We can also set other forms of autocorrelation function so as to include more nerghbor pixels. Different selection of the autocorrelation function results in different kinds of texture.

3. GIBBS DISTRIBUTION

Gibbs Distribution(GD) is an equivalent description method of Markov random fields. It is firstly used in the field of statistical physics. In the GD the local probability measure p{w} on sample space is given by

:
Antecedent condition m:
If p is ω_m then r is A_m
Antecedent:p is ω'

consequence: r is A '

:
:

[Fuzzy reasoning 2]
Antecedent condition 1:
If p is ω_1 then θ is B_1 else
Antecedent condition 2:
If p is ω_1 then θ is B_2 else

:
:

Antecedent condition m:
If p is ω_m then θ is B_m else
Antecedent:p is ω'

consequence: θ is B'

Fuzzy relation of there reasonings can be shown that

$$R_1 = \omega \rightarrow A = \omega \times A. \quad (3.2a)$$
$$R_2 = \omega \rightarrow B = \omega \times B. \quad (3.2b)$$

Furthermore,the composition rule of Fuzzy condition and antecedent to infer from consequence make following expression:

$$\omega' * (\omega \rightarrow A) = \omega' * (\omega \times A) = \omega \circ A' (\omega \times A) = A' \quad (3.3a)$$
$$\omega' * (\omega \rightarrow B) = \omega' * (\omega \times B) = \omega \circ B' (\omega \times B) = B' \quad (3.3b)$$

where Fuzzy relations R_1 ,R_2 are defined as product of the membership functions. A symbol '$\circ$' means max-min composition. Furthermore, 'else' in Fuzzy conditional proposition correspond to $\cup$(join). Let max be symbol '$\vee$' let min be symbol '$\wedge$' . Composition rule is as follows:

$$A' = \omega' \circ (\omega_1 \times A_1) \cup \omega' \circ (\omega_2 \times A_2) \cdots \cup \omega' \circ (\omega_m \times A_m) \quad (3.4)$$

$$\mu A'(r) = \{\mu w(p) \wedge (\mu_{w1}(p) \cdot \mu_{A1}(r))\} \vee \{\mu w'(p) \wedge (\mu_{w2}(p) \cdot \mu_{A2}(r))\} \cdots \vee \{\mu w'(p) \wedge (\mu_{wi}(p) \cdot \mu_{Ai}(r))\} \cdots \vee \{\mu w'(p) \wedge (\mu_{wm}(p) \cdot \mu_{Am}(r))\} \quad (3.5)$$

$$B' = \omega' \circ (\omega_1 \times B_1) \cup \omega' \circ (\omega_2 \times B_2) \cdots \cup \omega' \circ (\omega_i \times B_i) \cdots \cup \omega' \circ (\omega_m \times B_m) \quad (3.6)$$

$$\mu B'(\theta) = \{\mu w(p) \wedge (\mu_{w1}(p) \cdot \mu_{B1}(\theta))\} \vee \{\mu w'(p) \wedge (\mu_{w2}(p) \cdot \mu_{B2}(\theta))\} \cdots \vee \{\mu w'(p) \wedge (\mu_{wi}(p) \cdot \mu_{Bi}(\theta))\} \cdots \vee \{\mu w'(p) \wedge (\mu_{wm}(p) \cdot \mu_{Bm}(\theta))\} \quad (3.7)$$

The approximate position of the end poit can obtained by the above expression.

4. EXPERIMENT

4.1. A preprocess of two images of the initial point the end point

In a practical flow, the initial point of the tracers in the primary picture have done one correspondence to the end point or the tracers in the secondary pictures using Fuzzy reasoning.There is a quadrangular prisim as an obstacle in the flow (Photo.4.1,4.2). And the four circles are in the each corner as the marker for taking in agreement with two images The preprocess to extract position of the tracer using Fuzzy reasoning is given the following procedure.

[Procedure 1] A picture of an initial point in the flow are feeded by TV-camera to the image processor. The given image is performed by binarization. Furthermore, noise is removed by using the eight-neighbors of an isolated point.

[Procedure 2] The connecting components in image given by Procedure [1] are labeled and obtained a size of an area and the center of the gravity. The connecting components can be classified according to the area because the size of each areas is different from each others.And then,let each center of the gravity is coordinated.

[Procedure 3] [Procedure[1] and [2] are performed in regard to the picture of the end point and it should be obtained a gap of a position to the end point. And the position of the end point image is corrected by transferring and rotating of the image.

Photo.4.1 and 4.2 are showed the position of the initial point and the end point in Table 4.1 after the performing preprocess and computing.

4.2. Inference of the end point using Fuzzy reasoning

The end point corresponding to the initial point is obtained using Fuzzy reasoning from the position of the initial point and the end point of the tracer after the preprocessing. The flow is divide into sixsub-regions.The membership functions($\mu_{wi}(p)$,i=1,2,$\cdots$,6) meaning ratio at an initial point belongs to the region must be defined,since sub-regions are Fuzzy regions.In practice ,the regions are selected by a joystickas watching the image on the monitor and the regions are divided. $\mu_{wi}(p)$ in each region is given as follows,that is an effect of some neighborhoods is in proportion to the distance. Where x and y are the position of the initial point respectively. The membership functions on the distance and the angle in each region ($\mu_A(r)$,$\mu_B(\theta)$) are given depending on the flow velocity without an obstacle and the shape of the obstacle in the flow. The position of the end point is estimated by using equations (3.5) and (3.7), and by these membership functions and the position of the initial point of the tracers. And then, the practical end point being with correspondence.

4.3. A result of an experimental

Table 4.2 is a result the end point an image of the initial point in Photo 4.1 and is with correspondence to the end point. And the result is indicated by the graphical display as shown in Fig.4.1. Twenty-four tracers being with one to one correspondence,seven initial points be with miss correspondence,for the initial points of thirty-one tracers.

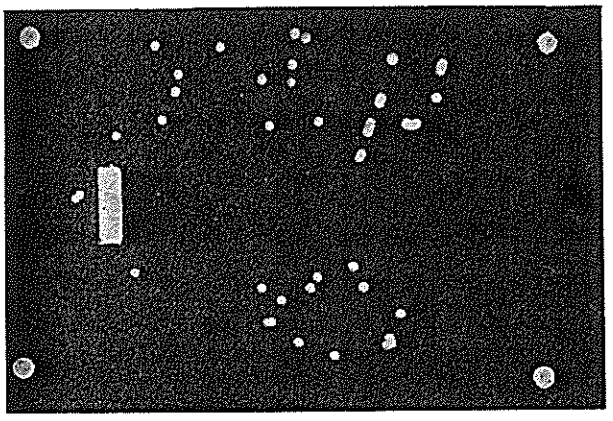


Photo.4.1
A primary picture of the tracer perticles.

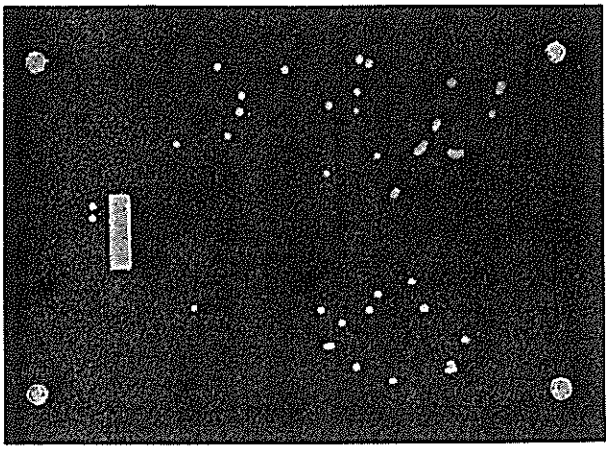


Photo.4.2
A secondary picture of the tracer perticles.

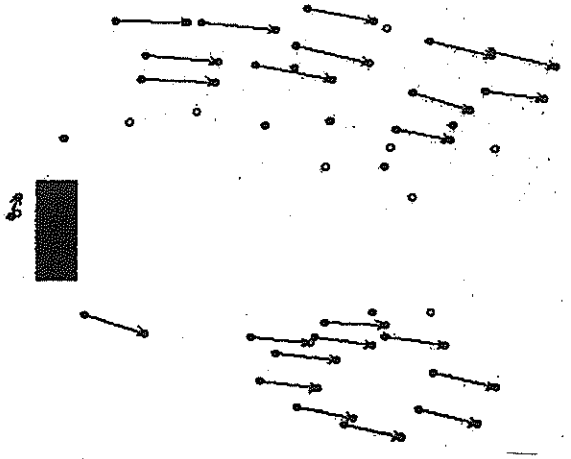


Fig.4.1 The result of Fuzzy reasoning.

Table 4.1
The tracer position of the initial point and the end point.

No.	X	Y	No.	X	Y
1	137	136	1	147	200
2	193	137	2	157	334
3	339	149	3	74	258
4	254	152	4	75	247
5	253	165	5	184	137
6	228	164	6	190	194
7	157	159	7	241	142
8	154	173	8	264	340
9	104	210	9	202	175
10	234	202	10	273	228
11	276	199	11	277	173
12	376	181	12	281	351
13	118	322	13	292	388
14	304	320	14	301	163
15	273	327	15	204	162
16	226	336	16	304	342
17	267	336	17	312	141
18	312	336	18	312	329
19	242	346	19	315	216
20	343	359	20	322	400
21	256	381	21	269	369
22	286	392	22	341	320
23	261	128	23	304	136
24	380	156	24	350	342
25	329	182	25	329	247
26	319	205	26	366	193
27	355	202	27	371	391
28	311	228	28	380	159
29	71	259	29	383	368
30	232	364	30	414	185
31	333	382	31	421	165
			32	353	211
			33	382	217

Table 4.2
The result of Fuzzy reasoning.

No.		No.	
I.P.	E.P.	I.P.	E.P.
1	5	17	16
2	7	18	24
3	28	19	12
4	14	20	29
5	—	21	13
6	11	22	20
7	15	23	23
8	9	24	31
9	—	25	26
10	—	26	32
11	—	27	—
12	30	28	—
13	2	29	4
14	—	30	21
15	18	31	27
16	8		

5. CONCLUSIONS

In order to analyze the flow visualized image, the distance and the direction of the motion of the tracer particle in a flow at a certain interval is obtained using an image processing technique with Fuzzy reasoning.
Firstly,the flow is divided into the Fuzzy regions and Fuzzy reasoning is giving the definition at the each Fuzzy region.
Second, the vector directed from an initial point to an end point is supposed, and the inferential method of the distance and the angle of the vector make it clear.
As processing the two pictures of the initial points and the end points the position of the tracers are detected.
After the flow is divided into the several region and Fuzzy reasoning is inferred by given membership functions which is expressed an approximate distance and angle. Consequently the initial points can be corresponded to the end points exactly in the region which is a little change of the flow. However the initial points do not be corresponded the end points in the region which is in vortex.
In the future, it is necessary to study more how to infer Fuzzy reasoning and to divide the region which is in vortex.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

Motion Field Estimation by 2-D Kalman Filtering *

J.N. Driessen and J. Biemond

Delft University of Technology,
Department of Electrical Engineering, Information Theory Group,
P.O. Box 5031, 2600 GA Delft, The Netherlands.

Abstract

In this paper a new pixel-recursive estimator is derived for the extraction of the motion field from two consecutive images. The estimator is based on a recursive 2-D AR vector model for the motion field and a nonlinear equation for the observation of the motion field in the time-varying intensities. Special cases of such models are predictors used in earlier proposed pixel-recursive algorithms. To reduce the computational complexity involved in the true 2-D Kalman filter, a sub-optimal one is proposed based on reduced-order model strategies. The update equation for the first two state elements in the filter is quite similar to earlier proposed updates. Theoretically, the advantage over the earlier proposed pixel-recursive estimators is the explicit assumption of motion field models and the incorporation of estimation error covariances. In practice some care has to be taken in updating these error covariances due to imperfect linearization of the nonlinear observation model.

1 Introduction

A motion field is the projection of 3-D object motion *onto* the image plane and it represents local displacements *in* the image plane. In an intensity image sequence the motion field relates intensities in consecutive frames along the so-called *motion trajectory* and in motion compensated processing of image sequences such as coding, filtering or interpolation the motion field is essential information. Prior to this processing, the motion field has to be estimated from the consecutive frames which is a difficult problem for two main reasons. A scene often consists of differently moving objects which causes the motion field to exhibit discontinuities at object boundaries and even regions in which the motion field is undefined due to uncovered object parts. Moreover, even within a moving object the estimation of the motion field is locally an ill-posed problem since only one intensity value variation is observed for two unknown vector components.

A possible solution to these problems is provided by a parameterization of the motion field as a function of 3-D object motion and surface parameters [1,2]. Such an approach turns the problem into a simultaneous parameter estimation and image segmentation problem. In such models, motion field discontinuities are included implicitely as the boundaries of segmented objects; uncovered regions can be found by testing the model hypothesis. Unfortunately, parametric models are more or less restricted to cover rigid motion of very smooth 3-D object surfaces. Another solution is the explicit assumption of smoothness of the motion field within a moving object [3,4]. Discontinuities are dealt with by adapting the smoothness constraint to intensity gradient information [3]. Intuitively, these techniques capture a more general class of motions than the parametric techniques. Since the algorithms developed here are intended to be used for real life sequences attention is focussed on these smoothness-based techniques.

The scope of this paper is to present a pixel-recursive motion field estimator technique that is based on a causal stochastic motion field model, a nonlinear observation equation and Kalman filter solution techniques. Originally, pixel-recursive techniques were introduced as adaptive estimators without explicit underlying modeling assumptions about the motion field: they consisted of a horizontal displacement vector predictor and a vector update that was based on the minimization of a very local functional [5]. In the following years, attempts were reported to incorporate more a priori model-knowledge into the estimators such as the assumption that the motion field can be modeled as a one-dimensional first order AR-process and estimated by a 1-D Kalman filter [6], that the update is a stochastic variable [7] and that the one-dimensional prediction can be improved by a two-dimensional prediction [8]. The estimator presented in this paper can be viewed as a generalization of these approaches.

2 Modeling Assumptions

In this section, the models which the estimator is based upon are presented. At first, the motion field model is described which is essentially a recursive stochastic model to cover structured but non-rigid motion. Finally, the motion field observation model is discussed which is essentially a non-linear model to cover large local displacement vectors.

2.1 Causal AR Motion Field Models

The motion field is assumed to be generated by the following causal 2-D AR vector process:

$$\vec{d}(x,y) = \sum_{(i,j)\in S} A_{ij}\vec{d}(x-i,y-j) + \vec{v}(x,y), \qquad (1)$$

with $\vec{d}(x,y)$ a motion field vector at location (x,y), with A_{ij} 2×2 matrices representing the AR-model parameters, with S a causal support region and with $\vec{v}(x,y)$ the driving noise which is assumed to be a zero-mean Gaussian random process with covariance matrix P_v. The AR-model matrices A_{ij} together

*This research is supported by the Netherlands Technology Foundation (STW).

with the noise covariance matrix P_v determine the smoothness of the motion field.

A convenient and elegant choice for the model matrices arises if it is assumed that the motion field has a decoupled separable autocorrelation function for each vector field component:

$$\begin{aligned} R_d(r,s) &= E[\vec{d}(x,y)\vec{d}^{\,T}(x-r,y-s)] \\ &= \begin{pmatrix} \sigma_x^2\rho_{hx}^{|r|}\rho_{vx}^{|s|} & 0 \\ 0 & \sigma_y^2\rho_{hy}^{|r|}\rho_{vy}^{|s|} \end{pmatrix}. \end{aligned} \tag{2}$$

Interestingly, a similar autocorrelation model has been proposed very recently by Namazi and Lee [9].

In natural scenes there is no reason for a difference between the correlation of the horizontal or vertical motion field component, so the correlation coefficients and the variances are assumed to be equal for both components:

$$\begin{aligned} \sigma^2 &= \sigma_x^2 = \sigma_y^2, \\ \rho_h &= \rho_{hx} = \rho_{hy}, \\ \rho_v &= \rho_{vx} = \rho_{vy}. \end{aligned} \tag{3}$$

If the horizontal and vertical sampling frequencies are equal, there is also no reason for a difference between the horizontal and vertical correlation coefficient in each vector component:

$$\rho = \rho_h = \rho_v \tag{4}$$

The AR model matrices for such a model are given by:

$$\begin{aligned} A_{10} &= \rho I, \quad A_{-11} = 0, \\ A_{01} &= \rho I, \quad A_{11} = -\rho^2 I \end{aligned} \tag{5}$$

and the noise covariance matrix is given by:

$$P_v = \sigma^2(1-\rho^2)^2 I. \tag{6}$$

The predictor of Tziritas [8] results from this model as a special case for the correlation coefficient equal to one and a zero noise covariance:

$$\vec{d}(x,y) = \vec{d}(x-1,y) + \vec{d}(x,y-1) - \vec{d}(x-1,y-1). \tag{7}$$

2.2 A Nonlinear Observation Equation

The motion field is observed as displaced intensities in an intensity image according to the following nonlinear observation model:

$$f(\vec{x},t) = f(\vec{x} - \vec{d}(x,y), t-dt) + w(\vec{x},t), \tag{8}$$

where $f(\vec{x},t)$ is the observed intensity at time t at spatial location (x,y), dt is the temporal distance between consecutive frames and where $w(\vec{x},t)$ is the observation noise which is assumed to be zero-mean Gaussian random process with variance σ_w^2. The assumption underlying model equation (8) is that the intensities along the motion trajectory are more or less constant. This assumption is not true in general and the noise term $w(\vec{x},t)$ accounts for small errors in this assumption and is *not* intended to account for noise present *in* the consecutive frames. This is indeed a simplification , however, it prevents the difficult combined motion estimation and image sequence filtering problem. Finally, notice that Eq. (8) provides only one equation in two unknown motion field vector components, which shows the ill-posedness of the estimation problem without *a priori* motion field model knowledge.

3 2-D Kalman Filter Solutions

In this section, we present a Kalman filter based on the models previously discussed. At first, to formulate the problem as a Kalman filtering problem, the motion field model and the observation equation are presented in a state-space formulation. Next, to reduce the computational complexity, the optimal Kalman filter has to be approximated based on sub-optimal approaches. Finally, it is shown that the update equation proposed here covers earlier proposed pixel-recursive update formulas as a degenerated case.

3.1 State-Space Formulation

To formulated the state-space equations, the image is assumed to be scanned line-wise starting at the most upper line and each line is assumed to be scanned from the left to the right. The state is defined as the set of vector elements of the motion field belonging to all previously scanned pixel locations that determine future vector elements of the motion field according to model (1). For a first-order AR vector model this results in the following state vector:

$$\begin{aligned} \vec{s}(x,y) &= [\vec{d}^{\,T}(x,y), \vec{d}^{\,T}(x-1,y), \cdots, \vec{d}^{\,T}(1,y), \\ &\quad \vec{d}^{\,T}(M,y-1), \cdots, \vec{d}^{\,T}(x,y-1)]^T, \end{aligned} \tag{9}$$

with M the number of pixels on a row. The state-space evolution equation is given by:

$$\vec{s}(x+1,y) = A\vec{s}(x,y) + C\vec{v}(x,y), \tag{10}$$

where the matrices A and C are the system matrices that have to be chosen appropriately. For a first-order AR motion field model, A is given by:

$$A = \begin{pmatrix} A_{10} & 0 \cdots 0 & A_{-11} & A_{01} & A_{11} \\ & & A_{sub} & & \end{pmatrix}, \tag{11}$$

where the matrices A_{ij} are the AR model matrices and where A_{sub} is a matrix containing zeroes and ones that only perform shifts of state elements. The driving noise $\vec{v}(x,y)$ only affects the first two state elements so the matrix C is given by:

$$C = \begin{pmatrix} I & 0 & \cdots & 0 \end{pmatrix}^T. \tag{12}$$

with I the 2×2 identity matrix and 0 the 2×2 zero matrix.

The state-space observation equation is given by:

$$f(\vec{x};t) = f(\vec{x} - H\vec{s}(x,y); t-dt) + w(\vec{x},t), \tag{13}$$

where H performs the extraction of the actual displacement from the state vector $\vec{s}(x,y)$:

$$H = \begin{pmatrix} I & 0 & \cdots & 0 \end{pmatrix}. \tag{14}$$

In the optimal extended Kalman filter, outlined in App. A, the first-order derivatives of the observation equation with respect to the state elements is needed. These derivatives are evaluated as follows:

$$\begin{aligned} F(\vec{x},\vec{s}(\vec{x})) &= \nabla_{\vec{s}} f(\vec{x} - H\vec{s}(\vec{x}); t-dt) \\ &= -H^T \nabla_{\vec{x}} f(\vec{x} - \vec{d}(\vec{x}), t-dt). \end{aligned} \tag{15}$$

Although the true extended Kalman filter is straightforward to derive, its implementation is computationally expensive due to the large dimension of the state which is in the case of a first-order model of order $O(2M)$.

3.2 Sub-Optimal Solutions

In the field of image restoration, two recursive sub-optimal 2-D Kalman filters that do not pose restrictions on the stationarity of the state-space models have been proposed [10,11]. The reduced update Kalman filter (RUKF) [10] assumes that the gain in the optimal Kalman filter is relatively close to zero outside a small region called the *local state*. The number of gain elements outside the local state is set to zero and since this number is quite large this reduces the computational load by applying a smart implementation. The reduced-order model Kalman filter (ROMKF) [11] is based on a reduction of the dimension of the state in the true state-space equations and Kalman filtering applied to the resulting state-space equations. The RUKF is a sub-optimal Kalman filter based on the original 2-D models, where the ROMKF is an optimal Kalman filter based on simplified state-space equations. The computational complexity of the ROMKF is less than the complexity of the RUKF with only a minor decrease in performance.

For a first-order AR motion field model the reduced-order model state is defined, similar to [11], by:

$$\begin{aligned}\vec{s}_R(x,y) &= [\vec{d}^T(x,y), \vec{d}^T(x-1,y), \vec{d}^T(x+2,y-1), \\ &\quad \vec{d}^T(x+1,y-1), \vec{d}^T(x,y-1)]^T. \end{aligned} \quad (16)$$

The state evolution equation is given by:

$$\vec{s}_R(x+1,y) = A_R\vec{s}_R(x,y) + B_R\vec{u}(x,y) + C_R\vec{v}_R(x,y), \quad (17)$$

where $\vec{u}(x,y)$ is an input variable that represents the most recent estimate of the displacement vector at the spatial location $(x+3, y-1)$:

$$\vec{u}(x,y) = \vec{d}_e(x+3,y-1) + \vec{v}_u, \quad (18)$$

where $\vec{v}_u$ accounts for the uncertainty in the estimate. The noise term $\vec{v}_R$ includes the driving noise $\vec{v}$ from the original model together with $\vec{v}_u$:

$$\vec{v}_R = \begin{pmatrix} \vec{v} \\ \vec{v}_u \end{pmatrix}. \quad (19)$$

Under the assumption that both components in the noise term are independent, the covariance matrix equals:

$$P_{v_R} = \begin{pmatrix} P_v & 0 \\ 0 & P_u \end{pmatrix}. \quad (20)$$

The model matrices are given by the following equations:

$$A_r = \begin{pmatrix} A_{10} & 0 & A_{-11} & A_{01} & A_{11} \\ I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \end{pmatrix}, \quad (21)$$

$$B_r = \begin{pmatrix} 0 & 0 & I & 0 & 0 \end{pmatrix}^T. \quad (22)$$

$$C_r = \begin{pmatrix} I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \end{pmatrix}^T. \quad (23)$$

$$H_r = \begin{pmatrix} I & 0 & 0 & 0 & 0 \end{pmatrix}. \quad (24)$$

with again I the 2×2 identity matrix and 0 the 2×2 zero matrix.

In this specific case the state dimension is of order 10 and the implementation can be done very efficient due to the large number of zeroes in the matrices and the large number of shift operations performed by the identity matrices in the system matrices. The computational complexity of this ROMKF is reduced substantially compared with the original Kalman filter.

3.3 Relation with Existing Algorithms

To show the relation with existing pixel-recursive algorithms, the updating equation for the first two state elements is considered into detail. The equation for the gain matrix is given by:

$$K = -P_b H^T \nabla f [\nabla^T f H P_b H^T \nabla f + \sigma_w^2]^{-1}, \quad (25)$$

where the spatial coordinates are omitted for convenience. Notice first that the inversion is a simple scalar inversion. A closer look at the matrix $HP_b(x,y)H^T$, tells us that the operation $H \cdot H^T$ extracts the most upperleft 2×2 submatrix of the covariance matrix. This matrix is refered to a P_d, since it represents the prediction error covariance matrix of, the actual displacement vector. The post multiplication of the matrix P_b by H^T extracts the first two columns of the state error covariance matrix. The first two elements of the gain, refered to as K_d, are thus given by:

$$K_d = -P_d \nabla f [\nabla^T f P_d \nabla f + \sigma_w^2]^{-1}. \quad (26)$$

The update of the first two elements results by multiplying $K_d(x,y)$ by the innovation term refered to as $dfd(\vec{x}, \vec{d}_b, t)$:

$$\begin{aligned} dfd(\vec{x}, \vec{d}_b, t) &= f(\vec{x}, t) - f(\vec{x} - H\vec{s}_b, t - dt) \\ &= f(\vec{x}, t) - f(\vec{x} - \vec{d}_b, t - dt) \end{aligned} \quad (27)$$

Let the covariance matrix be restricted to a diagonal matrix with σ_d^2 as diagonal elements. In that case the update for the first two elements of the state is given by:

$$\vec{u}_d = [|\nabla^T f|^2 + \mu]^{-1} dfd(\vec{x}, \vec{d}_b, t) \nabla f, \quad \mu = (\frac{\sigma_v}{\sigma_d})^2. \quad (28)$$

This equation is equal to the Wiener-based update for an observation window containing one element only.

4 Initial Experimental Results

Initial experimental results show that to obtain reliable smooth motion field realization from the 2-D AR vector model, the correlation parameter has to be fairly large (> 0.9) and the driving noise variance has to be fairly small ($< 10^{-3}$). This is in contrast with [9] who used much lower values for their estimator. In our experiments, the values according to the above specified ranges where used. Furthermore, small variances for the observation noise component where used (< 2.0).

Next, due to the large correlation, the small driving noise variance, the Kalman gain cannot be assumed to be relatively close to zero outside the local state. Therefore, in order to keep the modeling error in the ROMKF small, the state has to be

extended on the left hand side. If the ROM-state is extended with L vector elements, its state becomes:

$$\begin{aligned}\vec{s}_R(x,y) &= [\vec{d}^T(x,y),\cdots,\vec{d}^T(x-L-1,y),\\ &\quad \vec{d}^T(x+2,y-1),\vec{d}^T(x+1,y-1),\\ &\quad \vec{d}^T(x,y-1)]^T. \end{aligned} \tag{29}$$

The dimension of this state is increased with $2L$ since every vector has two components. Extension of the state at the right hand side has no effect which is due to the reduced-order modeling assumptions.

To more or less solve the boundary value problem, which we feel to be of even more significance in the nonlinear case compared with the linear case, each first row has been processed by a 1-D Kalman filter. Such an strategy has been used since the only reasonable boundary value seems to be zero, which will be very disturbing when this is not the true value. This is in large contrast with the boundary value problem im image restoration, where the available blurred and noisy data can be used to derive boundary values [10,11].

The final set of experiments were performed to investigate the errors in the filter due to linearization errors. In [6], it has been analyzed that the 1-D Kalman filter may become instable due to linearization errors and a very accurate linearization has been suggested. However, in their experiments linearization based on a least squares plane fit were used. It may be better to make a distinction between two operation modes of the extended Kalman filter used here: a convergence mode and a tracking mode. In the convergence mode a global linearization and in the tracking mode a local and accurate linearization has to be performed.

5 Discussion and Future Research

In this paper, a 2-D Kalman filter solution to the problem of estimating motion fields from consecutive frames has been proposed and analyzed theoretically. Existing pel-recursive techniques were shown to be degenerated cases of the Kalman filter. Sofar, experimental results concentrated on the convergence capabilities.

Future research will concentrate on performance comparison between the Kalman filter based algorithm and the existing pel-recursive algorithms. Major attention is focussed on elegant techniques to circumvent instabilities due to linearization errors and the inclusion of discontinuities in the motion field model.

A The Extended Kalman Filter

In this appendix the extended Kalman filter for a nonlinear observation equation is outlined briefly. Let the state-space equations be given by:

$$\begin{aligned}\vec{s}(x+1,y) &= A\vec{s}(x,y)+B\vec{u}+C\vec{v},\\ \vec{z}(x,y) &= \vec{f}(\vec{s}(x,y))+\vec{w}(x,y),\end{aligned} \tag{30}$$

where $\vec{h}(\vec{s}(x,y))$ is a nonlinear vector function of the state vector $\vec{s}(x,y)$. The extended Kalman filter is divided into a prediction and an update part. The prediction part follows the normal linear prediction equations:

$$\begin{aligned}\vec{s}_b(x+1,y) &= A\vec{s}_a(x,y)+B\vec{u},\\ P_b(x+1,y) &= AP_a(x,y)A^T+CP_vC^T,\end{aligned} \tag{31}$$

where the b means *before* updating and where the a means *after* updating. P_b and P_a are the state error covariance matrices before and after updating, respectively, or in other words they are the prediction and estimation error covariance matrices.

The update part of the extended Kalman filter differs from the linear Kalman filter and is given by:

$$\vec{s}_a(x,y) = \vec{s}_b(x,y)+K(x,y)[\vec{z}-\vec{h}(\vec{s}_b(x,y))], \tag{32}$$

$$K(x,y) = P_b(x,y)F^T[FP_b(x,y)F^T+P_v]^{-1}, \tag{33}$$

$$P_a(x,y) = [I-K(x,y)F^T]P_b(x,y),$$

with $K(x,y)$ the so-called Kalman gain, I the identity matrix with dimension equal to the state dimension and with F the *Jacobian* of the nonlinear vector function $\vec{f}(.)$ with respect to the state vector:

$$F=\begin{pmatrix} \frac{\partial f_1(\vec{s})}{\partial s_1} & \cdots & \frac{\partial f_1(\vec{s})}{\partial s_{N_s}} \\ \vdots & & \vdots \\ \frac{\partial f_{N_f}(\vec{s})}{\partial s_1} & \cdots & \frac{\partial f_{N_f}(\vec{s})}{\partial s_{N_s}} \end{pmatrix}, \tag{34}$$

with N_s and N_f the dimension of the state and the vector function $\vec{f}(.)$, respectively. A familiar choices for the location to evaluate the Jacobian is the the most recent estimate for the state: $\vec{s}_b(x,y)$.

References

[1] Murray, D.W. and B.F. Buxton, "Scene Segmentation from Visual Motion using Global Optimization", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 2, March 1987, pp. 220-228.

[2] Hötter, M. and R.Thoma, "Image Segmentation based on Object Oriented Mapping Parameter Estimation", *Signal Processing* 15 (1988), pp. 315-334.

[3] Nagel, H.H. and W. Enkelmann, "An Investigation of Smoothness Constraints for the Estimation of Displacement Vector Fields from Image Sequenecs", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol PAMI-8, no. 5, September 1986.

[4] Anandan, P, "A Computational Framework and an Algorithm for the Measurement of Visual Motion", *International Journal of Computer Vision*, vol. 2, no. 3, 1989, pp. 283-310.

[5] Netravali, A.N. and J.D. Robbins, "Motion-Compensated Television Coding: Part I", *Bell System Technical Journal*, BSTJ-58, no. 3, March 1979, pp. 631-670.

[6] Stuller, J.A. and G. Krishanamurthy, "Kalman Filter Formulation of Low-Level Television Image Motion Estimation", *Computer Vision, Graphics and Image Processing*, vol. CVGIP-21 (1983), pp. 169-204.

[7] Biemond, J., L. Looijenga, D.E. Boekee and R.H.J.M. Plompen, "A Pel-Recursive Wiener-based Displacement Estimation Algorithm", *Signal Processing* 13 (1987), pp. 399-412.

[8] Tziritas, G., "Displacement Estimation for Image Predictive Coding and Frame Motion-Adaptive Interpolation", *Visual Communications and Image Processing '88*, T. Russell Hsing Ed., Proc. SPIE 1001, pp. 936-941 (1988).

[9] Namazi, N.M. and C.H. Lee, "Nonuniform Image Motion Estimation from Noisy Data", *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-38, no. 2, February 1990, pp. 364-366.

[10] Woods, J.W. and C.H. Radewan, "Kalman Filtering in Two Dimensions", *IEEE Transactions on Information Theory*, vol. IT-23, no. 4, July 1977, pp. 473-482.

[11] Angwin, D.L. and H. Kaufman, "Image Restoration using Reduced-Order Models", *Signal Processing* 16 (1989), pp. 21-28.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

EFFECTS OF MOTION ESTIMATION ERRORS ON VOLUMETRIC AND PICTORIAL RECONSTRUCTION

Aldo GRATTAROLA, Sandro ZAPPATORE

DIST - Università di Genova
Via Opera Pia 11 A - 16145 - GENOVA - ITALY

The paper deals with an integrated system devoted to volumetric and pictorial object reconstruction from a set of bidimensional perspective views, acquired with a standard TV camera. The critical aspect issue of the system is the calibration of each view, i.e. the accuracy in determining position and orientation of the viewpoint. In the proposed system such a calibration is obtained via a motion recovery technique from corresponding points. The effects of calibration errors, produced by imperfect data, on the quality of the volumetric model built from occluding contours are analyzed by means of simulation. We finally propose a new regularization technique, based on physical constraints concerning the reconstruction geometry, significantly improving the results in very preliminary simulations.

1. INTRODUCTION

The recovery of 3D information from a sequence of 2D views is a basic issue in many applications, like computer vision, robotics, object recognition, image processing and computer graphics [1-3]. According to the application, the sequence of views is obtained by a set of static sensors, or by means of a relative motion between the objects and the sensors.

This paper deals with an integrated system devoted to volumetric and pictorial object reconstruction from a set of bidimensional perspective views, acquired with a standard TV camera and calibrated by means of a motion recovery algorithm.

The basic modules of the proposed system are:

- an intrinsic camera parameters estimator (off line) based on one of the available methods [4];
- an extractor of corresponding point features in the different views. The used approach selects the candidate points in each frame and finds the correspondences by a relaxation labeling procedure (modified from [5]). In this step the object-background segmentation is also performed;
- a calibration module for the extrinsic camera parameters based on a motion estimation algorithm providing, for each 2D view, the spatial position of the viewpoint and the camera orientation. The motion estimation is performed according to a modified linear approach [6], that improves the robustness of the results with respect to errors in the input data by imposing suitable regularization constraints;
- a volumetric reconstructor of the object under analysis by means of an occluding contour technique: the volumetric reconstruction is based on an efficient 3D model generated by intersecting the infinite conic-like volumes obtained from the silhouettes of the perspective views [7];
- a module that analyzes the pictorial information and integrates it with the 3D model with a resolution that is independent of the volumetric resolution, and can reach the detail level of the original images [7].

Relevant features of the proposed system are: the use of motion estimation to calibrate the perspective views, so that the 3D reconstruction can be performed also if the environment is not enough structured or under control of the observer; an efficient volumetric representation, exploiting a run-length coding technique to minimize both the storage requirement and the rendering computational burden; the separate representation of the pictorial information with resolution independent from the volumetric one; the capability of iteratively refining the 3D model when new views become available without restarting the whole procedure.

This paper does not describe the volumetric and pictorial reconstruction modules, that have already been discussed elsewhere (see, e.g., [7]) and focuses on:

i) effects of errors in the estimates of relative positions and orientations of the viewpoints on the quality of the volumetric and pictorial reconstruction;
ii) requirements on the original data to constrain such errors within a range providing an acceptable reconstruction.
iii) a new proposed regularization technique based on geometrical constraints devoted to improve the reconstruction quality.

In the following we first describe (sec. 2) the technique used to estimate the extrinsic camera parameters. Then we present (sec. 3) an analysis of the reconstruction sensitivity to calibration errors and we illustrate some simulation results. In the fourth section the basic steps of the proposed regularization technique are presented.

Work supported by the Italian Ministry of Education (MPI40%) and the National Research Council.

2. CAMERA CALIBRATION BASED ON MOTION ESTIMATION

The method used to calibrate the spatial position and orientation of the acquisition system consists of estimating the 3D motion leading the TV camera from a viewpoint to the next one. The motion parameters are estimated by means of an algorithm that exploits eight or more corresponding points, on the object under analysis, to build a set of linear equations whose solutions are the so called essential parameters [8] e_i, i=1, ..., 9:

$$A\,\underline{e} = 0 \qquad (1)$$

In eq. (1) A is a matrix of elements derived from the coordinates of the corresponding points and $\underline{e}$ is the vector of the unknowns. From the 3x3 matrix E obtained rearranging the elements of $\underline{e}$, the actual motion parameters can be computed by a singular value decomposition technique or other equivalent methods (see, e.g., [9]).

The matrix E describing the motion of a rigid body is characterized by the following specific algebraic structure [10]: 2 singular values coincide and the third one is zero. If A is affected by noise deriving from errors in the acquired data, the solution for E does not exhibit in general the previous properties. These latter cannot be imposed by explicitely constraining the image coordinates of the corresponding points, since their displacements induced by the varying perspective allow computing the motion parameters. On the other hand, the solutions of Eq. (1) can be constrained by means of two equations expressing the rigidity constraints: letting b_i's, i=1,2,3, be the rows of E, such constraints are

$$F_1(e_1,...e_9) = \mathrm{abs}(\det(E)) = 0$$

$$F_2(e_1,..e_9) = \mathrm{abs}(4[\,\|b_1\|^2\,\|b_2\|^2 + \|b_1\|^2\,\|b_3\|^2 + \|b_3\|^2\,\|b_2\|^2 - (b_1\cdot b_2)^2 - (b_1\cdot b_3)^2 - (b_3\cdot b_2)^2] - [(\|b_1\|^2 + \|b_2\|^2 + \|b_3\|^2)^2] = 0$$

in terms of the norms of b_i and of their scalar products.

The augmented system

$$\begin{cases} A\,\underline{e} = 0 \\ F_1^u(e_1,...e_9) = 0 \\ F_2^v(e_1,..e_9) = 0 \end{cases}$$

where the parameters u and v allow optimizing the weights of the constraints with respect the basic equations, becomes globally nonlinear and can be solved by an iterative procedure (of the Gauss - Newton kind) starting from a guess solution provided by Eq. (1). A number of simulations have proven that the augmented system yields estimates closer to the real motion parameters than in absence of regularization constraints, especially when few corresponding points are available and at poor resolution.

3. ERROR ANALYSIS

As previously stated, the object reconstruction from multiple views requires that these latter be exactly calibrated with respect to each other. Of course, in the proposed procedure, any error produced in the extrinsic calibration parameters by approximations in the relative motion estimate affects the result of the volumetric and pictorial reconstruction. We have started the investigation of this basic issue from the two following viewpoints:

i) effects of errors in the estimates of relative positions and orientations of the viewpoints on the quality of the volumetric reconstruction;
ii) requirements on the original data to constrain such errors within a range providing an acceptable reconstruction.

The first kind of effects has been analyzed by simulating various error levels in the relative position estimates within the reconstruction procedure. To keep the results fully under control, a synthetic "original" object has been generated, namely a pyramid with a squared basis, that is both simple to manipulate and characterized by features (e.g. vertices) enhancing the reconstruction errors. On the basis of the results obtained in analogous real cases, it has been decided to base the reconstruction on five suitable (synthetic) views: specifically, we have generated the top view and four evenly spaced lateral views of the pyramid. These data have been used first in association with the exact calibration data, to produce a reference reconstructed object, at a resolution of 40x40x100 voxels (along the length, depth and height axes, respectively). The calibration data have been subsequently perturbated by simulated random errors affecting both the coordinates of the optical center and the direction of the optical axis in each view, with respect to a unique reference system. A set of reconstructions have been performed using the original five views in association with the wrong calibration data, and the reconstruction errors have been evaluated in various ways. A simple global measure of the reconstruction defects is represented by the exceeding and missing volumes in the reconstructed object with respect to the reference. The percentage ratio of their sum over the correct figure is shown in Tab. 1 versus the errors in the viewpoint position and in the direction of the optical axis. For every value of ε_t and ε_θ (average percentage errors of the viewpoints coordinates and of the optical axis angles, respectively), the reported

ε_θ \ ε_t	0	1	2	3	5	7
0	-	0.121	0.127	0.142	0.192	0.254
1	0.198	0.201	0.213	0.232	0.285	0.338
2	0.318	0.314	0.317	0.326	0.356	0.399

Table 1

Average percentage volume error versus percentage errors of the viewpoint coordinates, ε_t, and of the angle the fining the camera orientation, ε_θ.

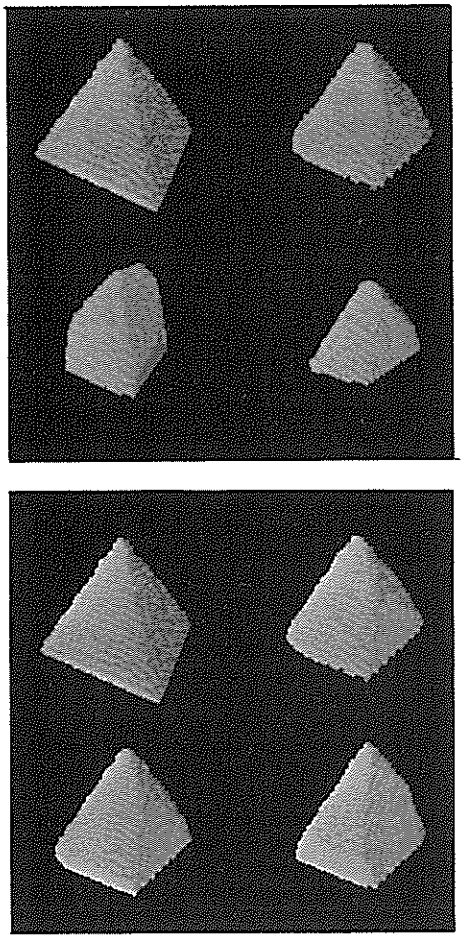


Figure 1

Reconstruction results from simulated imperfectly calibrated views. In each figure, the object reconsructed from perfectly calibrated data is shown at top left. The other reconstructions are obtained from data affected by the following percentage calibration errors: TOP) t.op right ε_t=0, ε_θ=1; bottom left ε_t=0, ε_θ=2; bottom right ε_t=0, ε_θ=3; BOTTOM) t.r. ε_t=1, ε_θ=1; b. l. ε_t=3, ε_θ=1; b.r. ε_t=7, ε_θ=1. The resolution is in all cases of 40x40x100 voxel.

volume error is the average over many simulations, performed by randomly varying the components of the calibration errors. It should be noticed that the actual values of the volumetric error in single simulations may differ significantly (even more than 50%) from the displayed averages, especially for large errors.

A less quantitative but perhaps more immediate evaluation of the reconstruction errors can be visually performed by examining the Fig. 1, that show some reconstruction results obtained from imperfectly calibrated data. These examples are chosen from the simulation runs characterized by error values close to the average reported in Tab. 1

It can be noticed that, in general, the reconstruction results are more haevily affected by the miscalibration of the direction of the optical axis than by the error in the optical center, due to the higher sensitivity of the procedure to the former parameter. Moreover, the results are not acceptable for calibration errors outside the range of 1%, especially on the direction. We have therefore investigated (by extending the study and the simulations reported in [6]) which requirements on the input data to the motion estimation procedure constrain the calibration errors within this range. The relevant parameters are the number of corresponding points, their distribution in space, the kind and amount of 3D motion and the resolution of the original 2D data. We have considered in particular the dependence on the number of correspondences, by choosing an image resolution of 512x512 pixel and amounts of relative motion corresponding to those leading from one view to another in the previously described simulation, and by averaging out the effects of the spatial distribution of the points over many simulations with randomly generated coordinates of the corresponding points. The results are summarized in Fig. 2, where the average errors in the location of the optical center and in the direction of the optical axis are plotted against the number of corresponding points n, in the two cases of linear and nonlinearly constrained estimation. It is apparent from the general error behaviour that, for the chosen values of the other parameters, it is not difficult to constrain ε_t and, more important, ε_θ within the 1% range, by choosing n suitably higher than the minimum value of 8. The results confirm once more that better performances, i.e. lower errors for a given value of n, are usually provided by the technique exploiting the rigidity constraint.

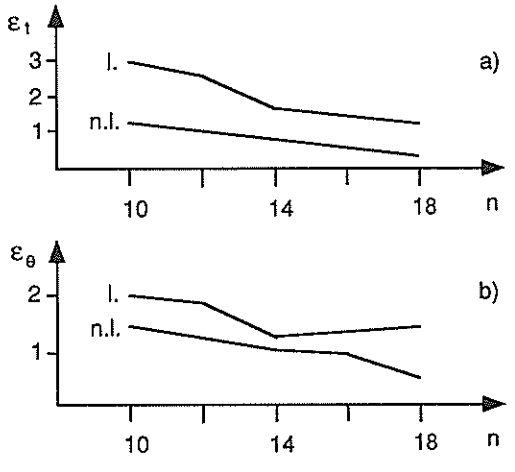


Figure 2

Average percentage errors of the viewpoint coordinates ε_t (a) and of the camera orientation angles ε_θ(b) versus the number n of corresponding points used in the linear (top curves) and nonlinearly constrained (bottom curvers) motion estimation algorithm.

4. RECONSTRUCTION IMPROVEMENT

As we have seen, the basis of the volumetric reconstruction procedure is determining the intersection volume of the generalized cones defined by the projection of the object in the available perspective views. Consider the i-th generalized cone C_i consisting of the infinite straight lines connecting the vertex (the viewpoint) with all the points belonging to the object projection in the corresponding view. The system geometry implies that, for any other cone C_k, $k \neq i$, every line of C_i must have at least a point in common with C_k, i.e. it must intersect at least one of the lines of C_k. In other words, no path exists internal to C_i, leading from its vertex to any point internal to C_i on the other side of the object, and not touching at least one point belonging also to the volume defined by C_k.

Due to the unavoidable errors, in particular on the calibration parameters, such a constraint is not satisfied, in general, for the available data. The most visually noticeable effect of cone "misregistration" is the increasing "erosion" of the reconstructed object when the error increases, as shown in Fig. 1. Our hypothesis is that, if the errors are sufficiently small, imposing the above constraint on the available data reduces the reconstruction error while improving the calibration parameters.

To reduce the problem dimensionality we notice that the errors on the extrinsic parameters are in general greater than those connected with image quantization, object-background segmentation and approximated intrinsic parameter estimation. Assuming then that these latter can be neglected, at least as a first approximation, we seek a regularization procedure to improve the estimate of the six extrinsic parameters of each view according to the previously mentioned constraint. The dimensionality of the problem is still high, involving 6N variables if N is the number of available views, and the equations representative of the 3D geometrical model are not simple, so that we have discarded the analytical formulation of the proposed system, in favor of a numerical algorithm, presently under test, based on the iteration of simple steps to search the minimum of a cost function along the steepest descent direction.

The chosen error function is obtained as follows. Let us consider the generic i-th view V_i representing the object projection in binary form separated from the background, and let us project the viewpoint corresponding to a different view V_k, $k \neq i$, onto the projection plane of V_i. If the projection point falls within (or very close to) V_i, V_k cannot be used in conjunction with V_i, otherwise the cone defined by V_k is projected onto the projection plane of V_i, where an angle is obtained. Two distances, one for each edge of the angle, are defined: if the edge does not intersect V_i, d is the minimum distance between the two, otherwise it is the maximum distance between the edge and the points of V_i outside the angle. The relative error between V_i and V_k is defined as $E_{ik} = d_{ik1} + d_{ik2}$ and the total error with respect to V_i is $E_i' = \sum_{1}^{N}{}_k E_{ik}$, for $k \neq i$ and not including the views that cannot be used for the condition on the viewpoints. The total error for the complete set of views is $E_T = \sum_{i=1}^{N} E_i'$. The proposed procedure to minimize E_T is the following:

i) choose suitable values for the steps Δ by which each calibration parameter is varied during the search (e.g. as a function of the error range);

ii) compute E_T for the $3^6 \cdot N$ combinations of parameter values obtained by varying the 6 parameters of each view by 0, $+\Delta$ and $-\Delta$, while those of the other views are unchanged (for N=5 this means computing 3645 values of E_T);

iii) modify the parameters corresponding to the minimum value of the E_T so computed and go to ii), provided the error reduction is above a fixed threshold.

The whole procedure can be iterated with decreasing step sizes Δ, until a stop condition (e.g. on the error decrement) is met.

REFERENCES

[1] D.H. Ballard, C.M. Brown, Computer Vision, Prentice-Hall, Englewood Cliffs, New Jersey, 1982.

[2] P.J. Besl, R.C. Jain, "Three-Dimensional Object Recognition", Computing Surveys, vol. 17, pp. 75-145, 1985.

[3] V. Cappellini, R. Casini, M.T. Pareschi, C. Raspollini, "From Multiple Views to Object Recognition", IEEE Trans. Circuits and Systems, vol. CAS-34, pp. 1344-1350, 1987.

[4] R.Y. Tsai, "An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision", Proc. of the IEEE Comp. Soc. Conf. on CVPR, Miami Beach, Florida, pp. 364-374, 1986.

[5] S.T. Barnard, W.P. Thompson, "Disparity Analysis of Images", IEEE Trans. PAMI, vol. PAMI-2, pp. 333-340, 1980.

[6] C. Braccini, G. Gambardella, A. Grattarola, S. Zappatore, "Motion estimation of rigid bodies: effects of the rigidity constraints", in Signal Processing III: Theories and Applications, I. T. Young et al., Eds. New York, NY: North -Holland, pag. 645-648, 1986.

[7] C. Braccini, G. Gambardella, A. Grattarola, M. Milanta, M. Sassoli, "Pictorial Reconstruction of Three-Dimensional Objects through Multiple Views", in J.L. Lacoume et al., Eds., Signal Processing IV: Theories and Applications, North-Holland, Amsterdam, pp. 1461-1464, 1988.

[8] R.Y. Tsai, T.S. Huang, "Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces", IEEE trans. PAMI, Vol. PAMI-6, N° 1, pag. 13-26, 1984.

[9] J. Weng, T.S. Huang, N. Ahuja, "Error analysis of motion parameter estimation from image sequences", in Proc. of First Int. Conf. Comp. Vision, London, pag. 703-707, 1987.

[10] T.S. Huang, O.D. Faugeras, "Some Properties of the E Matrix in Two-View Motion Estimation", IEEE Trans. PAMI, Vol. 11, N° 12, pp. 1310-1312, 1989.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

On a statistical model for moving pictures

Peter Vogel

Philips Kommunikations Industrie AG, Thurn-und-Taxis-Str. 14,
8500 Nürnberg 10, West Germany

Typical Moving pictures for visual telephony are analyzed. The Concept of a spherically invariant random process (SIRP) proves to be very helpful for modelling image signals. This concept, which has already been used for speech coding, is promising for image coding as well.

1. Introduction

In image processing, signal models are mostly restricted to second order moments, e.g. autoregressive random fields [1]. They do not consider multivariate probability density functions (pdfs) of the signal. However, high dimensional pdfs are indispensable for a precise source model, e.g. for coding purposes. Such pdfs arise from a SIRP model, for example, which has already been used for description [2] and coding [3] of speech signals. As will be shown, they are also expedient in the description of the local behaviour of image signals.
The main intention is for improvements to a hybrid DPCM/transform coder for low bit rate video telephony (64kbit/s) standardized by CCITT (Fig. 1). Since this scheme incorporates motion compensated prediction, statistical dependencies in temporal direction are utilized by this scheme. Statistical properties of the remaining prediction error in spatial direction are investigated in the following. Rate distortion functions based on this properties have already been evaluated [4].

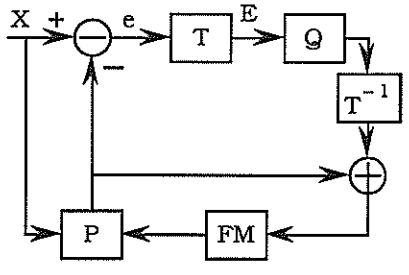
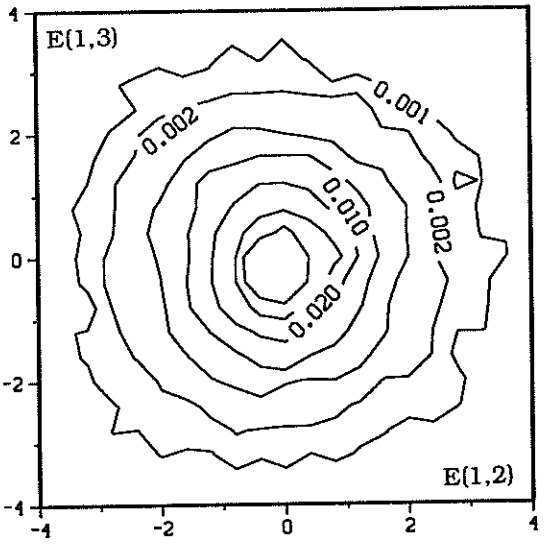


Fig. 1 Hybrid DPCM/transform coder for moving pictures

X Actual frame to be encoded (10 frames per sec, 288x352 luminance pixels per frame)
e Prediction error
E Prediction error after transform
FM: Frame memory for previous reconstructed frame
P: Linear predictor
T: 2-dimensional discrete cosine transform (block size 8x8)
Q: Quantizer

2. Modelling of the prediction error

The prediction error is partitioned into spatial square blocks e of size n=mxm, m=8. These blocks are subjected to the DCT for blockwise decorrelation (Fig. 1). The resulting block of coefficients is denoted by E. For the 2-dimensional pdf of closely neighbouring coefficients concentric ellipses are found for the contour lines (Fig. 2). Due to the normalization of the coefficients to variance 1 they appear as circles. Since the DCT is orthogonal, i.e. performes a rotation of the coordinate axis, the contour lines before the DCT are also ellipses.



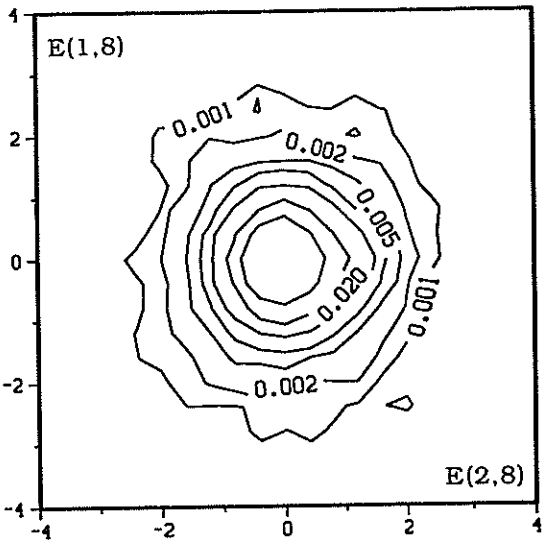


Fig. 2 Contour lines for 2-dimensional pdf of DCT coefficients E(1,2),E(1,3) and E(2,8),E(1,8) normalized to variance 1.

The spherical characteristic indicates vanishing correlation between DCT coefficients. In fact, the decorrelation property was found to be true for all DCT coefficients.
The spherical characteristic holds for a Gaussian pdf, for example. The Gaussian assumption would imply statistical independence between coefficients due to the decorrelation property. Statistical independence between the DCT coefficients is often asserted but is false, as will be outlined.
A contradiction can be obtained by the pdf of

$$(1) \qquad \sigma_n^2(e) \triangleq \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{m} e_{i,j}^2$$

which is the block energy per sample (Fig. 3).

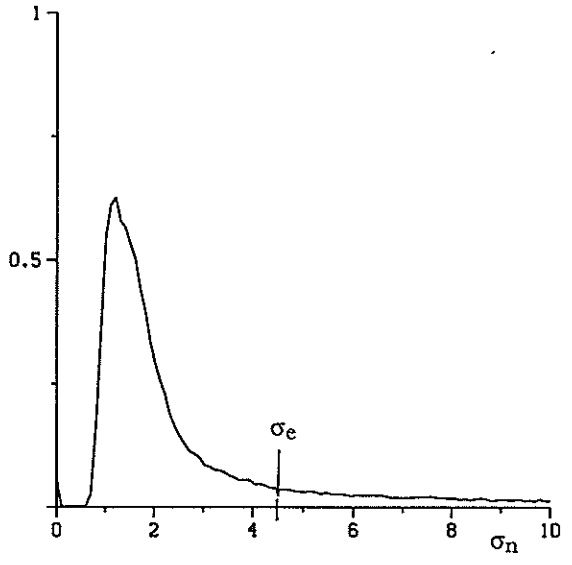


Fig. 3 pdf of block energy per sample
σ_e is the standard deviation of the prediction error process.

In case of statistical independence, most of the probability would lie on the expected value σ_e (Fig. 3). Although σ_e is approximated by σ_n when n tends to infinite, the probability is not concentrated on σ_e. Thus statistical dependence is proven in spite of nearly decorrelated DCT coefficients. Fig. 3 reveals that σ_n is concentrated at a smaller value than σ_e. This shows that large and small coefficients are not mixed in an arbitrary way, but that coefficients with small values often appear together with small coefficients. A physical interpretation of this phenomenon becomes obvious considering e a prediction error block. Since the main part of a picture for visual telephony consists of background and slow motion, many prediction error blocks possess only small energy.

The spherical characteristic of contour lines in Fig. 2 suggests a SIRP model. This model is expedient for statistical dependency expressed by the pdf of the block energy. A SIRP can be characterized by a composite source as follows [2]. Depending on the value of a positive random variable σ_{SIRP}, a Gaussian process with standard deviation σ_{SIRP} and autocorrelation function (ACF) $\sigma_{SIRP}^2\, C(k)$ is switched towards the output of the source. Here C(k) denote the correlation coefficients of the prediction error process. This SIRP characterization proves to be very helpful for generation of SIRPs [2] as well as for coding SIRPs [3]. It implies that the n-dimensional conditional pdfs are Gaussian, i.e.

$$(2) \qquad f_n(\underline{e}\,|\,\sigma_{SIRP}) \overset{SIRP}{=} f_n^G\{\sigma_{SIRP}^2\, C(k)\}\,(\underline{e}) \quad .$$

In the following determination of σ_{SIRP} is discussed.

For a given value of σ_{SIRP}, e is Gaussian distributed with standard deviation σ_{SIRP} by (2). With respect to this pdf,

$$(3) \qquad \sigma_n(\underline{e})\,\big|\,f_n^G\{\sigma^2\, C(k)\} \approx \sigma$$

holds for n=64 where σ stands for σ_{SIRP}. Here it is assumed that the correlation coefficients C(k) are vanishing for k=64. Consequently, σ_n defined in (1) can be used as an estimate of σ_{SIRP} for n=64 (good approximation of the σ_{SIRP}-pdf by the σ_n-pdf for n=40 is outlined in [3], pp. 134).

For a SIRP (2) holds for all integer n. This yields

$$(4) \qquad \sigma_{SIRP} \overset{SIRP}{=} \lim_{n \to \infty} \sigma_n(\underline{e}) \quad .$$

An experimental evaluation of (4) would be senseless, because this would result in a deterministic value which is σ_e, the standard deviation of the prediction error process. This follows from the ergodicity of the prediction error process, which is assumed during all measurements carried out in this paper. Contrarily, a SIRP is not ergodic because σ_{SIRP} is a random variable (unless the SIRP is a Gaussian process). Thus a SIRP is only used here for modelling the local behaviour of the prediction error process up to n=64 samples.

Replacing σ_{SIRP} in (2) by σ_n yields the estimate

$$(5) \qquad \hat{f}_n(\underline{e}|\sigma_n(\underline{e}) = \sigma) \triangleq f_n^G\{\sigma^2\, C(k)\}(\underline{e}).$$

This estimate is not exact, since (5) does not vanish outside the sphere

$$\sigma_n(\underline{e}) = \sigma$$

as

$$f_n(\underline{e}|\sigma_n)$$

does. However, (5) is concentrated next to this sphere due to (3).

From (5) an estimate

$$(6) \qquad \hat{f}_n(\underline{e}) \triangleq \int f_{\sigma_n}(\sigma) f_n^G\{\sigma^2\, C(k)\}(\underline{e})\, d\sigma$$

is also obtained for unconditioned multivariate pdfs. Statistical dependency expressed by the pdf of the block energy is directly incorporated in (6) in terms of this pdf. Both (5) and (6) will be substantiated by experimental results.

By (5), the ACFs under σ_n- condition differ only in a constant factor which agrees with experimental results.

Since the DCT maintains the Gaussian property, by (6), a univariate coefficient pdf is composed of Gaussian pdfs. In particular, the value at position 0 is obtained by integrating

$$f_{\sigma_n}(\sigma) / \sigma.$$

Since small σ_n-values occur (c.m.p. Fig.3), the value at position 0 is high. This agrees with the spiking characteristic of univariate coefficient pdfs (Fig. 4.1).

From (5), it follows that the conditional univariate coefficient pdfs conditioned by σ_n are Gaussian. This agrees with Figs. 4.2-4.

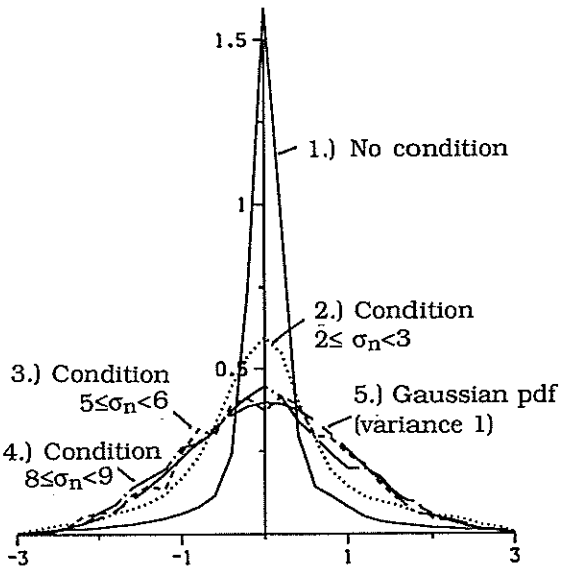


Fig. 4 Pdfs of DCT coefficient E(1,3) normalized to variance 1.

3. Conclusions

The experimental results given in Figs. 2 and 4 suggest modelling of moving pictures with the aid of SIRPs. By (5) and (6), multivariate pdfs of the prediction error are composed of Gaussian pdfs with an identical correlation coefficient structure. As a result, these pdfs are constant on ellipsoids.
In [4] it has been shown that purely horizontal and vertical frequency coefficients occur simultaneously only rarely. This kind of statistical dependency excludes a SIRP model. However, a SIRP model is expedient to express statistical dependency discovered by the pdf of block energy (Fig. 3).

[1] Chellappa,R., Kashyap, R.L., Texture synthesis using 2-D noncausal autoregressive models, IEEE Trans. ASSP (1985), 194-203.

[2] Brehm,H., Stammler,W., Description and generation of spherically invariant speechmodel signals, Signal Processing (1987), 119-141.

[3] Trottler,K., Informationstheoretische Untersuchungen zur Vektorquantisierung von sphärisch invarianten Sprachmodellprozessen, Diss., tech. Fak. der Univ. Erlangen-Nürnberg, 1987.

[4] Vogel,P., On the determination of rate distortion functions for moving pictures,ITG-Fachbericht 107, VDE-Verlag Berlin (1989), 35-41.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Multi-Resolution Image Segmentation In Higher Dimensional Feature Spaces Using Local Transforms

Caspar Horne*
Signal Processing Laboratory
Swiss Federal Institute of Technology
EPFL-Ecublens, CH-1015 Lausanne, Switzerland

Abstract

The problem of using multiple features for unsupervised image segmentation is addressed. The high dimensional feature space that results from many feature extraction processes is reduced in dimensionality in a local manner, resulting in a lower complexity in decision making. The locality is provided by the segmentation method that uses a combined region analysis and region estimation approach. Such an approach allows well developed supervised dimensionality reduction techniques to be used in an unsupervised manner. In this paper the approach is illustrated by using discriminant analysis, which reduces the dimensions of the feature space to a single one, so that the region properties are represented in a local subspace with maximum discriminatory power.

1 Introduction

The partitioning of an image into regions of homogeneous property is an important step in machine vision and consequently has received considerable attention of many researchers over the past two decades. The many different approaches to this problem that have been developed during this time can be divided into two broad categories, namely edge based and region based techniques.

In the edge based techniques the regions are defined by discontinuities established by applying local operators to feature images. The main problem in these techniques is the selection of those discontinuities that correspond to regions boundaries and the rejection of those that are inside regions. To get to such a robustness usually the data is processed over several scales. The problem arises here how to combine the output of the different operators working at different scales. A second problem, one relevant to the work described here, is the design of operators that work not only over a range of different scales, but that work also with high dimensional signals, such as those that are obtained by texture feature extractors. A set of scalar operators can be used working on projections onto one dimensional signals, but the problem of how to combine them arises.

Region based approaches directly work on homogeneity properties to establish the regions. Here the main problem is one of scale, that is, the establishment of the homogeneity of both small and big regions. Recently a set of techniques has evolved that process the data over a range of scales. The most well known techniques are the linked pyramid algorithm of Burt et al. [1], and the quadtree segmentation algorithm of Spann and Wilson [2]. These techniques showed the power of multi resolution processing, but suffer from a lack of generality, the former algorithm requiring knowledge of the number of regions present, the latter requiring assumptions about the minimum region size.

To surmount these difficulties a technique was developed by Spann and Horne [3,4] which is able to surmount these difficulties. Here a strategy is adopted where the regions are simultaneously analyzed and estimated going from a high resolution to a low resolution. The technique uses a local clustering algorithm where the amount of clustering is controlled by the local information present in the analysis window. Such an approach allows the automatic spawning of seeds (compact descriptions of regions) at resolution levels

*The support of Thomson CSF is gratefully acknowledged

corresponding to their natural size, and has been shown to give accurate segmentations, even in extremely low inter-region signal to noise ratios.

The work described in this paper extends this work to images that can only be accurately segmented using high dimensional feature spaces. Examples of such images are satellite images where the features are the spectral bands provided by the satellite, and textured images, where the features are computed by a texture feature extraction process. In such high dimensional spaces the problem arises which features to use. To discriminate neighboring regions it is often advantageous to discard noisy features and take only the representative ones. Clearly this feature set will be different from region to region. Therefor the feature selection is embedded in a segmentation algorithm that provides locality with respect to the regions.

The organization of the paper is as follows. Section 2 describes the segmentation algorithm. Section 3 and section 4 describe the local transforms of the feature space and section 5 shows some results on a three dimensional feature space. Finally section 6 presents conclusions and a discussion.

2 Multiresolution segmentation

The segmentation algorithm is based on a linked pyramid structure [1], which allows regions of any shape and size to be represented and processed in a multiresolution representation. The representations at each level of this pyramid are computed in a non linear manner. Instead of convolving the image with a set of filters that are (band)limited in the frequency domain, and then subsampling the image, the representations at different resolutions are obtained by a simultaneous region analysis and region estimation procedure. The algorithm proceeds from the base of the pyramid, which corresponds to the highest resolution, to the top of the pyramid, thus constructing the pyramid level by level in bottom up fashion.

The algorithm is based on the idea that a region can be represented in a hierarchical structure in such a way that at each resolution in the hierarchy the dimensions of the region are reduced by approximately the same amount as the total image dimensions. Allowing only integer dimensions, regions then disappear at resolutions in the hierarchy corresponding to their original dimensions. In Figure 1 a one dimensional example of such a structure representing two regions, containing respectively three and five pixels, is shown. From this figure it can be seen that a simple structure as the quadtree (which corresponds to a binary tree in 1-D) is not sufficient for representational purposes, and that a more complicated structure as the linked pyramid is needed.

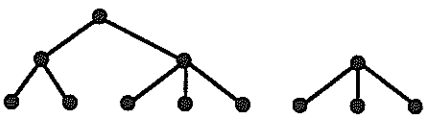


Figure 1: Pyramid structure

Now if the regions are known beforehand, then the complete pyramid can easily be constructed. Once constructed, the complete pyramid representing the whole image consists of a set of sub pyramid structures, each distinct region being represented by its own pyramid structure. The top of each pyramid will then be the node containing all descriptive region information, such as gray level or textural information, while the structure of the pyramid, together with the location of its top node describes all spatial information. If we describe segmentation as "detecting what is where", then the top node contains the "what", and the structure of the pyramid indicates the "where".

In the general segmentation problem, the regions of an image are not known beforehand, and have to be estimated by the segmentation procedure. This is done by a local clustering method. Here the nodes at each level of the pyramid are grouped together, the grouping being controlled by a dynamically estimated local threshold. This procedure makes it possible that nodes that do not belong to a bigger group of nodes, become automatically the top of the pyramid, and thus the "seed" node that represents a region.

3 Multiple features

The use of multiple features as input for the segmentation process results in a much greater complexity in decision making, as compared to the use of only one feature, such as only gray level information. This "curse of dimensionality" often makes it advantageous to work in a lower dimensional feature space. The dimensionality of the feature space can be reduced by using a projection in a lower order space. This lower order subspace has to contain all the relevant information that is necessary for the segmentation process, so that all the regions can be discriminated sufficiently.

A key observation here is that in order to discriminate between the regions they have to differ significantly in properties. Such an observation seems obvious in a one dimensional space where all regions properties can be described in the same feature space. In a

high dimensional feature space, however, region properties can be described in different feature subspaces. Now looking at the image at the region level we can see that each region is best represented in its own subspace, and that it is not only the value of the properties, but also the selection of particular properties that characterizes a region best.

In literature various techniques have been developed for reducing the dimensionality of the feature space to obtain a more manageable problem. Here we propose to use these transforms in a local way, without using a priori knowledge about classes or regions present in the image. Once these transforms are computed the output can be used for a stage of feature reduction or for a local weighting of a distance function, distance being the distance between classes in feature space.

4 Discriminant analysis

The most often used feature reduction method for unsupervised segmentation is the Karhunen-Loeve transform. Although this transform is optimal for minimum error representation, this is not necessarily the transform that gives the highest discriminating power. A more appropriate transform for segmentation purposes can be obtained by using Fisher's linear discriminant functions. These functions are optimal for a large variety of separability criteria. The Fisher's linear discriminant function maximizes the between-class scatter as compared to the within-class scatter.

Following [5] we can denote the samples feature vector by $\mathbf{x}$, then

$$y = \mathbf{w}^t\mathbf{x} \quad \text{with } ||\mathbf{w}|| = 1 \tag{1}$$

is the projection of $\mathbf{x}$ onto one dimension. Suppose now that in the two-class case the two classes $\mathcal{X}_1$ and $\mathcal{X}_2$ are known. In this case a within-class scatter matrix $\mathbf{S}_W$ and a between-class scatter matrix $\mathbf{S}_B$ can be defined as

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 \tag{2}$$

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t \tag{3}$$

where

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\mathcal{X}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \tag{4}$$

The projection vector $\mathbf{w}$ can be computed such that the criterion function

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}} \tag{5}$$

is maximized, resulting in

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{6}$$

In a general segmentation problem the number of classes are not known, and neither are the classes themselves. Fisher's linear discriminant functions are in that case not computable, because the area over which to compute the scatter matrices is unknown. The main problem is thus to estimate these regions, which is the same problem as the segmentation problem.

In the segmentation algorithm described previously in section 2, the region analysis is performed simultaneously with the region estimation procedure. At each moment of the algorithm each node represents one region. This means that at each moment and at each node of the pyramid the information present is restricted to one class. Thus at each pair of nodes the problem is reduced from a general n-class problem, with the classes and their number unknown, to a two-class problem, with the two classes known. The scatter matrices $\mathbf{S}_W$ and $\mathbf{S}_B$ from Equations 2 and 3 can then easily be computed.

5 Results

To illustrate the method we constructed a testimage of size 256 * 256 pixels consisting of two natural textures, shown in Figure 2 on the left. The textures were taken from the Brodatz album, D29 "beach sand" and D9 "grass lawn". The textures were arranged according to the maskimage shown in the same figure, on the right.

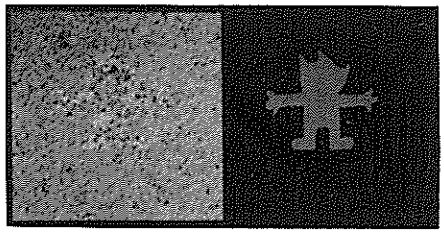


Figure 2: Testimage (left) and maskimage (right)

From this textured image three texture features were computed. The features are computed by convolving local masks with the image. The masks are three masks of the Unser feature set [6] and are approximations of horizontal, diagonal, and vertical edge detectors. In Figure 3 the three feature images are shown, together with the segmentation result. It can

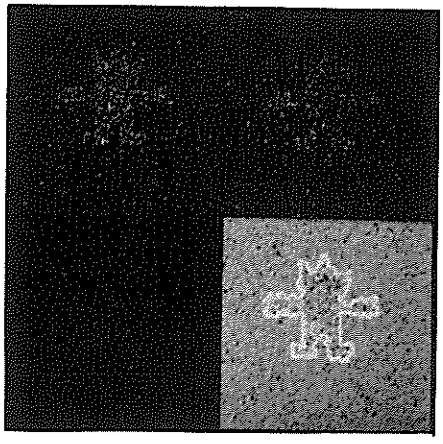


Figure 3: Texture features, and segmentation result

be seen that all three feature images contain information, so that the problem is really three dimensional, and that the information is very noisy.

The segmentation result shows the pixels having one or more neighbors belonging to a different region in white, superimposed on the original textured image. Due to the noise in the features, the boundaries are somewhat irregular, and have a somewhat blocky appearance. This blocky appearance is mainly caused by the way the boundary regions are displayed, the white boundary region being more than one pixel large.

The segmentation algorithm has found the two regions, corresponding to the two regions of the maskimage. Considering the low inter region signal to noise ratio, the boundaries are very well placed, and the overall result is excellent.

6 Conclusion

A segmentation algorithm has been presented that partitions images into homogeneous regions using high dimensional feature spaces. Due to the locality with respect to the regions local transforms of the feature space can be used. This locality is provided by the segmentation algorithm, which uses a simultaneous region analysis and region estimation process. Using local transforms of the feature space gives the advantage that features can be selected that are optimal with respect to the regions, and that can differ from one region to another. The method has been tested using discriminant analysis techniques and results have been presented.

It is anticipated that the method can used on the basis of other local transforms as well, such as the Karhunen Loeve transform, which is optimal for representational purposes. A further possibility will be the use of features that are computed on a multiresolution basis. In that case a different number of features can be used at each level of the pyramid, and each feature can be used at a level corresponding to the resolution of the feature extraction process. It is intended to report on progress of these topics in the future.

References

[1] P.J. Burt, T.H. Hong, and A. Rosenfeld. Segmentation and estimation of region properties through co-operative hierarchical compution. *IEEE Trans. Sys. Man. and Cyb.*, SMC-11:802–809, 1981.

[2] M. Spann and R. Wilson. A quad-tree approach to image segmentation. *Pattern Recognition*, 18:257–269, 1985.

[3] C. Horne and M. Spann. Region extraction using a dynamic thresholding pyramid. In *Proc. of the SPIE Conf. on Visual Comm. and Image Processing '88, Vol. SPIE-1001*, Cambridge, MA, USA, November 1988.

[4] M. Spann and C. Horne. Image segmentation using a dynamic thresholding pyramid. *Pattern Recognition*, 22:719–732, 1989.

[5] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* Wiley-Interscience, New York, 1973.

[6] M. Unser. Local linear transforms for texture measurements. *Signal Processing*, 11:61–79, 1986.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Texture Boundary Detection Based on LVQ Method

Ari Visa
Helsinki University of Technology
Laboratory of Information and Computer Science
Rakentajanaukio 2 C, SF-02150 Espoo, Finland

This paper concerns images containing stochastic textures. A new image segmentation method is shortly described. It's behaviour is studied at region boundaries. The power of the method is demonstrated on realistic stochastic textures. The suggested method is based on multiresolution representation of co-occurrence matrices, feature maps and Learning Vector Quantization (LVQ). The edge detection is achieved by region recognition. Each region is assumed to consist of unique texture type. The reported results are promising.

1 Introduction

The importance of texture recognition is increasing within the image processing field. There are four main ares of image processing in which texture plays an important role: classification, image segmentation, realism in computer graphics and image coding. Classification and segmentation can also be interpreted as a part of image analysis. All these fields imply that texture has many practical applications. There are three main problems in the texture analysis: 1) How can a textured region be described? 2) Given a textured region, to which of a finite number of classes does the sample belong? 3) Given a scene, how can the boundaries between the major textured regions be established? A proposition is made to the third question but the other questions are also discussed. First a definition to texture is given. Texture could be defined as a structure composed of a large number of more or less ordered similar elements or patterns without one of these drawing special attention. So a global unitary impression is offered to the observer. A texture can be a strictly ordered array of identical subpatterns, for instance a chessboard [Br 68]. Such a texture is called deterministic and it can be described by the characteristics of one subpattern or primitive and by the placement rules defining the spatial distribution of the primitives. The impression of a pattern can also obey some statistical laws. The resulting structure might resemble noise on a television screen [Br 68]. Such a texture is said to be stochastic. One should notice that deterministic textures can be heavily disturbed in their repetition and their primitives might be similar, but not at all identical. First there is the problem with texture description. Several methods to describe texture are known [Va85]. However, stochastic textures are different in character from other ones. For instance syntactical methods are not successful on stochastic textures. To classify stochastic textures the gray tone co-occurrence matrix method (GTCM) [Ha 73] has been found to be superior both theoretically [Co 80] and empirically [We 76]. The reported disadvantages of the GTCM method are discarding shape information in image and fixed resolution.The first one is not a drawback when stochastic textures are considered. The second one can be alleviated by using multiresolution representation. The segmentation of a textured image is not a new problem. There have been several attempts to solve the problem. Zucker and Rosenfeld [Zu 75] made segmentation based on local picture properties and histograms. Deguchi [De 78] developed a two dimensional linear estimation technique to characterize and segment textured images. Conners [Co 84] and Raafat [Ra 88] have also studied the problem. Conners has used cooccurrence matrices and texture measures derived from co-occurrence matrices but he used a statistical segmentation. The segmentation procedure consideres three kind of regions at each level of the segmentation: uniform, boundary and unspecified. At every level the procedures differentiated uniform regions from boundary and unspecified regions. Raafat introduced the idea of texture distance measure for classifying textures and directing the region growing process. There is, however, the learning problem. Attempts to apply neural networks on image compression [Lu 89], edge detection [Fu 88] and texture recognition [La 89] have lately reported.

2 TEXTURE SEGMENTATION

The texture boundaries are detected by a segmentation procedure. A direct approach to edge detection is difficult because edges have a different meaning in texture context than in ordinary images. The idea of the method is to use co-occurrence matrices in a multiresolution way. The co-occurrence matrices are considered as feature vectors. The learning samples define a feature map. The teaching of the feature map is accomplished by neural networks methods. The classification is done in feature space. The classification is based on distances.
A GTCM is computed in given direction and distance. The computation is done in three distances. The GTCM is computed as follows : Let $f(x,y)$ be a picture with $x \in 1,2,...NX, y \in 1,2,...NY$, where NX, NY are the

dimension of the picture in the $x-$ and $y-$ directions, respectively. The picture is digitized to NG gray levels, $G = 0, 1, ...(NG-1)$. Let D be a set of displacement vectors,

$D = \{d|d = (dx, dy),$
$0 < dx < NX,$
$-NY < dy < NY\}.$

Thus, $M_d(i,j)$ is defined to be the matrix whose (i,j)th element is the number of times that gray levels i and j occur in the relative position d. The entries in the GTCM, $M_d(i,j)$ of f with respect to a particular $d = (a,b)$ are defined as follows:

$$M_d(i,j) = \#\{((x1,y1),(x2,y2)) : \quad (1)$$

$f(x1,y1) = i, f(x2,y2) = j$, and $d = (a,b), d \in D$, such that $x2 = x1 + a$ and $y2 = y1 + b\}$ where $\#$ denotes the number of elements in the set.

A triplet $(M_{d1}(i,j), M_{d2}(i,j), M_{d3}(i,j))$ is extracted and considered as a pattern vector. To reduce the memory demand the co-occurrence matrices are usually replaced by some measures. In this case so is not done. In stead the co-occurrence matrices $(M_{d1}(i,j), M_{d2}(i,j), M_{d3}(i,j))$ are requantized and used as feature vectors.

The texture segmentation is executed partly in classical way partly with neural netwoks methods. A Feature map of N reference vectors are chosen [Ko 83]. The topology of the feature map doesn't matter. The distance metric can be calculated in many ways. In this case Euclidean distance is employed. N should be large enough that each texture class contains sufficient number of reference vectors. Each texture class corresponds with a texture region. The reference vectors are taught to each class in supervised learning mode. Suitable areas of image are selected and taught to the segmentation procedure. More samples, variations of reference samples, are further selected and taught to the segmentation procedure. Some of extracted feature vectors are used directly as reference vectors and the others are used to fine tune the feature map. The learning process is realized by learning vector quantization (LVQ) [Ko 86]. The LVQ method consists of two parts: First the closest reference vector m_c is localized on the map. Secondly the reference vector is fine tuned. The closest reference vector m_c is located by nearest neighbour method. N distances $D_j = ||x - m_j||$ are calculated to localize m_c to a feature vector x.

The m_c is minimum of

$$D_j = ||x - m_j||. \quad (2)$$

During the fine tuning phase labelled samples and a feature map of N elements are needed. The exact form of the fine tuning part is:

$$m_c(t+1) = m_c(t) + \alpha(t) * (x(t) - m_c(t)) \quad (3)$$

if $x(t)$ and the closest unit $m_c(t)$ belong to the same class,

$$m_c(t+1) = m_c(t) - \alpha(t) * (x(t) - m_c(t)) \quad (4)$$

if $x(t)$ and the closest unit $m_c(t)$ belong to different class,

$$m_i(t+1) = m_i(t) \quad (5)$$

for $i \neq c$,
where function $alpha(t)$ is decreasing function with properties:

$\sum_{t=-\infty}^{\infty} \alpha(t) = \infty, \sum_{t=-\infty}^{\infty} \alpha(t)^2 < \infty.$

The feature map determines how many regions can be found in an image. Depending on the number of reference vectors in each class it is possible to choose some reference vectors to the inner regions and the others to the boundaries.

During segmentation process the LVQ method works in a sense of nearest neighbour classifier. The image is processed in raster scan manner from upper left to lower right corner. A $n * n$ window glides over the image. The feature vectors are calculated within the window. The extracted feature vector is compared with the feature map. The classification is given by the feature map. The classification results in reference vectors on the feature map. The reference vectors belong to a class. The classmembership is transferedinto the image. All the classmemberships create the segmentation. Besides the segmentation it is possible to get more detailed information of the regions because there are more reference vectors than classes.

Classification in feature space tends to result in small regions with uncertain and rugged boundaries. This depends on the feature selection and on the size of window. To improve segmentation and region boundaries either the features or the window size had to be adjusted. There is, however, another solution too. One can adjust decision surfaces in feature space by teaching new samples. The LVQ method is capable to fine tune the class boundaries or decision surfaces in feature space. This will result in smoother region boundaries in image.

3 EXPERIMENTS AND RESULTS

The method has been first tested on simulated images. The image size has been 512*480 pixels and the whole gray level range 0 .. 255 has been used. The number of reference vectors N on the map has been 16. The feature map has been taught further by 116 samples that have been extracted from the image. The stochastic textures have been generated by published methods [Cr 83].

One test image containing four textures is shown, Figure 1. The image consists of four regions that have uniform, logarithmic gaussian, gaussian and logarithmic gaussian distributions. Grain size is less than 5 pixels.

The image has been segmented by the given segmentation procedure. The window size has been $50 * 50$ pixels. The window size has been selected due to resolution and due to representativeness of the sample. The detailed segmented image is shown in Figure 2. It can be seen that the ex-

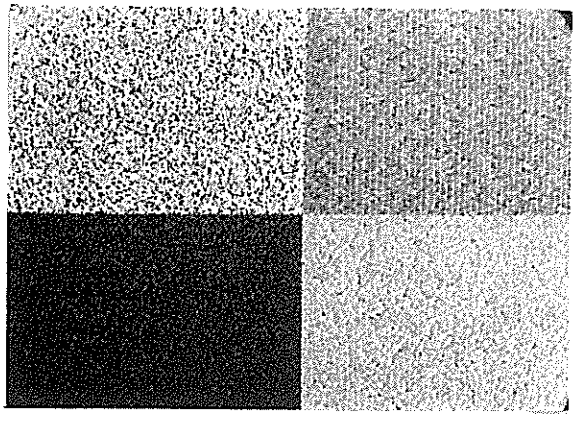


Figure 1: Test image consisting of four generated textures

pected regions have been detected relatively well. So has been the case with boundaries, too. The stochastic texture with uniform distribution has been hard. The place of the texture boundary is quite correct. There is an uncertain region of half window size around the correct border. This uncertain region can be decreased by smaller windows.

To test the suggested method on real textures some radiographs of paper formation have been captured. An image containing four captured stochastic textures have been composed, Figure 3. The four stochastic textures have been defined by a human observer.

The image containing real stochastic textures is more vivid. This can be noticed considering the segmented image, Figure 4. The regions have been detected. They are not perfect. The problem has been the inhomogeneity of real samples. The feature selection might also play a role. If regions are distinguishable considering certain features boundary detection works quite well.

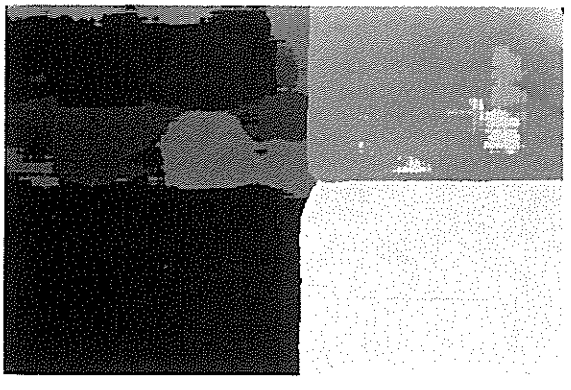


Figure 2: Generated test image after segmentation

4 DISCUSSION AND CONCLUSIONS

The reported results are mainly based on artificial images. The reason is that it is otherwise difficult to determine the exact position of region boundary. The real stochastic tex-

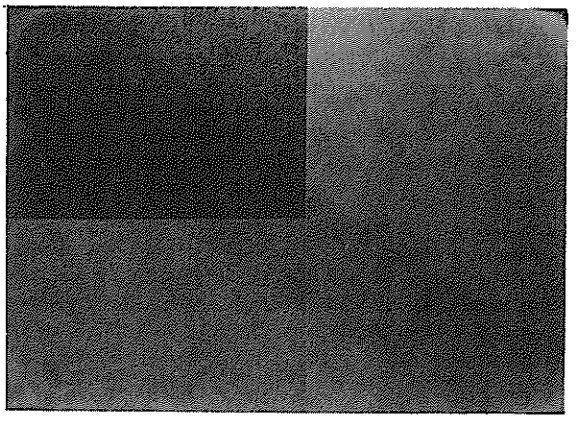


Figure 3: Test image containing four real stochastic textures

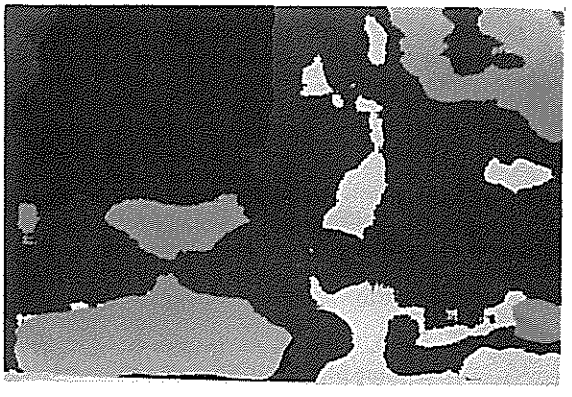


Figure 4: Test image containing four real textures after segmentation

tures are seldom homogeneous. The comparison between the detected and the known boundary is then difficult. The capability of the new method has been shown on realistic images, too. The problem with real stochastic textures is texture description. The problem is also known as feature selection problem. First the resolution of the actual texture should be chosen so that the recognition is meaningful. Then the size of the actual window should be chosen so that the extracted features differ from each other when it is necessary. This should be weighted against need of memory, computing power and uncertainty of boundary localization.

The selected window scans over the whole image which means a large amount of computations. The reported results are based on four regions. Images with eight and sixteen regions have been tested but they have needed many hours to run on a 386 based SUN workstation. This problem can be alleviated by parallel processing for which the suggested method is suitable. The problem how estimate the number of regions is not discussed here. It is, however, possible [Vi90]. The size of the feature map will determine the upper limit of detectable regions.
Several stochastic textures are taught before hand and a map is created. This map is used to classification later. The uncertainty at region boundaries can be partly controlled by the LVQ method and further teaching. The remaining problem is the size of the learning set and speed of the convergence.
To summarize a new method to segment textured images is represented. The definition and detection of region boundaries are done by segmentation. The quality of boundary detection can be controlled by a fine tuned learning procedure. The results are also promising when the method is applied to real stochastic textures.

Acknowledgement The author wishes to thank Professors T. Kohonen and O. Simula for support.

References

[Br 68] Brodatz, P., Textures: A Photographic Album for Artists and Designers, Reinhold, New York, 1968.

[Co 80] Conners, R., Harlow, C., A Theoretical Comparison of Texture Algorithms, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No 3, 1980, 204-222.

[Co 84] Conners, R., Trivedi, M., Harlow, C., Segmentation of a High-Resolution Urban Scene Using Texture Operators, Computer Vision, Graphics, and Image Processing, 25, 1984, 273-310.

[Cr 83] Cross, G., Jain, A., Markov Random Field Texture Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5 ,No 1, January, 1983, 25-39.

[De 78] Deguchi, K., Morishita, I., Texture Characterization and Texture-Based Image Partitioning Using Two-Dimensional Linear Estimation Technique, IEEE Transactions on Computers, Vol C-27, No. 8, 1978, 739-745.

[Fu 88] Fukushima, K., A neural network for visual pattern recognition, IEEE Computer, 31(3),March, 1988, 65-75.

[Ha 73] Haralick, R., Shanmugam, K., Dinstein, I., Textural Features for Image Classification, IEEE Trans. on System, Man, and Cybernetics, Vol. SMC-3, No.6, November, 73, 610-621.

[Ko 83] Kohonen, T., Self-Organization and Associative Memory, Springer-Verlag, Berlin, Heidelberg, New York, Tokio, 1983.

[Ko 86] Kohonen, T., Learning Vector Quantization for Pattern Recognition, Helsinki University of Technology, Report TKK-F-A601, 1986.

[La 89] Lampinen,J., Oja, E., Self-Organizing Maps for Spatial and temporal AR models,Proceedings of The 6th Scandinavian Conference on Image Analysis, Oulu, Finland,June 19-22, 1989, 120-127.

[Lu 89] Luttrell, S.P.,Image compression using a multilayer neural network, Pattern Recognition Letters, 10, July, 1989, 1-7.

[Ra 88] Raafat, H., A Texture Information-Directed Region Growing Algorithm for Image Segmentation and Region Classification, Computer Vision, Graphics, and Image Processing, 43, 1988, 1-27.

[We 76] Weszka, J., Dryer, C., Rosenfeld, A., A Comparative Study of Texture Measures for Terrain Classification, IEEE Trans. on System, Man, and Cybernetics, Vol. SMC-6, No. 4, April, 1976, 269-285.

[Zu 75] Zucker, S., Rosenfeld, A., Picture Segmentation by Texture Discrimination,IEEE Trans. on Computers, C-24, No. 12, 1975, 1228-1233.

[Va85] Van Gool, L., Dewaele, P., Ooterlink, A., Survey Texture Analysis Anno 1983, Computer Vision, Graphics, and Image Processing, Vol. 29, 1985, 336-357.

[Vi90] Visa, A., Identification of Stochastic Textures with Multiresolution Features and Self-Organizing Maps, in 10th International Conference on Pattern Recognition, IEEE Computer Society Press, 1990,in print.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A MODEL BASED IMAGE SEGMENTATION METHOD

Anu Langinmaa

Technical Research Centre of Finland, Graphic Arts Laboratory
Tekniikant. 3, SF-02150 Espoo
tel. Int + 358 0 4561, telefax Int. + 358 0 463848

A model based image segmentation method has been developed to segment a contact image of paper. When the segmented image is further processed the contact area size distribution of the even areas of the paper is obtained. This distribution can be used to estimate the printability of the paper.

1. Introduction

Paper roughness is defined as small scale variation of paper surface (Fig. 2). It is a structure property of paper which is of importance especially when paper is printed using the gravure printing method. In the printing press the paper is pressed against the cylinder in the print nip and it is important that the pressing surface reaches the paper surface, otherwise the print quality suffers. From the practical point of view it is thus essential to measure the paper roughness under pressure which simulates the real print circumstances, i.e., what happens in the print nip. Various measurement devices for this purpose have been developed [2]. One lack of the existing measurement devices is that the results are not processed in the signal processing sense. It is, e.g., not possible to get information about the size distribution of the even areas which is an interesting issue in the printing point of view.

2. The problem

The problem was to segment automatically a contact area image of paper surface. The image was acquired using a FOGRA/KAM surface roughness measuring device [3] shown in Fig. 1. The operation principle of the device is as follows. The specimen is pressed against a prism using a pressure which corresponds with the pressures used in printing presses. The specimen is illuminated at an angle which is greater than the angle of total reflection. When the prism is in optical contact with the paper (paper is even) diffuse reflection occurs. When there is no contact total reflection occurs. Light cell measures the amount of optical contact between the prism and the specimen and gives a reading which tells the contact percentage. The evener the paper the greater the reading. When the prism is looked from above a contact area image is seen where the even areas are light and rough areas dark. In this work a contact area image was acquired from above using a video camera and this image was further processed.

The image was digitized into 512x512 pixels. In Fig. 3 a pressure image of a newspaper specimen is shown. The area is 0.9x0.9 cm2. So one pixel corresponds to about 20 um. The dark areas which represent the rough areas are referred to as background areas and white areas as contact areas. The con-

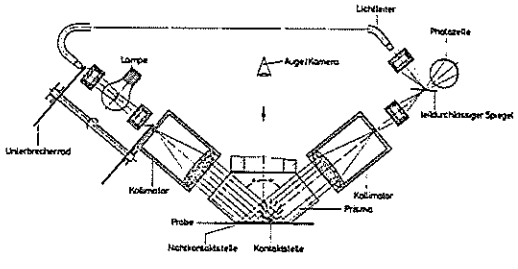


Fig. 1 FOGRA/KAM roughness measurement device

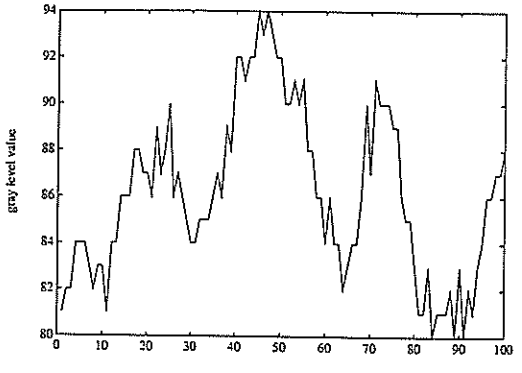


Fig. 2 Cross-section of newspaper, pressure 10 kP/cm2

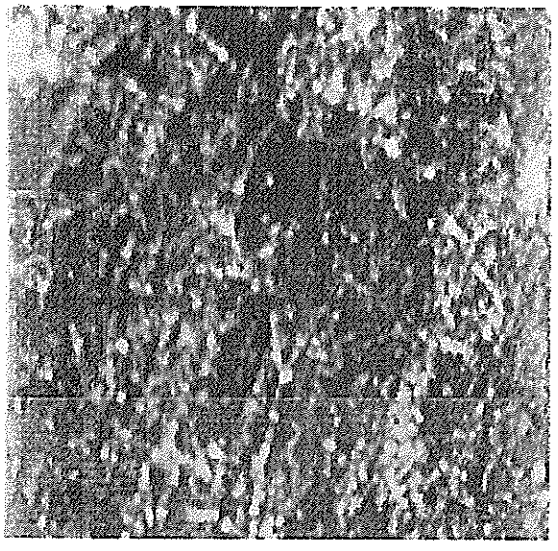


Fig. 3 Newspaper I, pressure 10 kP/cm^2

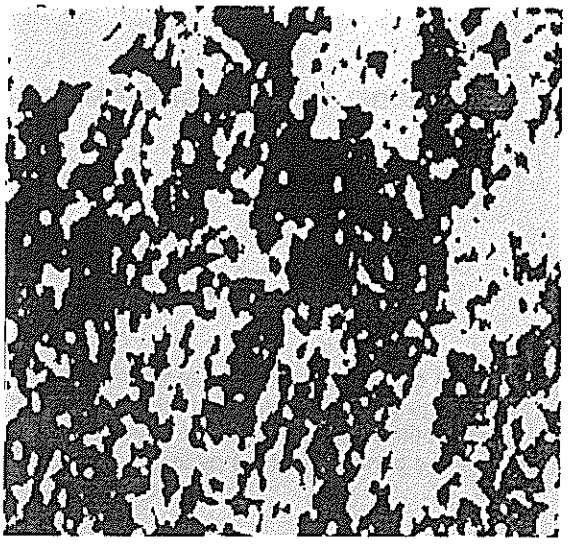


Fig. 4 Segmented newspaper I, pressure 10 kP/cm^2

trast of the image is not especially good, at most 30 gray level values as can be seen from the cross section shown in Fig. 2. By higher pressure and/or more even paper type the percentage of white areas as well as the overall gray level value of the image increases.

3. The suggested method

In Fig. 3 and 7 two typical examples of an original image are shown, in Fig. 3 newspaper and in Fig. 7 super calendered (sc) gravure paper. The applied pressure is 10 kP/cm^2. The images have been equalized to improve hardcopy quality.

The starting-point of the work was the histogram of the image. In Fig. 5 and 6 we see histograms computed for newspaper and sc paper samples. In each image several histograms are shown. In Table 1 we see four characteristic values computed from the histogram of a typical sample. The chosen characteristic values are mode, average, median and standard deviation.

Characteristic value	Newspaper I pressure, kP/cm^2 10	50	SC paper I pressure, kP/cm^2 10	50
average	84,7	102,7	81,2	106.0
median	84	84	80	104
mode	82	98	78	98
std deviation	5.2	8.8	6.0	17.1

Table 1. Characteristic values of Newspaper I and SC paper I

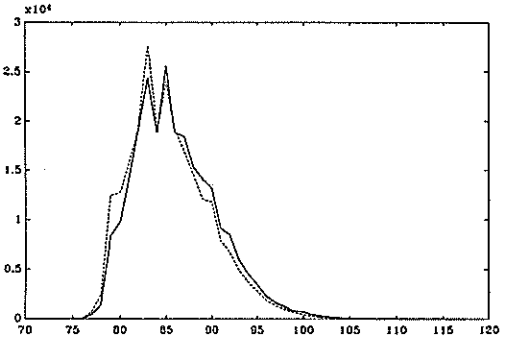


Fig. 5 Histogram of newspaper, pressure 10 kP/cm^2, 3 samples

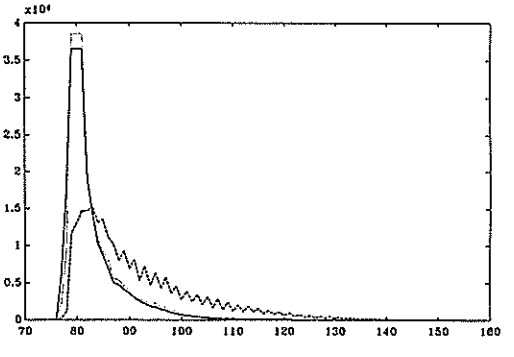


Fig. 6 Histogram of SC-paper, pressure 10 kP/cm^2, 3 samples

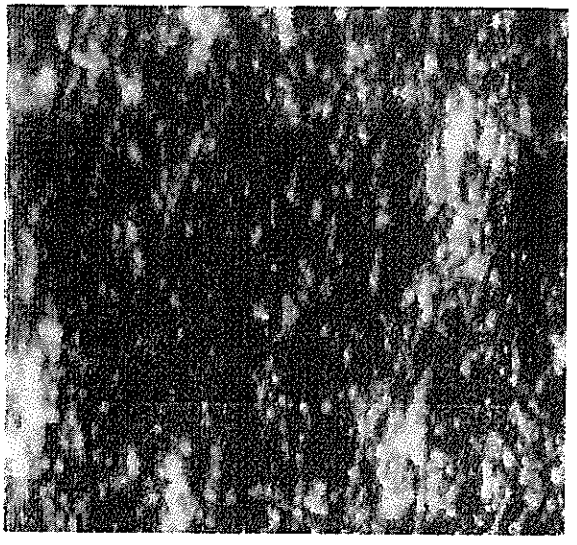


Fig. 7 SC paper I, pressure 10 kP/cm^2

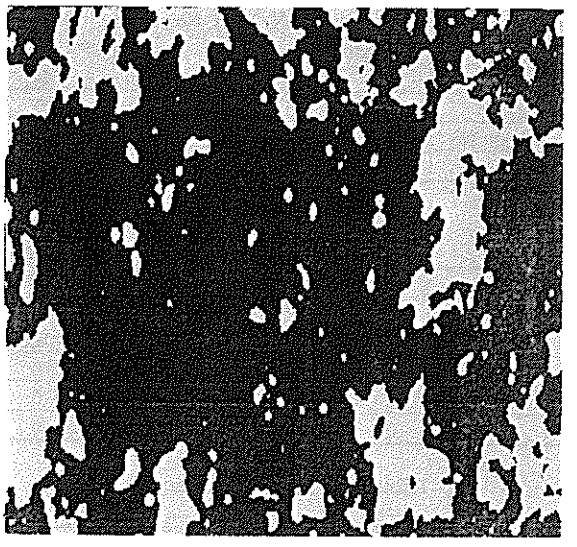


Fig. 8 Segmented SC paper I, pressure 10 kP/cm^2

Several paper types were tested and it was noticed that every paper type has a very characteristic histogram which is approximately the same for all samples. The characteristic values describe the shape of the histogram. If average, median and mode are about the same the distribution is about gaussian. If mode is much greater (smaller) than the average value the distribution is a on the right (left) tailing one. Standard deviation characterizes the size and dynamics of the contact areas. Standard deviation increases when the paper gets more even or the applied pressure is higher.

A histogram tailing on the right (Fig. 5) can be modelled as a sum of a gaussian distributed background area and contact area. In the case of a very even paper (e.g. art paper) under great pressure the tail is on the left and represents the background of the paper.

The aim of the work was to extract the contact areas from the image. Because of speed requirements thresholding on the basis of the image histogram was an appealing alternative. Unfortunately the image histograms are not bimodal. Thus standard threshold selection methods [4,5] which select a valley between two peaks in the histogram could not be applied. Some experiments to estimate the underlying background and contact area distributions to determine the threshold value using their characteristic values were carried out. However, it turned out that it was impossible to estimate the distributions reliably enough. Thus some other method had to be applied.

The solution was to suppose that the background distribution is about gaussian and estimate its' mean by the mode of the histogram. The standard deviation of the background area was estimated using the whole histogram. Thus the threshold value is

$$\text{threshold} = \text{mode} + x^*(\text{standard deviation})$$

where x is a confidence interval coefficient which depends on paper type and pressure.

The hypothesis was tested by first thresholding a small amount of samples of each specimen by hand and determining the appropriate confidence interval coefficients. It turned out that in practice x was 0.0-1.0. The value of x for a given paper type and pressure was about constant. Then totally 158 samples were thresholded using the acquired confidence interval values and the results were compared to the result of thresholding by hand. The samples consisted of newspaper (two types), sc gravure paper (two types) and light weight coated paper (LWC paper). Each automatic thresholding was a given a mark which corresponds to its' quality. The results are shown in Tables 2, 3, 4 and 5.

In Fig. 4 and 8 we see the newspaper and sc gravure paper images of Fig. 3 and 7 thresholded using the suggested method. The thresholded images have been lowpass filtered using a 3x3 median filter to get rid of insignificant details. The region size distribution of the even areas in Fig. 4 and 8 are shown in Fig. 9 and 10. The region size distribution has been obtained by labelling the thresholded and median filtered image and then computing the size of each region in "square pixels".

quality	Newspaper I, N=39 pressure, kP/cm^2			Newspaper II, N=24 pressure, kP/cm^2		
	10	15	50	10	15	50
good	9	7	6	11	9	0
acceptable	4	6	6	1	3	0
bad	0	0	1	0	0	0

Table 2. Newspapers I and II

quality	SC paper I, N=26 pressure, kP/cm^2			SC paper II, N=36 pressure, kP/cm^2		
	10	15	50	10	15	50
good	10	10	0	15	10	0
acceptable	3	1	0	3	8	0
bad	0	2	0	0	0	0

Table 3. SC papers I and II

quality	pressure, kP/cm^2		
	10	15	50
good	8	7	5
acceptable	3	3	4
bad	0	1	2

Table 4. LWC paper, N = 33

quality	pressure, kP/cm^2		
	10	15	50
good	53	43	11
acceptable	14	21	10
bad	0	3	3

Table 5. Total, N = 158

4. Discussion and conclusions

Model based image segmentation techniques have been applied to contact area extraction of paper.

The image acquiring and digitizing was performed using a DT2851 image processing card. The computations were carried out using a ALR/386 computer equipped with mathematical coprocessor.

The method has been applied to characterize newspaper, sc gravure paper and LWC paper. The applied pressures have been 10, 15 and 50 kP/cm^2.

The method turned out to be fast. It is also reliable enough for newspaper . It does, however, not always work correctly when applied to sc gravure and LWC paper. The obtained contact area size distribution is of importance when the printability of the paper is concerned.

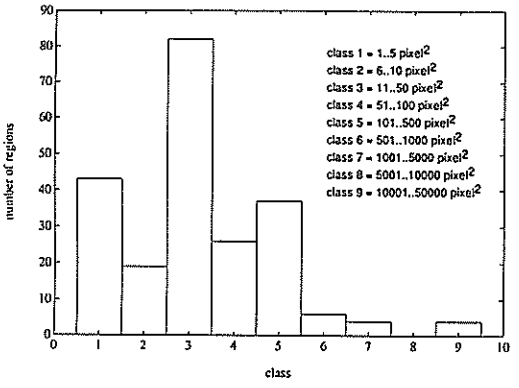


Fig. 9 Contact area size distribution of Fig. 4

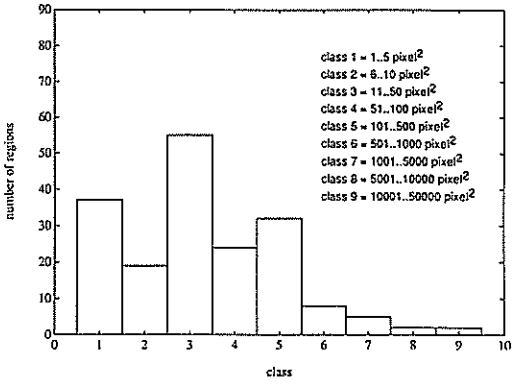


Fig. 10 Contact area size distribution of Fig. 8

5. Acknowledgments

This work has been funded by the Technology Development Centre of Finland.

6. References

[1] J.A. Bristow and P. Kolseth, Paper Structure and Properties, Marcel Dekker Inc. 1986.

[2] W.I.Wilt, Paper testers for instaneous measurement of smoothness and porosity. Tappi 39, 1956.

[3] FOGRA/KAM, Kontaktanteilmessgerät. Operating instructions.

[4] M. Haralick and L.G. Shapiro, Image Segmentation Techniques, Computer Vision, Graphics and Image Processing 29, 1985.

[5] J.S. Weszka, A Survey of Threshold Techniques. Computer Graphics and Image Processing 7, 1978.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

STUDY OF STONES BY IMAGE PROCESSING

R. Harba*, G. Jacquet* and M. Rautureau**

* Groupe de Recherche sur l'Etude des Milieux Ionisés
** Laboratoire de cristallographie

U.F.R. faculté des sciences
rue de Chartes BP 6759
45067 Orléans CEDEX 2 FRANCE

ABSTRACT :

This paper describes a method to extract two basic components of a stone from three images of the same scene made by scanning electron microscopy. This is realized by taking into account prior information that is to say that these three images would be superposable as a jigsaw puzzle for an ideal case. So, a discriminating fonction D(T) which represents the reliability of the superposition is defined versus a vector T whose components are the thresholds of the three images. It is planned to minimize D(T) in order to yield the best vector T. Our results are compared with classical ones and also with extraction made by an expert.

INTRODUCTION :

Up to now, the primary expertise of historical monuments is essentially visual, and it cannot be either reproduced or even quantified. So, the failure to recognize some parameters related to composition, porosity and anisotropy of stones does not permit a correct evaluation of their growing old in order to decide when, why and wherefore to restaure it.

A very precise knowledge of stones is necessary to evaluate some important parameters. For example, silice particles agglomerate to calcite ones and compose the cement. Therefore, silice proportion reflects the strength of the stone. Chemical and physical reactions of the stone to water depend on the quantity and the shape of porousness.

Quantitative studies are made. For example, a chemical analysis gives the proportions of the various components. Another solution is to observe a thin section of stone by electronic microscopy [1] : a spectral analysis also gives these proportions. But, only an image processing leads to information on the shape and on the relative position of the components. The present study reports preliminary investigations of such a processing on very porous stones (about fifty per cent of porousness) called tuffeau often used in the past for building monuments which are mostly composed of siliceous and calcareous particules.

MICROSCOPIC STONE IMAGES :

Samples are made of thin sections of stones of a few square centimeters large and thirty micrometers thick, where porousness is filled up by synthetic resin : induration is necessary because the tuffeau is a polyphased material with little coherence. Such petrographic sections are observed by scanning electron microscopy, yielding a polaroïd image (fig. 1a) in which the secondary and backscattered electrons are analysed. This good quality image is called porousness image and is composed of a lot of particles :

i) the very black ones are calcite, little calcareous fossils and siliceous particles which are quartz.

ii) the medium grey ones are quartz or mica principaly made of silicium.

iii) the small and clearer ones are silice or clay (siliceous particle).

iv) A very few of them (about one per cent) are iron oxydes or iron sulfurs.

v) porousness is the white parts of the image and is easilly separable from the particles.

It seems difficult to separate all these particles by only taking into account the porousness image. In spite of this, it can be expected from the histogram to be composed of four modes. But acquisition problems do not allow such a perfect case (fig. 1b). As seen, only the white distribution which is porousness is separated from the three others. A correct

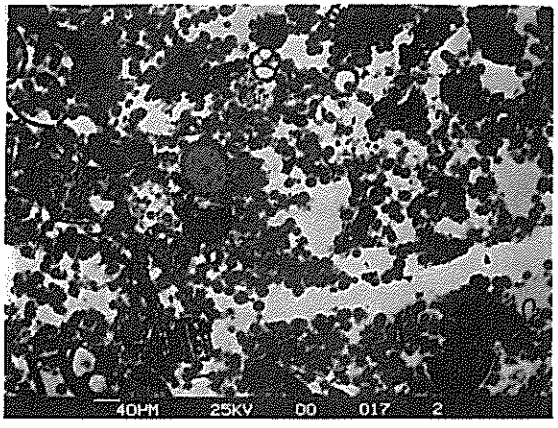


figure 1a : Porousness image and ...

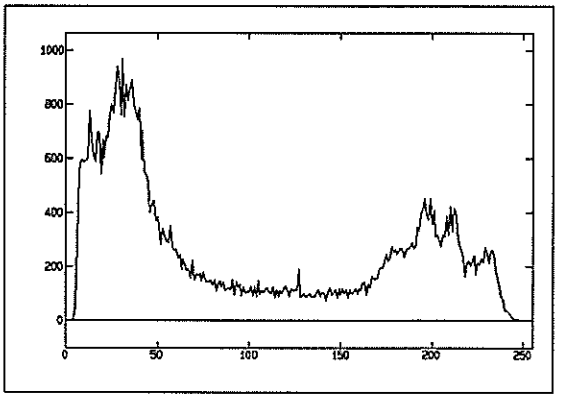


figure 1b : Histogram of the porousness image.

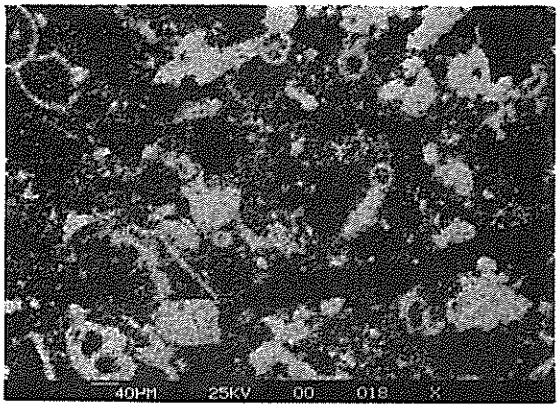


figure 2a : X-ray calcium image and ...

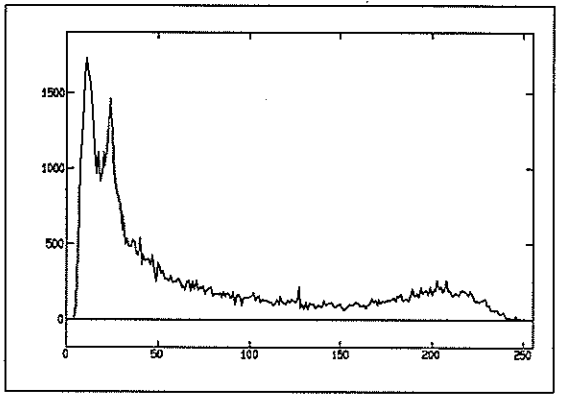


figure 2b : Calcium histogram.

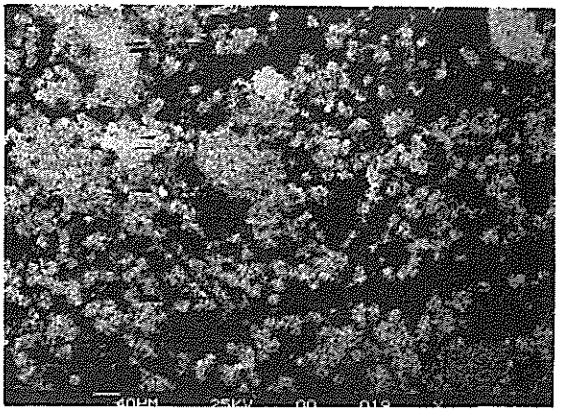


figure 3a : X-ray silicon image and ...

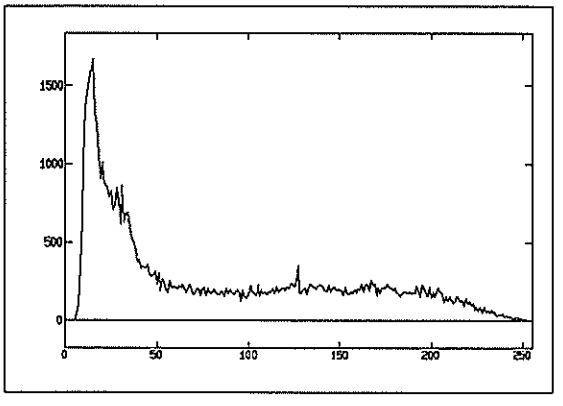


figure 3b : Silicon histogram.

extraction must rely on different information.

Scanning electron microscopy gives extra indications by analysing photons. From the same section, two other noisy X-ray images are available :

i) an X-ray calcium image (fig. 2a) where the Kα ray is taken into account. Particle with calcium are in white, everything else in black.

ii) an X-ray silicon image (fig. 3a) where the siliceous particles are in white.

The interesting thing is that mostly all the particles of the stone are composed either of silicon or of calcium with different concentration. So, they appear only on one image at a time and permit to separate the siliceous particles from the calcareous ones.

As the information of the electron microscope are not directly converted in digital images, the polaroid photos must be well adjusted one to the others during acquisition. In addition, technical difficulties for lighting and for the adjustement of the camera induce a lost of information.

CLASSICAL THRESHOLD VERSUS EXPERT THRESHOLD:

As seen in the three histogramms, the average grey level of particles and background are different enough. So, a global threshold technique can be used [2][3] for each image. The method used is explained in [4]. Each component of the mixture is supposed to be normally distributed by the mean of μ, a standard deviation σ and a *a priori* probability P. First, an initial threshold vector is crudely obtained from smooothed histograms. Each component is given by the position of the minimum between the two maxima of the related histogram. Then, an updated threshold is computed according to the Bayes minimum error rule by an iterative method. This is done separately on the three images and this gives a T vector with three components : Tp for porousness image, Tc for calcium image and Ts for silicon one. The T vector is (88,58,53).

For comparison, partition of the three images is made by an expert and the T vector is equal to (104,144,130).

Tp is quite well but Tc and Ts are not. This is due to the gaussian hypothesis on the histograms and spread of particles in X-ray images : they appear bigger than they are [5]. This important experimental artefact disturbs the method and another approach seems necessary.

IMAGES SUPERPOSITION :

As told before, the three images are issued from the same scene and each image is only related to a particular component : porousness, calcium and silicon. If the extraction was perfect, these images would be superposable as a jigsaw puzzle. This does not suppose any sort of distribution of the histograms. Then, a criterion related to the reliability of this complementarity will be explained. In this purpose different levels are affected to the two components of binarized images :

i) porousness image : threshold Tp, level 1 for porousness and level 0 for the background.

ii) calcium image : threshold Tc, level 2 for calcium, level 0 elsewhere.

iii) silicon image : threshold Ts, level 4 for silicon, 0 for the background.

The quality of the superposition is tested by making an addition of the three binary ones. It is planned to find the T vector leading to the best attainable complementarity.

In an ideal case, the result will be an image with only three levels corresponding to well classified pixels :

i) Np pixels for porousness at level 1.

ii) Nc for calcium (level 2).

iii) Ns for silicon (level 4).

In the actual case, in addition to the three well classified pixels, five badly classified ones occur at intermediate levels. These are:

i) No at level 0 : pixels belonging to no component.

ii) Npc at level 3 : those belonging to porousness and calcite.

iii) Nps at level 5 : porousness plus silicon.

iv) Ncs at level 6 : calcium plus silicon. It should be noticed that this class includes a very few calcium silicate particles which are well classified. It will be easy to display them once the extraction will be performed.

v) Npcs at level 7 : porousness plus calcite plus silicon.

Most of these badly classified pixels appear at the boundaries of the different particles but silicate of calcium and particles whithout the analysed components can be seen.

PROCESSING METHOD :

The noise of the grey level X-ray images is lowered by a median filter of size 3 by 3. To find the optimal threshold T = (Tg,Tc,Ts), use is made of a discriminating function D(T) which represents the reliability of the superposition. It may be the sum of the badly classified pixels :

$$D(T) = No + Npc + Nps + Ncs + Npcs$$

To place this optimization problem without constraint[4], D(T) is represented figure 4a and 4b for a given Tg and for Tc and Ts varying from 56 to 251 step 13. As seen, the shape at this scale seems to present only one minimum. But with a step one and close to the minimum, local minima appear. The chosen method must take this into account and in our case, gradient and simplex methods have been tested [6].

PRACTICAL RESULTS AND COMPARISONS :

The minimum of D(T) has been found equal to 8754 and in an automatic implementation, D(T) is a good evaluation of the partition quality. The T vector equal to 148, 140 and 64. Expert, classical and optimization results are compared

in figure 5a and 5b. As seen, badly classified pixels are minimum for our method and the T vector is close to the expert appreciation for X-ray images but not for the porousness one. This is due to particles which are not siliceous neither calcareous. The expert recognizes these particles and choses the good threshold. A similar approach can be taken. The first one is to make iron and sulphur X-ray images and every particle will be analysed. An other way is to extract these particles using mathematical morphology ; this method is under active investigation.

Afterwards, the final work is to assign badly classified pixels to one of the three components. For this purpose, the neighbourhood of each badly classified pixel is watched in each of the three binary images and this pixel is classified.

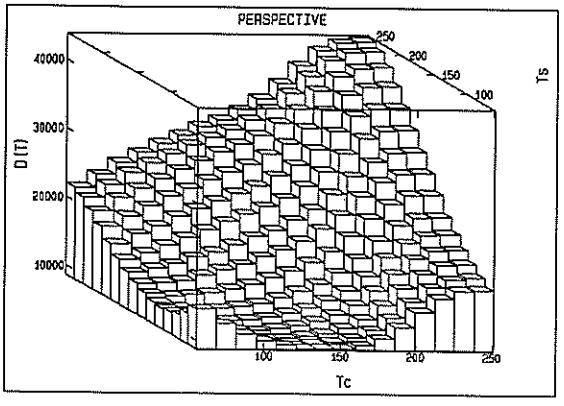


figure 4a : Tridimentional shape of D(T) versus Ts and Tc for a given Tp (147).

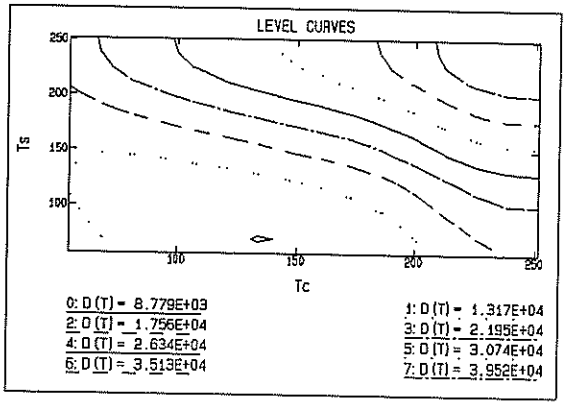


figure 4b : Level curves of D(T) versus Ts and Tc for a given Tp (147).

	Tp	Tc	Ts	D(T)
Expert	104	144	130	9142
Baye. Min.	88	58	53	21166
Our method	148	140	64	8754

figure 5a : Comparison of the three methods. D(T) is the sum of the bad classified pixels.

	porousness	calcium	silicon
Expert	25482	12120	18892
Baye. Min.	18447	11907	14006
Our method	19337	11936	25509

figure 5b : Comparison of the three methods. The number of pixels of each class is given.

CONCLUSION :

This paper describes a method to extract particles of a stone using complementary images of a same scene by taking into account the superposition of these images. This method gives the best result in the term of superposition and does not suppose any kind of distribution of the histograms. This method can be improved taking into account the few particles whithout calcium or silicon.

REFERENCES :

[1] D. Jeulin " Mathematical morphology and material image analysis " ed Scanning Microscopy International vol 2 p 165 184 (1988).

[2] Y. Nakagawa and A. Rosenfeld " Some experiment on variable thresholding " Pattern recognition vol 11 pp 191 204 (1979).

[3] P.K. Sahoo, S. Soltani A.K.C. Wong and Y.C. Chen " A survey of thresholding techniques " Computer vision, graphics and image processing 41 pp 233 260 (1988).

[4] J. Kittler and J. Illingworth " Minimum error thresholding " Pattern Recognition Vol. 19 pp 41 47 (1986).

[5] J.P. Eberhart " Méthodes physiques d'étude des mineraux et des matériaux solides " Doin éditeurs (1976).

[6] R. W. Daniels " An introduction to numerical methods and optimization techniques " North-Holland editor (1978).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Texture Synthesis Using Nonhomogeneous Gaussian Markov Random Fields Model

Zou Cairong Wang Taijun He Zhenya

Dept. of Radio Engineering, Southeast University, Nanjing , P.R.China

This paper is mainly concerned with the texture synthesis using nonhomogeneous Gaussian Markov Random Fields (GMRF) Model. The existing texture synthesis method using MRF model is discussed. Two aspects of MRF are considered to synthesize the texture. Suppose the autocorrelation function of the texture image is exponentialy distributed, we get simple recursive form of the algorithm.A more general texture synthesis method using Gibbs distribution is also described.

1. INTRODUCTION

Texture synthesis and texture analysis are important contents of computer vision, remote sensing and image analysis [1-3]. We can segment the natural scenes using texture measure. For example , we can discriminate among rivers, grass lands and streets from an aeronautical image.The surface orientation and depth can also be computed according to the texture change. For texture,there is no precise definition. Generally, we think of texture as something that is random, periodic or has some definite structure and placement. There are mainly two methods in the study of texture analysis, they are statistical estimation and structure description. Because texture covers the entire image, the statistical models become a powerful method in both texture synthesis and texture analysis. Markov Random Field model has found wide application in the field of image processing and image analysis, such as image restoration, image segmentation, image smoothing and texture analysis [4-8]. Among them ,texture synthesis and texture analysis using MRF model attracted many researchers. There are mainly two methods in the study of texture analysis using MRF. One is proposed by R.L. Kashyap et al.[9], they use Gaussian MRF model because it results in a diffirence equation in spatial domain. They derived a texture synthesis algorithm which use the Fast Fourier Transform , we must choose the optimal neighbors when we analyze the texture. Another is used by A.K. Jain and G.C.Cross[10]. They compute a local conditional probablity according to the binomial distribution. A neighborhood system model is used. Using Metropolis algorithm they obtain many different texture. D.Geman and S. Geman point out the method used by A.K. Jain and G.C. Cross is really one example of Gibbs Distribution[4]. But untill now, there is no report that apply Gibbs Distribution to the synthesis of texture. In this paper we describe a fast texture syhthesis method . We fist discusse the one dimensional case, then the two dimensional case.The homogeneous and nonhomogeneous case are applied to produce the texture. We synthesize texture using nonhomogeneous random fields because there exist a lot of nonhomogeneous natural texture. We also describe a texture synthesis algorithm which use Gibbs Distribution, three kinds of optimization methodws (ICM, SA, DP) can be used.

2. TEXTURE MODEL

We first discuss one dimesional case. The local probablity expression for 1th Markov property is

$$Pr\{y(n)\,|\,\text{all } y(m), m=n\}=Pr\{y(n)\,|\,y(n-1)\}$$

Where y(n) is one dimensional random signal. If y(n) is also Gaussian distributed we can obtain the equation as follows

$$R(t_1,t_3)*R(t_2,t_2)=R(t_1,t_2)*R(t_2,t_3)$$

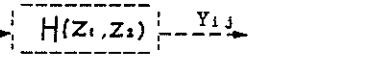
Where $R(t_i,t_j)$ is autocorrelation functon of y(t). Assuming y(t) is stationary, that is to say , $R(t_i,t_j)$ depends on the difference (t_i-t_j) only. It is easily shown that

$$R(t_i,t_j)=K*\exp\{a(t_i-t_j)\}$$

This equation is an equivalent expression of 1th order Gaussian Markov process .Extending this one-dimensional case to two-dimensional case, we have

$$R(t_{mn},t_{kl})=A_mA_nB_kB_l*\exp\{a(m-k)+b(n-l)\}$$

If y_{ij} is homogeneous we can use Z-transform to get the sample function of a linear system which outputs the signal y_{ij} while the input sinal is i.i.d. Gaussian white noise N_{ij}.



For stationary homogeneous case we have

$$Z\{R(x,y)\}=H(z_1,z_2)*H(z^1,z^2)*Var$$

Where Var is the variance of input white noise, $H(z_1,z_2)$ is Z-transform of sample function.It is easily shown that

$$Z\{(R(x,y)\}=\frac{Var(1-M^2)(1-N^2)}{(1-MZ_1)(1-NZ_2)(1-MZ_1^{-1})(1-NZ_2^{-1})}$$

Where M=exp(-a),N=exp(-b),and we select

$$H(z_1,z_2)=\frac{1}{(1-MZ_1^{-1})(1-NZ_2^{-1})}$$

and

$$Var=K(1-M^2)(1-N^2)$$

Thus we the random value y_{ij} is the combination of neighboring elements and i.i.d. white noise. It is a unilateral Gaussian Markov random filed. The bounded recursive procedure is

$$Y(1,1)=W(1,1)$$

$$Y(i,1)=MY(i-1,1)+W(i,1)\qquad 1<i<NL+1$$

$$Y(1,i)=NY(1,i-1)+W(1,i)\qquad 1<i<ML+1$$

$$Y(i,j)=MY(i-1,1)+NY(i,j-1)-MNX(i-1,j-1)+W(i,j)\qquad 1<i<NL+1,\ 1<j<ML+1$$

Where NL and ML are image lattice size. It is obvious that the resulted bounded texture image is both Gaussian and Markovian.The texture synthesis results using homogeneous GMRF model are illustrated in Fig.1. For nonhomogeneous case we first write $Y(j)=\{y_{1j},y_{2j},y_{3j},\ldots,y_{nj}\}$.Since the random field is Markovian, we suppose the two dimensional signal y_{ij} is computed according to a recursive procedure. So it is an unilateral Markov random field model .That is to say Y(j) can be calculted form the combination of Y(j-1) and W(j), where $W(j)=\{w_{1j},w_{2j},w_{3j},\ldots,w_{nj}\}$. A recursive method that extends the one dimensional case to two dimensional case is used. Fortunately, there exists algothm for the generation of two- dimensional random signals. E.Hryniewicz proposed the numerical generation method for two dimensional Gaussian random fileds[14]. The recursive procedure is stated as follows

$$Y(1,1)=a_1b_1W(1,1)$$

$$Y(i,1)=a_i/a_{i-1}MY(i-1,1)+a_ib_i(1-M)W(i,1)\qquad 1<i<NL+1$$

$$Y(1,i)=b_i/b_{i-1}NY(1,i-1)+a_1b_iW(1,i)\qquad 1<i<ML+1$$

$$Y(i,j)=a_i/a_{i-1}MY(i-1,1)+b_j/b_{j-1}NY(i,j-1)-a_ib_j/(a_{i-1}b_{j-1})MNY(i-1,j-1)+a_ib_j\ 1-M\ 1-N\ W(i,j)\qquad 1<i<NL+1,\ 1<j<ML+1$$

The simulation result using nonhonogeneous GMRF model is illustrated in Fig.2. We can also set other forms of autocorrelation function so as to include more nerghbor pixels. Different selection of the autocorrelation function results in different kinds of texture.

3. GIBBS DISTRIBUTION

Gibbs Distribution(GD) is an equivalent description method of Markov random fields. It is firstly used in the field of statistical physics. In the GD the local probability measure p{w} on sample space is given by

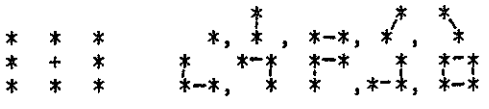
$$p\{w\}=\exp(-U(w)/T)/Z$$

Where U(w) is called energy function , T is contant and Z is partition function. The energy function with respect to configration w is

$$U(w)= \quad Vc(w)$$

where denotes the all the cliques, and Vc(w) is potential fuction which is a nonlinear combination of clique elements. The different choice of neighborhood system results in diffrent associated cliques. The cliques corresponding to 1th and 2nd neighborhood system are illustracted as follows

1th neighborhood cliques



2nd neighborhood cliques

A clique is defined from a neighborhood system. The first clique class has only one pixel. The other cliques have pixels which is certainly in a same neighbor. The energy function associated with cliques is expressed as

$$U\{w\}=\sum_{i,j}[Aij\ Yij+\sum_{k,l}BijYijYkl+.....]$$

The coefficients Aij ,Bij et al is arbitrary .Based on the local neighbor we can define these coefficients. For example, if the pixel values in a neighbor are also the same we define Aij=C1, or we define Aij=C2. Actually when we use Gibbs Distribution we always adopt simple cliques which have pixels not more than two. If the image is binary and cliques with only one pixel is applied it is easily shown the local probality measure will have the same form as A.K.Jain and G.C.Cross give in their paper[10]. So we have an idea that texture synthesis is carried out using Gibbs Distribution. We maxmize the local probablity measure p{w} directly, and this is equivalently minimize the energy function U{w}. From another point of view, we think of texture as a realization of a system which reachs to minimal energy. It is rule that any system will tend to have less energy if there is no external power acting on it. Three kinds of techniques could be used to minimize U{w}, these methods are ICM algorithm proposed by Besag[12] ,Simulated Annealing algorithm which is similar to the Stochstic Relaxtion used by G.Geman and S.Geman[4] and Dynamic Programming. By the selection of parameters Aij ,Bij or even Cij we obtain different texture results.

Here we put forward a texture synthesis scheme using Gibbs Distribution.The parameter estimation in Gibbs Distribution will be reported in another paper.

4. NUMERICAL SIMULATION

We do the computer simulation according to the recursive procedure .The homogeneous and nonhomogeneous cases are included. For the first case, we list the coefficients chosed in the program in Tab.1. For the another case we choose the same coefficients while the A(i) and B(i) take the forms of

$$A(i)= 0.5*(255-i)/127$$

$$B(i)= 0.5*(255-i)/127$$

TAB 1. Parameters(M,N)

(0.1, 0.7)	(-0.1, 0.89)
(0.068,0.327)	(0.7834,0.012)

5. CONCLUSIONS

In this paper we use the homogeneous and nonhomogeneous GMRF model to carry out the synthesis of texture. Numerical simulation shows that texture synthesis using recursive procedure listed in this paper is effective. We also present a general texture synthesis scheme using MRF model. The parameter estimation in GMRF model and GD model will be reported in other papers.

REFERENCES

[1] P.Bouthemy and P.Lalande, " Motion Detection in An Image Sequence Using Gibbs Distributions", Proc. of 1989 Int. Conf. on ASSP., pp.1651-1654.

[2] J.W. Woods, " Two-Dimensional Discrete Markovian Fields",IEEE Trans. on Information Theory, Vol.IT-18, No.2, March 1972, pp.101-109.

[3] H.Derin, H.Elliott, R.Cristi and D. Geman,"Bayes Smoothing Algorithms for Segmentation of Binary Images Modolled by Markov Randim Field", IEEE Trans. on Pattern Analysis and Machine Intelligence,Vol.PAMI-6, No.2, Nov. 1984, pp.707-720.

[4] S.Geman and D. Geman, " Stochastic Relaxtion, Gibbs Distributions and Bayesian Restoration of Images", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-6, Nov. 1984, pp.721-741.

[5] T.Simchony and R.Chellappa,"Stochastic and Deterministic Algothms for MAP Texture Segmentation", Proc. of 1988 Inter. Conf. on ASSP, pp.1120-1123.

[6] H. Hassner and J. Sklansky, " The Use of Markov Random Fields as Models of Texture", Computer Graphics and Image Processing , Vol.12,1980,pp.357-370.

[7] R. Chellappa and. R.L. Kashyap," Digital Image Restoration Using Spatial Interaction Models", IEEE Trans. ASSP.,Vol.ASSP-30,June 1982, pp.461-472.

[8] M. El-Gabali, M. Shridhar and M. Ahmadi, " Segmentation of Noisy Images by Markov Random Fields with Gibbs Distribution", Proc. of 1987 Int. Conf. on ASSP., pp.551-554.

[9] R.Kashyap and R. Chelappa,"Estimation and Choice of Neighbors in Spatial Interaction Model of Imagess",IEEE Trans.on Information Theory, Jan. 1983,pp.60-72.

[10]G.C. Cross and A.K. Jain," Markov Random Field Texture Models", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.PAMI-5, 1983,pp.25-39.

[11]F.Schmitt and D.Massaloux,"Texture Synthesis Using a Bidimensional Markov Model", Proc. of Int. Conf. on ASSP. pp.593-596.

[12]J.Besag, "On the Statistical Analysis of Dirty Pictures", Journal of Royal Statistic Society B, Vol.48, No.6, 1986,pp.256-302.

[13]R. Cristi , M. Shridhar and M. V. Prasadarao, " Segmentation of Multilevel Images Using Gibbs Distribution", Proc. of 1985 Int. Conf. on ASSP. pp.901-904.

[14]Z.Hryniewicz," Numerical Generation Methods of Homogeneous and Nonhomogeneous Two-Dimensional Gaussian Random Fields", SIAM Journal of Applied Mathematics, Vol.39, No.1, August 1980, pp.169-172.

[15]C. Srinivas and M.D. Srinath, " Compound Gauss Markov Random field Model for Image Segmentation and Restoration", Proc.of 1989 Int. Conf. on ASSP., pp.1586-1589.

[16]H. Derinand H. Elliott, " Modelling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.PAMI-9, No.1,1987, pp.39-55.

[17]C.S.Won and H.Derin, " Segmentation of Noisy textured Images Using Simulated Annealing", Proc. of 1987 Int. Conf. on ASSP,pp.563-566.

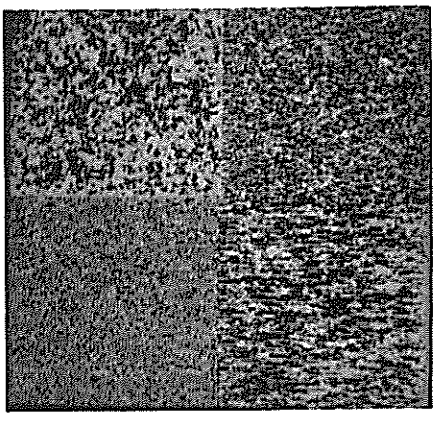


Fig.1 Texture Synthesis Reselts

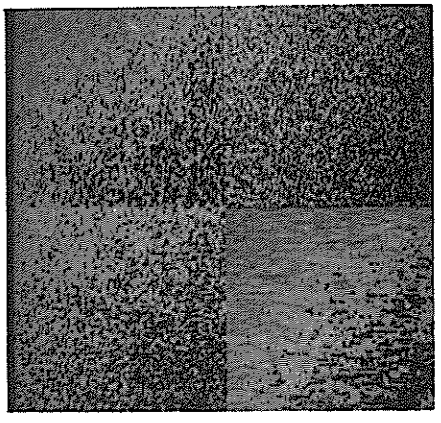


Fig.2 Texture Synthesis Results

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

CHARACTERIZATION OF EXTRUDED PRODUCTS USING TEXTURE ANALYSIS METHODS.

J.SEROT[1] - S.LELANDAIS[1] - D.BERTRAND[2] - P.ROBERT[2]

(1) LRII/ATI - IRESTE - CP 3003 -
44087 NANTES CEDEX 03 - FRANCE - Tel: 40 68 30 00.
(2) Laboratoire de Technologie Appliquee à la Nutrition
INRA - rue de la Geraudiere -
44072 NANTES CEDEX 03 - FRANCE - Tel: 40 67 50 00.

This paper deals with three image processing and texture analysis methods to discriminate some agronomic food products obtained by puffing extrusion process. First we describe the products, texture analysis methods and our choices for this work. Two methods using a statistical description of the texture are tested. One is the "grey level run lengths method" and the other is the "local histogram method". Then we test a third method using a structural description of the texture. Finally we present results and conclusion of this work.

INTRODUCTION

Extrusion process has been used from fifty years to product food like biscuits, sweets, breakfeast cereals, soups,....This technic is quick and simple, but the process isn't easy to control and the model is complex. We project to obtain some features of the product at the end of the process and to search correlation between these features and extrusion cooking conditions (temperature, pressure, hydration rate, exit flow,...). So we try to propose a closed loop regulation of the process. The problem is: can we obtain, by image analysis, some features of extruded products allowing to discriminate different parameters values of extrusion process ?

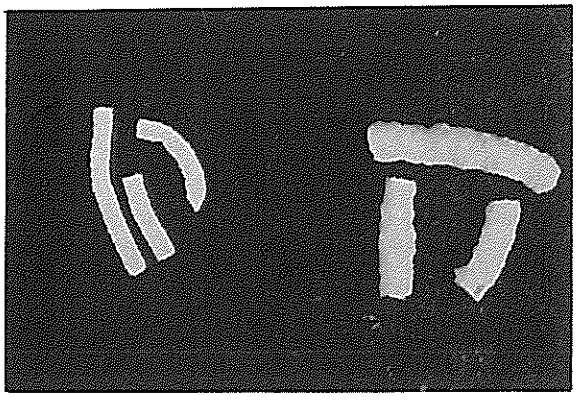


Figure 1

On figure 1, we present a photography of the product on which we work (three samples from two different classes). It looks like white "sausages" of 5 to 10 millimeters in diameter. We make a longitudinal section at half diameter and we observe it under a microscope (figure 2). The product consists of dark alveoles and a white compact substance. Size, number and organization of alveoles are typical of extrusion process parameters values. We have chosen to analyse the texture. In fact the image on figure 2 is a macroscopic texture and we want to know the characteristics of elementary primitives and primitives disposition laws.

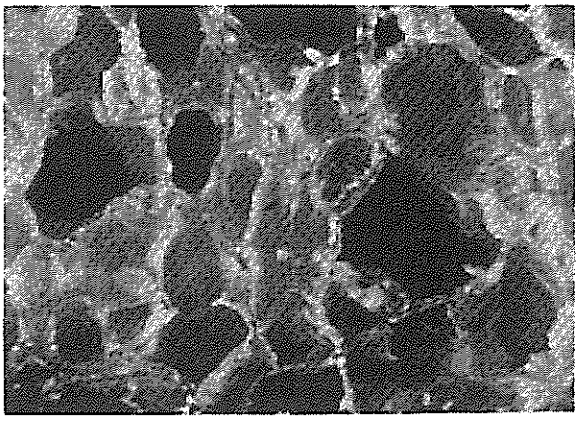


Figure 2

Many works have been published in the field of texture analysis {HAR-73,HAR-79,GAG-83} but results examples are very often given for microscopic textures. Classical methods are frequential analysis, spatial analysis (autocorrelation technic, mathematical morphology, gradient methods, linear transform, ...) second order statistics analysis (co-occurence matrix) and structural analysis.
We have choosen three methods to extract characteristics of this macroscopic texture. Two of them use spatial analysis: grey level run lengths and local histogram. The third gives a feature vector with first and second order statistics. It's a structural analysis of primitives which uses the concept of generalized co-occurence matrix to obtain the second order statistics.
Before we detail these three methods and theirs results, we introduce study conditions. We have seven classes of extruded products. For each class, the process parameters are typical. We take four samples of each class. We make a longitudinal section of each of the 28 samples. The flat side of the section is placed under a binocular lens. A CCD camera is on the lens. For the lighting we use two optical fibers to obtain a low-angled light. With this lighting we can see the edge of the alveoles and their depth.

The acquisition card is on a "PC like" micro-computer. The images have 512 by 512 pixels, each pixel is coded on 8 bits. For our application we reduced the size of the images to 256 by 256 pixels.
The three methods are programmed in C language.

ANALYSIS BY GREY LEVEL RUN LENGHTS

A grey level run {GAL-75} is a set of maximal collinear connected pixels having the same grey level. Each grey level run can be characterized by its grey level, its length and its direction. For a given picture, and using each of the 3 main directions ($\theta = 0°$, 45° and 90°), we can compute a grey level run length matrix L_θ in which $L_\theta[i,j]$ specifies the number of times a run of length j and grey level i appears in the picture. Let Ng be then number of grey levels and Nr be the total possible number of runs (so that the matrix is NgxNr). Five parameters are computed to summarize the information contained in each matrix L_θ:

+ Short Run Emphasis:

$$\sum_{i=1}^{Ng} \sum_{j=1}^{Nr} L_\theta[i,j]/j^2 \;/\; \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} L_\theta[i,j]$$

+ Long Run Emphasis:

$$\sum_{i=1}^{Ng} \sum_{j=1}^{Nr} j^2 * L_\theta[i,j] \;/\; \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} L_\theta[i,j]$$

+ Grey Level Non-Uniformity:

$$\sum_{i=1}^{Ng} \left(\sum_{j=1}^{Nr} L_\theta[i,j]\right)^2 \;/\; \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} L_\theta[i,j]$$

+ Run Length Non-Uniformity:

$$\sum_{j=1}^{Nr} \left(\sum_{i=1}^{Ng} L_\theta[i,j]\right)^2 \;/\; \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} L_\theta[i,j]$$

+ Run Percentage:

$$\sum_{i=1}^{Ng} \sum_{j=1}^{Nr} L_\theta[i,j] \;/\; \sum_{i=1}^{Ng} \sum_{j=1}^{Nr} L_\theta[i,j]$$

ANALYSIS BY LOCAL HISTOGRAM

The aim of this method {LOW-83} is to extract textural information from grey-level histograms computed on squared neighbourhood centered around a pixel. Each local histogram is defined by a two-components vector consisting of its modulus and its phasis. The modulus M(h) of a local histogram h is defined as the "distance" between this histogram and a "flat" histogram h_0. This distance is computed as the entropy difference between the two histograms, so that:

$$M(h) = \sum_i \left(n_i - n^2 / R \right) \mathrm{Log}\,(n_i)$$

where n_i is population of the ith class of the histogram,
R the number of grey levels,
$n = \sum_i n_i$.

The phasis $\theta(h)$ is defined as the index of the maximum component of h.
Thus for each sample, we construct 2 images: one for the modulus ans the other for the phasis. For both we compute the mean, variance and entropy of their grey level histogram.

STRUCTURAL ANALYSIS

In this approach, texture is viewed as composed of "primitives" (connected regions satisfying certain properties) placed in a certain spatial arrangement. To describe this texture one needs to describe both the primitives and the spatial dependance or interaction between these primitives. In fact, on each sample, we extract the texture primitives (alveoles). Then a set of attributes (shape parameters) is computed on each primitive. First order statistics of these attributes values give informations about the distribution of the primitives on the sample. Second order statistics provide a mean of quantizing the spatial relationships and dependances between these primitives.

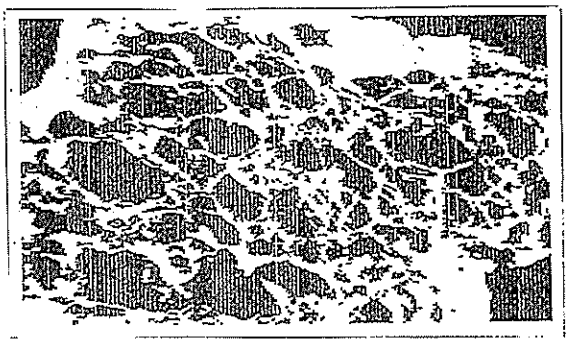


Figure 3

*) Primitives extraction
The basic idea is to segment the image into primitives (the visually connected cells) and background by local thresholding of the grey level. First the original 256x256 image is divided into 4 128x128 sub-images. On each of these sub-images we compute an optimal binarisation threshold using an iterative bimodalisation process {PEL-78}. A threshold is then computed for each pixel using a bilinear interpolation. Figure 3 shows the result of this binarisation. In order to reduce the influence of noise and artefacts on this image, we apply a morphological filter. This filter, described in {JOL-85} performs a segmentation of the image into 4 classes: nucleus, nucleus neighbourhood, jonction between neighbourhood and protuberances. Only the first and second classes are kept in the output images (figure 4).

*) Measures on the primitives
The extracted primitives are memorized and for each one, in addition to the position (x,y) of the gravity center, 5 shape attributes are computed:

+ A: Area (number of pixel in the primitive);
+ P: Perimeter (number of boundary pixels);
+ CPC: Compactness (A/P*P);
+ ECC: Eccentricity (ratio of the major to the minor axis of inertia);
+ OR: Orientation (angle that the major axis makes with a fixed orientation).

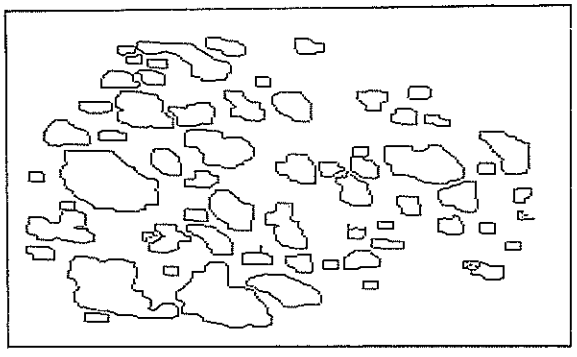


Figure 4

*) First order statistics
These are the first and second moments (mean and variance) of each attribute value histogram for one sample. Each texture is then characterized by 5x2 parameters.

*) Second order statistics
The definition of these statistics is based upon the concept of generalized co-occurence introduced in {HAR-79}:
Let Q be a set of primitives, T the set of primitives attributes, f a fonction assigning to each primitive of Q the value of a given attribute X of T. Let $S \subseteq QxQ$ be a binary relation pairing all primitives satisfying a certain spatial relation. A generalized co-occurence matrix PX is defined by:

$$PX[i,j] = Card\{ (q_1,q_2) \in S \mid f(q_1) = i, f(q_2) = j\}$$

These statistics are computed in 3 steps:
First, for each primitive qi, we list the set of its neighbours $\{q_j\}_i$. The neighbourhood relation is defined as follows:
On each sample, we say that two primitives q_i and q_j, with centroïds (x_i,y_i) and (x_j,y_j) and areas A_i, A_j are neighbours if $d(q_i,q_j) < d_T$, where d is an application from QxQ to R which associates to each pair (q_i,q_j) the value:

$$\max\{ [(x_i-x_j)^2+(y_i-y_j)^2]^{1/2}-(A_i/n)^{1/2}-(A_j/n)^{1/2},0\};$$

and d_T = mean "equivalent radius" of the primitive on the sample:

$$d_T = 1 / \text{card}\{q_i\}_i \, (\Sigma_i (A_i/n)^{1/2}.$$

Then, by using this neighbourhood relationship, we compute the co-occurence matrixes of the five shape attributes. For each of these attributes, the values are first divided into 8 disjoint intervals using an equal-probability quantizing algorithm. Each attribute has its own partition, but same partitions are used for all the samples. The size of the co-occurence matrixes is also 8x8. Finally, each matrix PX is normalized by dividing each entry by $\Sigma_i \Sigma_j PX[i,j]$.

At last, the second order statistics are computed using the following formulas:

+ Angular second moment:

$$ASM(X) = \Sigma_i \Sigma_j PX[i,j]^2.$$

+ Entropy:

$$ENT(X) = -(\Sigma_i \Sigma_j PX[i,j] * Log(PX[i,j])).$$

+ Inverse difference moment:

$$IDM(X) = \Sigma_i \Sigma_j (PX[i,j] / (1 + (i-j)^2)).$$

+ Contrast:

$$CON(X) = \sum_{k=1}^{N-1} (k^2 * \Sigma_i \Sigma_j PX[i,j]) \quad \text{with } k = |i-j|.$$

COMPARATIVE RESULTS

The data test set used in our experiment consists of 28 samples corresponding to 7 classes of texture (4 samples per class). Each sample is 256x256 pixels. In order to evaluate the ability of each method to discriminate between these classes of textures, we apply a classification algorithm using the feature vector computed on each sample. The classifier works on the basis of a Principal Components Analysis (PCA) followed by a Discriminant Analysis (DA) to assign each of the sample, described by its feature vector, to one of the 7 classes. By this way, and knowing the a-priori class of each sample, the classification accurency can be measured as the good classification rate.
For the first two methods (Grey level run length and Local histogram) the samples are median-filtered (to reduce noise) and equally requantified on 8 grey levels {SER-89}.
Table n°1 summarizes the results for the 3 methods.
For the Grey Level Run Length (GLRL) method, the feature vector consists of 3x5=15 components (5 parameters and 3 directions).
For the Local Histogram (LH) method, 3 neighbourhood sizes are selected (9x9, 17x17 and 33x33 pixels) and the dimension of the feature vector is 3x2=6 in each case (3 parameters computed on 2 resulting histograms).
For the structural analysis (SA), the dimension of the feature vector is 32 (11 first order parameters, including the total number of primitives, and 21 second order parameters, including the average neighbouring rate).

METHODS	Classification Accurency
GLRL	89%
LH Window-size=9	51%
LH Window_size=17	54%
LH Window_size=33	49%
SA Using only 1st order stat.	82%
SA Using only 2nd order stat.	93%
SA Using both 1st and 2nd order	75%

Table 1

It appears that the best results are obtained when using the second order features computed on the texture primitives. First order features on primitives perform a bit worse. The performances decrease when these two groups of features are used simultaneously in the classification stage.
Moreover, a PCA performed on this feature vector shows a very weak correlation between all the 2nd order variables.
The Grey Level Run Length performs almost as well as the second order statistics. In this case the PCA shows that the 15 components of the feature vector are strongly correlated. In fact a further Factorial Analysis performed on this set of variables showed 2 non-correlated groups of correlated variables.
The bad performances of the Local Histogram method, retrospectively, is not surprising since this method does'nt involve any second order-type information.
These results may be to compare with the ours of another french staff which works on similar products {MAL-88}. Authors have tested also three methods: grey run lengths, co-occurence matrix and structural analysis. For them, only the first method allows to discriminate between different classes of products. This conclusion is near our own, but we think that the bad results of the other methods can be explained and even corrected. For the co-occurence matrix, they choice a too little shifting to analyse a macroscopic texture. For structural analysis, they tested only a feature vector obtained with first order statistics, so it's not enough to well describe these kinds of textures.

CONCLUSION

The structural analysis based upon the second order statistics of primitives seems to be the best way to characterize that kind of "cellular" textures produced by the puffing extrusion process. However, its main drawback lies in its heavy computationnal cost. Further work should be done in order to optimize the primitives extraction process, which still requires a lot of time.
On the other hand, the Grey Level Run Length analysis gives satisfying results at a much smaller computationnal cost. This confirm previous results about the usefulness of this method in texture characterization. Moreover, we showed that it should be possible to strongly reduce the number of features required by this method. A real-time application could also be considered.

BIBLIOGRAPHY

GAG-83: A.GAGALOWITZ
"Vers un modèle de texture", Thèse de doctorat d'Etat es Sciences Mathématiques, Université de Paris 6, 1983.

GAL-75: M.GALLOWAY
"Texture analysis using gray level run lengths"
Computer Graphics and Image Processing, N° 4, 1975.

HAR-73: R.M.HARALICK, K.SHANMUGAM, I.DINSTEIN
"Textural features for image classification",
IEEE Trans. on Systems, Man and Cybernetic,
Vol SMC-3, N°6, 1973.

HAR-79: R.M.HARALICK
"Statistical and structural approaches to texture"
Proceedings of IEEE, Vol 67, N° 5, 1979.

JOL-85: J.M.JOLION, P.PREVOT
"Analyse d'images de structures cellulaires dans les matériaux fatigués", Congrès AFCET-INRIA,
RdF et IA, Grenoble, Nov. 1985, pp 1051-1062.

LOW-83: G.E.LOWITZ
"Can a local histogram really map textures information"
Pattern Recognition, Vol 16, N° 2, 1983.

MAL-88: C.MALOIGNE-FERNANDEZ, A.SMOLARZ, E.VANHECKE, J.M.BOUVIER
"Caractérisation de produits alimentaires extrudés par des méthodes d'analyse d'images texturées"
Applications de l'I.A. à l'Agriculture et l'Agrochimie, Premières Journées Internationales de Caen, Septembre 1988.

PEL-78: S.PELEG, A.ROSENFELD
"Determining compatibility coefficient for curve enhancement relaxation processes"
IEEE Trans. on Systems, Man and Cybernetic,
Vol SMC-8, N° 7, 1978, pp 548-555.

SER-89: J.SEROT
"Caracterisation de materiaux extrudes par analyse de texture", Rapport de DEA, LRII-ATI/IRESTE,
Université de NANTES, Septembre 89.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

IMAGE FEATURES EXTRACTION BY RADIAL TOMOGRAPHIC ANALYSIS

Giovanni Jacovitti (*), *Roberto Cusani* (**)

(*) INFOCOM Dpt. University of Rome 'La Sapienza',
Via Eudossiana 18, I-00184 Rome (Italy)
(**) Electronic Engineering Dpt., University of Rome 'Tor Vergata',
Via Orazio Raimondo, I-00173 Rome (Italy)

ABSTRACT

A new image decomposition technique based on radial projections of local polar maps is presented. This approach leads to a multichannel spectral analyzer constituted by a set of polar separable filters. These filters possess some nice properties, and are suited for direct extraction of fundamental features such as edges, lines, crosses, etc.
More complex applications including corner detection and texture recognition are allowed by the joint use of multiple filters.

INTRODUCTION

Research in image processing for computer vision and redundancy reduction has recently been focused on space-frequency selective decomposition.
In fact, the Fourier transform is representative of the dynamical overall behavior of the image, but it is unable to characterize local features. On the other hand, spatial domain operators, such as edge and line detectors, are not suited for efficient characterization of other features, such as corners, textures, surfaces, etc.
Window Fourier transforms, and specifically the Gabor transform, have been proposed as general tools for image analysis and representation [1]. This approach has been supported also by some analogies of the Gabor functions with the receptive fields in the visual mammalian cortex. A similar decomposition is performed by the wavelet transform [2].
In this contribution, we present a new method for joint spatial and spectral analysis. The approach originates from concepts developed in the area of automatic object recognition based on polar maps of images [3]. It has been ascertained that simple objects are sufficiently well characterized by radial tomographic views of their polar maps. This fact does suggest that they could be applied for discriminating image features, at a microscopic level.

One attractive point of this approach is that radial tomographic projections are inherently robust to magnification factors, while spatial rotations are converted into shifts.
However, direct application of polar mapping around any point of the image is unpractical because of its high computational cost, so that we have developed a technique for evaluating the radial tomographic view using shift-invariant operators in a multichannel scheme.
A set of these operators defines a family of functions having some important properties. In particular, some of these functions can be individually employed for the extraction of fundamental features, such as edges and lines.

THE RADIAL TOMOGRAPHIC PROJECTIONS

Let us refer to fig.1, where a simple pattern is displayed alongside its polar mapping with respect to a conventional center x_0, y_0 in the plane x, y. Referring to the polar coordinates:

$$r = \sqrt{(x-x_0)^2 + (y-y_0)^2} \qquad \theta = \tan^{-1}\left(\frac{y-y_0}{x-x_0}\right) \qquad (1)$$

the Radial Tomographic Projection (RTP) of the image $I(r,\theta)$ referred to the point x_0, y_0 is defined as:

$$\mathcal{R}(\theta; x_0, y_0) = \int_0^{r_M} w(r)\, I(r, \theta)\, dr \qquad 0 \le \theta \le 2\pi \qquad (2)$$

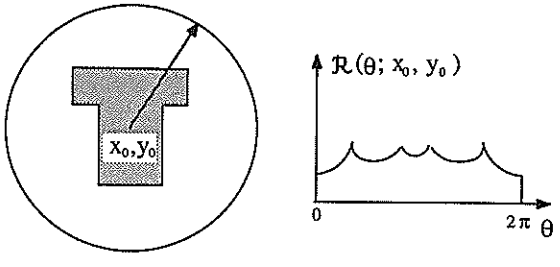


Fig.1 - Radial tomographic view.

where r_M is the radius of the projection circle, and w(r) is a radial weighting profile (RWP).
The mono-dimensional signal $\mathfrak{R}(\theta; x_0, y_0)$ is periodic modulo 2π. In particular, we observe that the RTP of a rotated image $I(r, \theta-\theta_0)$ is $\mathfrak{R}(\theta-\theta_0; x_0, y_0)$.
Moreover, the choice of suitable RWPs confers scale invariance to the RTP shape. For instance, posing w(r)=1 the RTP of a magnified image $I(\alpha r, \theta)$ is $\alpha^{-1}\mathfrak{R}(\theta; x_0, y_0)$ (if the windowing effects of the circular domain are negligible).
Owing to its periodic nature, the RTP signal can be represented by a Fourier series:

$$\mathfrak{R}(\theta; x_0, y_0) = \sum_{n=-\infty}^{\infty} R_n(x_0, y_0)\, e^{jn\theta}\, d\theta \qquad (3)$$

where the Fourier coefficients $R_n(x_0, y_0)$ are given by:

$$R_n(x_0, y_0) = \frac{1}{2\pi}\int_0^{2\pi} \mathfrak{R}(\theta; x_0, y_0)\, e^{-jn\theta} d\theta \qquad (4)$$

Each coefficient constitutes a new image in the plane x_0, y_0 and, for real images, the coefficients of order n and -n are conjugate in pairs. The set of coefficients forms a new local spectral description of the image, useful for characterizing local features, as described in the next section.
Substituting now def.2 into eq.4 yields:

$$R_n(x_0, y_0) = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{r_M} \frac{w(r)}{r}\, e^{-jn\theta}\, I(r, \theta)\, r\, dr\, d\theta \qquad (5)$$

Referring now to an input image $I(x,y)$, in absolute Cartesian coordinates, the above integral corresponds to a convolution product:

$$R_n(x_0, y_0) = \int\int I(x-x_0, y-y_0)\, f_n(x,y)\, dx\, dy \qquad (6)$$

where:

$$f_n(x,y) = \begin{cases} \dfrac{w(\sqrt{x^2+y^2})}{\sqrt{x^2+y^2}}\, e^{-j n \tan^{-1}(\frac{y}{x})} & 0 \le \sqrt{x^2+y^2} \le r_M \\ 0 & \text{otherwise} \end{cases} \qquad (7)$$

Thus, the evaluation of the RTP signals for each reference point in the input image can be carried out by a set of polar separable filters defined by eq.7, which we will refer to as Harmonic Angular (HA) "filters" or "functions".
The HA functions posses some remarkable properties. The fundamental property is the following

Property: The Fourier transform of a n-th order HA function is given by:

$$F(\rho, \phi) = (-j)^n\, e^{-jn\phi}\, H_n\left(\frac{w(r)}{r}\right) \qquad (8)$$

where ρ, ϕ are polar coordinates in the Fourier plane and $H_n(\,)$ indicates n-th order Hankel transform.
This property says that the spectrum of the HA functions is itself polar separable, with the same angular modulation, and that the bandwidth generally increases with the order n. For the sake of brevity, proof, details and some corollaries derived from this fundamental property are omitted here and are reported in [4].
We outline here that the features of the HA filters depend on the radial profile w(r). Many choices can be made, depending on the application at hand. For instance, profiles of the kind: $w(r) = r^s$ define through eq.2 the Mellin transform [5] and lead to pattern size independence. In object recognition by optical template matching, some particular HA filters with object dependent profiles have been also employed [6,7]. In this work, we made use of wideband HA filters, controlling their RWP in the frequency domain. In fig.2 the image of these HA filters along with their Fourier transforms is displayed. The radial isomorphism in the two domains is a consequence of the particular radial profile, given by $w(r) = r^n e^{-r^2}$ (for the n-th order).

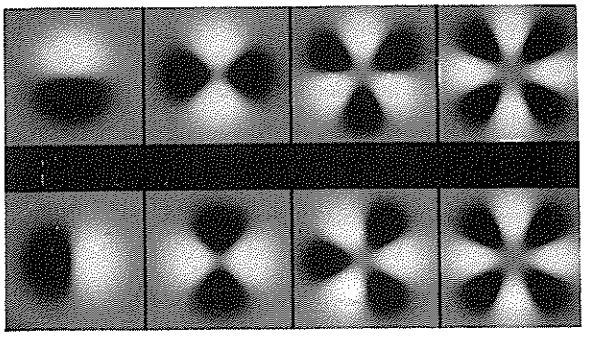


Fig.2 - Real part of the HA filters with $w(r) = r^n e^{-r^2}$, n=1,2,3,4 (top row, from left to right) and real part of their Fourier transform (bottom row, from left to right).

APPLICATIONS OF THE HA FILTERS

In comparison with the Gabor functions, the HA functions perform local harmonic decomposition as well, but in a angular rather than in a usual orthogonal sense. Moreover,

their bandwidth is much larger than the one of the Gabor functions.
The outstanding aspect of the HA functions is that they are strictly related to basic features, and possess individual significance.
To show that, let us first consider the RTP of an edge. It consists of a pulse in the interval $[0,2\pi]$. The size and the sharpness of this pulse change as long as the projection circle shifts orthogonally to the edge. When the edge crosses the center, then the duty cicle of the pulse is 50%. Thus, the first order RTP harmonic, namely the output of the first order HA filter, presents a pronounced peak in this position. Moreover, rotations of the edge pattern around the center corresponds to shifts of the pulse in the RTP domain, and in particular to a simple phase shift of the first order output. Therefore, the first order HA function is well suited for contour extraction and orientation estimation.
In fig.3 the magnitude of the first order output is compared with the output of a conventional edge extractor.

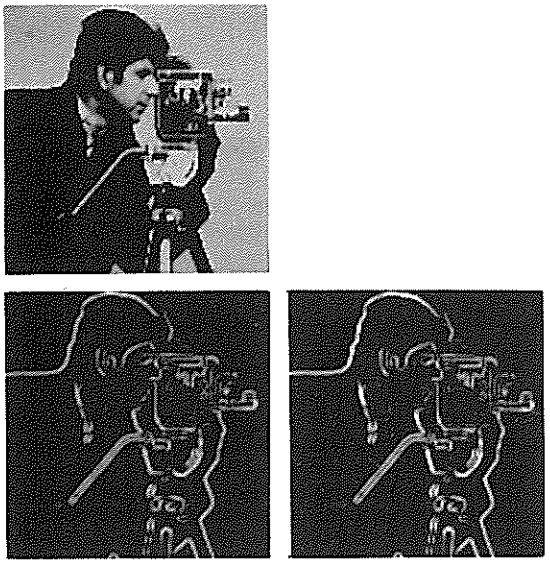


Fig.3 - Test image(top); magnitude of the first order RTP harmonic (bottom left) and Sobel edge extraction (bottom right)

The phase output in gray levels is displayed in fig.4 along with a simple application of directional edge extraction. Here, the magnitude output is suppressed for phase values outside a narrow angular sector (ten degrees). This method appears simpler and more accurate than conventional techniques based on edge gradient or directional filtering.
With the same reasoning, it is easily verified that lines (close parallel and complementary edges) produce a pair of pulses in the RTP. As a consequence the amplitude of the

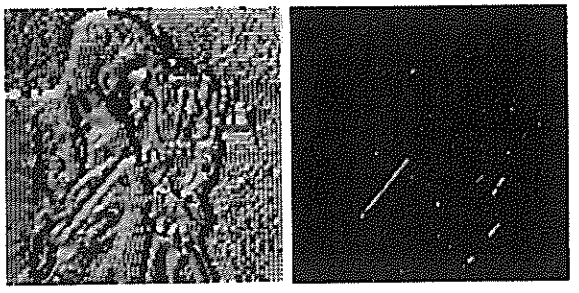


Fig.4 - Phase of the first order RTP harmonic (left) and directional edge extraction (right) of the image of fig.3.

second order RTP harmonic (output of the second order HA filter) presents a peak when the line crosses the center of the projection circle. The resulting phase is twice the orientation angle of the line.
In fig.5 an example is shown.

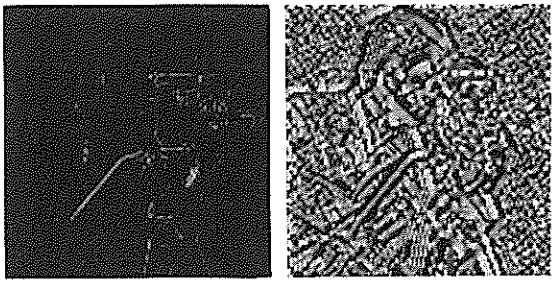


Fig.5 - Magnitude and phase of the second order RTP harmonic of the image of fig.3

Increasing the RTP harmonics order, it is easily seen that third order HA filters are suited for detecting trihedral vertices, while fourth order HA filters can be used for detection of crosses. Such capabilities can be exploited in some textures recognition problems.
For instance, textures originating from honeycomb structures are characterized by the presence of thriedral vertices. In fact, the reptile skin of fig.6 is immediately recognized by a third order HA filter. It is revealed by a sparse grid of spikes into the magnitude output. In a similar way, textures made by textiles are revealed by spikes of the magnitude output of a fourth order HA filter, as shown in the same figure.
It is not difficult to devise more sophisticated recognition schemes based on the joint use of multiple RTP harmonics. Let us consider for instance the classical problem of corner detection. The solution to this problem is straightforward in the framework of radial tomographic analysis. In fact, a

corner is characterized by high values of the moments (or moments of the derivative) of the RTP when its vertex is placed on the center of the projection circle. For instance, the presence of a corner can be revealed by the fourth order moment (L4 norm) of the RTP:

$$\int_0^{2\pi} \left[\mathcal{R}(\theta;x_0,y_0)\right]^4 d\theta \qquad (9)$$

or by the fourth order moment of the RTP derivative:

$$\int_0^{2\pi} \left[\frac{d}{d\theta}\mathcal{R}(\theta;x_0,y_0)\right]^4 d\theta \qquad (10)$$

In fig.7 an example of corner detection based on the first four RTP Harmonics is reported.

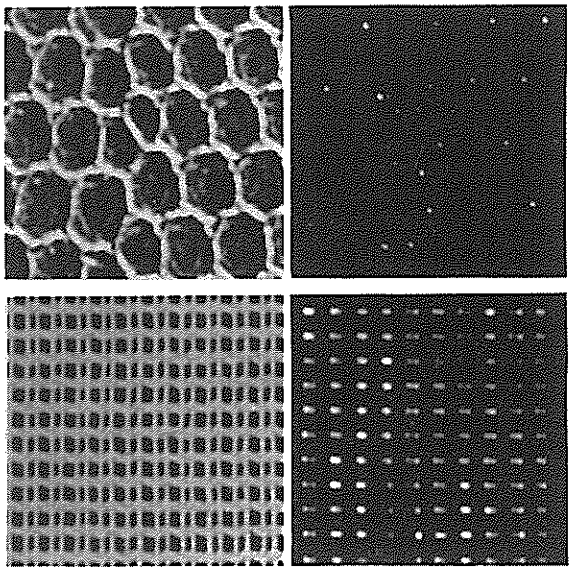


Fig.6 - Snake skin texture and third order HA magnitude output (top); textile texture and fourth order HA magnitude output (down)

CONCLUSIVE REMARKS

Radial tomographic analysis provides new tools for image feature extraction. Most features, ranging from edges, lines, corners to textural fields are handled into a single flexible framework.

Even thought the modeling of the mammalian visual mechanism is beyond the scope of our work, a resemblance of some low order HA functions to receptive fields in the visual cortex is evident, and could suggest new interpretations.

Further work about this subject will cover multiresolution RTP schemes and image coding applications.

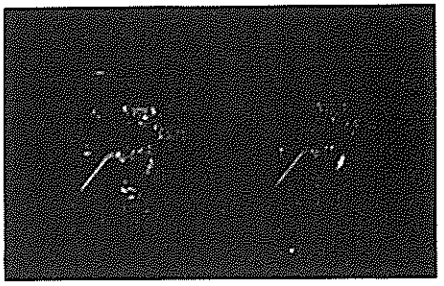


Fig.7 - Fourth order moment of the RTP (left) and of its derivative (right) of the image of fig.3

REFERENCES

[1] J.G. Daugman, "Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression", *IEEE Trans. on Acoustic, Speech and Signal Proc.*, Vol.ASSP-36, no.7, July 1988.

[2] S.G. Mallat, "Multifrequency Channel Decompositions of Images and Wavelet Models", *IEEE Trans. on Acoustic, Speech and Signal Proc.*, Vol.37, no.12, Dec. 1989.

[3] R. Cusani, G. Jacovitti, "A Double Tomographic Approach to the Estimation and Classification of Single Objects", *Proc. ICASSP-89*, Glasgow, Scotland, May 1989.

[4] G. Jacovitti, R. Cusani, "Image analysis by local angular spectra", in preparation.

[5] C. Braccini, G. Gambardella, "Form-Invariant Linear Filtering: Theory and Applications", *IEEE Trans. on Acoustic, Speech and Signal Proc.*, Vol.34, no.6, Dec. 1986.

[6] R. Wu, H. Stark, "Rotation-Invariant Pattern Recognition Using Optimum Feature Extraction", *Applied Optics*, Vol.24, no.2, 15 January 1985.

[7] J. Rosen, J. Shamir, "Circular Harmonic Phase Filters for Efficient Rotation-Invariant Pattern Recognition", *Applied Optics*, Vol.27, no.14, 15 July 1988.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

HIERARCHICAL DOCUMENT SEGMENTATION SYSTEM

Dr. G. S. D. Farrow and Professor C. S. Xydeas,

Multimedia Information Systems, Department of Electrical Engineering,
University of Manchester. Manchester. M13 9PL.

The increased use of electronic document production and handling systems brings with it the problem of integrating existing paper format documents. Given a scanned binary image of a document page, the problem is one of segmenting the page into its fundamental components typically typewritten text, geometric diagrams and photographics. A novel segmentation strategy is proposed incorporating a hierarchy of partitioning and classification schemes. This enables classification of unambiguous data by relatively simple algorithms whereas ambiguous data is classified using increasingly complex techniques.

1. INTRODUCTION

The electronic handling of documents is becoming increasingly common in the office environment. To facilitate the handling of such electronic documents a new information interchange standard has been formulated, namely the Office Document Architecture (ODA) standard [ODA85]. A problem therefore emerges concerning the integration of existing and future paper documents into such an electronic environment. For this task the concept of Automated Document Entry (ADE) [HOR85] has been introduced, the stages of which are illustrated in Figure 1. A crucial stage in the ADE process, and one which represents a challenging research problem, is the segmentation of a binary document page into three fundamental constituents, namely text, geometric diagrams and photographics. Having done this, recognition and coding of the document components may be undertaken. This involves the application of character recognition algorithms to areas identified as text, the vectorisation of areas identified as geometrics and the compression of regions identified as photographics. The final task is the specification of the document in terms of the ODA standard.

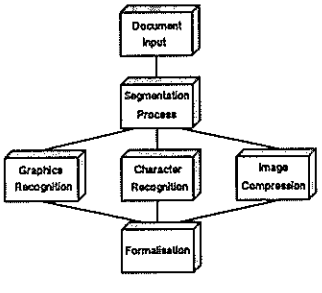


Figure 1 Stages Involved in Automated Document Entry (ADE)

This paper is concerned primarily with the document segmentation problem and a general purpose system is described which is applicable to free-format document pages.

2. SEGMENTATION SYSTEM FRAMEWORK

For the solution of the document segmentation problem a working framework is defined and is illustrated in Figure 2. The segmentation process is divided into 3 distinct phases. An initial partitioning of the page is produced with the objective of producing distinct regions of homogeneous data. The next phase is the classification of the regions produced by the partitioning phase into one of the 3 classes. Inbuilt into the framework is a verification procedure to confirm the results produced by the classification stage. An important feature of the framework is the provision of a feedback mechanism in which knowledge about the components of the page accumulated in the initial stages, can be used as input to the later stages. The system is also hierarchical in that initially relatively simple partitioning and classification schemes are to be employed. The aim is to extract unambiguous data from the page with relative computational ease. For more complex document architectures which may contain ambiguous data, progressively more complex algorithms are to be applied until a satisfactory segmentation of the page is achieved.

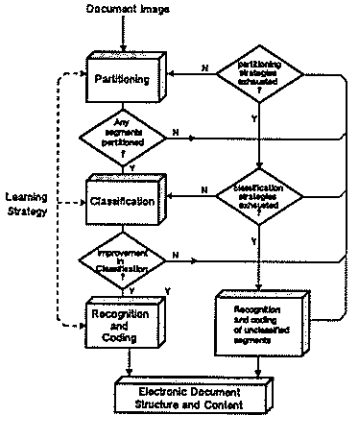


Figure 2 Segmentation Framework

A document segmentation system which conforms to the framework outlined is presented. It is shown to be successful for a variety of document architectures containing a mixture of text, geometrics and photographics. However, development of the system is ongoing and it is envisaged that further algorithms will be incorporated into the system within the hierarchical framework discussed. In particular, schemes for the verification phase are currently under investigation.

3. SYSTEM IMPLEMENTATION

3.1 Hybrid Partitioning Scheme

The aim of the partitioning phase of the segmentation system is to produce homogenous regions of (as yet unclassified) data. It is a very important step as misclassification of the data occurs in general, not from the actual ambiguity of the data itself, but from incorrect partitioning resulting in regions of mixed data types. A number of partitioning strategies have been attempted in the past with varying degrees of success.

The run-length smoothing algorithm (RLSA) proposed by Wahl et. al. [WAH80] is a low complexity algorithm which exploits the rectangular block structure of typical document pages. The algorithm works by producing two masks, one from eliminating horizontal white runs less than a threshold rh and one from eliminating vertical white runs less than a threshold rv. These two masks are then combined using a logical AND operation to produce a final mask. Although the values rh and rv are not critical, the algorithm will fail for the case of the document being skewed and for non-block structured page architectures.

Another method of decomposing the page into homogenous blocks is the x-y tree proposed by Nagy and Seth [NAG84]. In this approach, the entire page is decomposed into a hierarchical structure using recursive cuts in the horizontal and vertical black pixel projection profiles of a given image. The root node of the structure represents the entire document page. Descendants of a node are obtained by successive horizontal or vertical partitioning. Tsuji et. al. [TSUJI87] have attempted to prevent full decomposition of the page to the character level by analysing the correlation between adjacent peaks of the horizontal projection profile. If the correlation is high then this implied that the image is a segment of typewritten text and further decomposition prevented. The problems with this approach are again failure with skewed images and inability to cope with free format document pages. Furthermore, a particular problem associated with both the RLSA and the x-y tree methods is their failure to separate different data that is contained within a rectangular box such as that illustrated in Figure 5.

A more reliable partitioning technique which is independent of skew angle and page structure is that of connected component extraction using boundary tracing. The disadvantage of the technique is that it is much slower in isolating components of the page than recursive projection cutting.

We have adopted a hybrid partitioning scheme which exploits the speed advantage of recursive projection cutting and the reliability of connected component extraction. The scheme can be described as a bottom up procedure in which all connected components are first isolated and separated into symbols and non symbols based on a size criteria. The symbols represent possible text characters and these are then grouped to form larger regions.

The algorithm proceeds as follows. First decomposition of the page is performed using recursive projection cuts. This produces a set of rectangular subimages which cannot be further subdivided by projection cuts. The size of subimage is now compared to a threshold. If the dimensions are less than the prescribed threshold then further decomposition is impossible or unnecessary and the subimage is labelled as containing a symbol. If the dimensions are bigger than the prescribed threshold then a number of situations must be accounted for. Firstly the subimage may contain a large connected component in which further partitioning is impossible. More importantly, the subimage may contain text characters which cannot be partitioned due to skew or local distortion or it may contain a mixture of data types enclosed in a box structure. To account for the latter situations, further analysis must be performed on these subimages.

This is done using a component extraction algorithm which uses windowed boundary tracing [JOH83]. Using a window size of the same dimensions as the above threshold, any connected components less than the window size are guaranteed to be separated out from larger components. We choose a threshold window size which is able to contain the largest character of a 24pt font. Ideally we expect to separate all typewritten text less than this size. In practice characters will be joined and isolated text may appear in the non symbol area. This is of no consequence since these will be identified at the classification stage. Thus, by the combined use of the above two algorithms, a page can be partitioned into symbols and non symbols as prescribed by the threshold.

3.2 Symbol Merging

For all symbols located in the original doucment image a solid rectangle of dimensions large enough to circumscribe the symbol is drawn into a buffer image of the same dimensions as the original. Components larger than the prescribed window size are copied to another buffer image. The symbols and non symbols are thus separated. The aim now is to group the symbols into larger regions such that a classification of these regions can be made. This can be achieved via a simple application of the RLSA.

A problem may arise with photographic areas in that, although they consist mostly of large solid black areas, there are also smaller regions which will be separated as symbols, such that there is an overlap between symbol and non symbol areas. Similarly geometric areas, typically consisting of unconnected line segments and text characters may be split resulting in overlapping regions. To eliminate the overlap, run length smoothing is also applied to the non symbol image buffer which creates a solid mask area. XORing this mask with the symbol then eliminates any overlap between the symbol and non symbol areas.

3.3 Region Classification

The next phase in the segmentation process is the classification stage in which we label blocks created during

the merging phase as one of the 3 data types. The approach adopted here is a classical statistical pattern recognition approach. With this technique a vector of statistical features, $\bar{x}$ is extracted from a block to be classified. For each of the 3 classes we calculate the function

$$F_i(\bar{x}) = \sum_{j}^{N} c_{ij} \times j + c_{i0}$$

where N is the number of features extracted.
F_i is the function for the ith class.
$\bar{c}_i$ is a set of weights and c_{i0} is a constant.

A region belongs to the class for which $F_i(\bar{x})$ is minimised.

c_i and c_{i0}, the class weights and constant respectively are derived from the class covariance matrix and class means vector. These are estimated via a training process in which feature vectors are extracted from a set of M images representing samples from each of the 3 classes. For the classification problem in hand it is felt that the information content of the image is localised in the black pixels. For this reason we choose statistics which are based on black pixel run lengths only. A total of seven features were extracted and used in our experiments. These are

1. Mean horizontal black run length
2. Standard deviation horizontal black pixel run length
3. Mean vertical black pixel run length
4. Standard deviation vertical black pixel run length
5. Black-black pixel transition probability (horizontal direction)
6. Black pixel probability
7. Maximum Black pixel run length

4. TEXT/HANDWRITING DISCRIMINATION

Using the partitioning strategy outlined above we expect typewritten text to be extracted as symbols. There are two circumstances in which regions found in the non symbol mask will be classified as text. This situation may arise if several text characters are joined together such that the combined size is greater than the symbol/non symbol threshold. A more important problem which must be accounted for is the fact that the statistics of typewritten and handwritten text are very similar and hence no discrimination can be achieved using the statistical classifier described. Within the hierarchical framework outlined in Section 2, we now apply a more powerful but computationally more complex algorithm which allows us to achieve the required discrimination.

The discrimination is achieved via the application of an adaptive Wiener filter to these non symbol regions classified as text. The operation of this filter is now outlined. The filtering process consists of tracking a fixed size window across the given segments.

Suppose we have an input mask window, containing L elements. In general the output y(i) of the filter is given by:

$$y(i) = W_o + \sum_{j} m(i,j)W_1(j) + \sum_{j}\sum_{k} W_2(j,k)m(i,j)m(i,k)$$

+ Higher order terms .....

where W_o is a constant and W_1, W_2 are a series of 1st and 2nd order weights and m(i,j) is the jth element of the mask.

To eliminate the effect of white space (having no information content) on the filtering process, we choose to apply the filter only when it is centred on black pixels. A training process is undertaken to optimise the weights in which the desired output of the filter is fixed to '1' for a segment of typewritten text and '0' for a segment of handwritten text. The weights are chosen such that the mean squared residual between desired and actual filter outputs is minimised.

To assess the performance of filters of different orders and having different numbers of coefficients, the filters are applied to the original training samples and the mean squared residual is evaluated. Results are presented in Table 1 for a series of linear and 2nd order filters. From the results presented it is apparent that a linear 5x5 filter offers a good performance/complexity tradeoff and is therefore incorporated into the system hierarchy.

It is envisaged that the filters may also be trained to discriminate between other pairs of the 3 classes. There exists also the possibility of implementing similar filters via a multilayer network approach.

Filter Size	1st Order Terms	2nd Order Terms	MSR
3x3	8	0	0.1911
5x5	24	0	0.144
7x7	48	0	0.1391
9x9	80	0	0.1369
15x5	74	0	0.139
3x3	8	4	0.1841
3x3	8	8	0.1795
3x3	8	28	0.1722

Table 1: Comparison of Filters for Text/Handwriting Discrimination

5. EXPERIMENTAL RESULTS

The performance of the above schemes are now illustrated for some sample document images. Figure 4 illustrates the application of the scheme to the non block structured page segment in Figure 3 for which straightforward application of run length smoothing fails. The first stage of the segmentation procedure consists of separating symbols and non symbols into 2 separate buffer images. This is shown in Figure 4(b). Run length smoothing is now applied to both these images and the result after XORing to prevent overlap is shown in Figure 4(c). Blocks from both the resultant masks are now classified using the statistical classifier and Figure 4(d) shows classified areas separated out from the original image. The sequence shown in Figure 5(a)-(d) illustrates performance for a page segment containing a rectangular box structure surrounding 2 different data types. Note the

removal of symbols from the photographics area by the XORing process shown in Figure 5(c). It is seen that text is successfully extracted from the boxed area. This principle may be used to extract text from areas classified as geometrics. Regarding Figure 6, it is seen that the initial classification identifies regions of typewritten text and geometrics. Given that geometrics does itself contain some typewritten text, using the algorithms in the segmentation system, we are able to extract text from the geometric segment.

6. DISCUSSION

A system for the segmentation of free-format document pages has been presented. A reliable bottom-up approach is adopted in which all connected components of the page are first isolated regardless of any skew or local distortion that may be present. By using a hybrid partitioning scheme, a page is decomposed in the most efficient way possible.

The hierarchical framework outlined allows for the partitioning and classification of unambiguous data using relatively simple techniques whereas more difficult problems, such as text and handwriting discrimination, are solved using computationally more complex and powerful techniques. Further algorithms are to be incorporated into the segmentation system within this framework.

7. ACKNOWLEDGEMENTS

The work has been funded by the Office Systems Division of International Computers Limited, Bracknell, England.

8. REFERENCES

[ODA85] Standard ECMA-101, "Office Document Architecture", 1985.

[HOR85] W. Horak, F. Tartanson and G. Coulouris, "Handling of Mixed Text/Image/Voice Documents Based on Standardised Office Document Architecture", ESPRIT '84, Status Report of ongoing work, pp. 395-410, 1985.

[WAH80] F. M. Wahl, K. Y. Wong and R. G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents", Computer Graphics and Image Processing, Vol. 20, pp. 375-390, 1980.

[TSU87] Y. Tsuji, J. Tsukumo and K. Asai, "Document Image Analysis for Reading Books", Advances on Image Processing, SPIE, Vol. 804, pp. 237-243, 1987.

[NAG84] G. Nagy and S. Seth, "Hierarchical Representations of Optically Scanned Documents", Proc. 7th International Joint Conference on Pattern Recognition, pp. 347-349, 1982.

[JOH83] O. Johnsen, J. Segen and G. L. Cash, "Coding of Two Level Pictures by Pattern Matching and Substitution", Bell Systems Technical Journal, Vol. 62, No. 8, pp. 2513-2545, 1983.

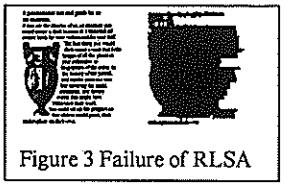


Figure 3 Failure of RLSA

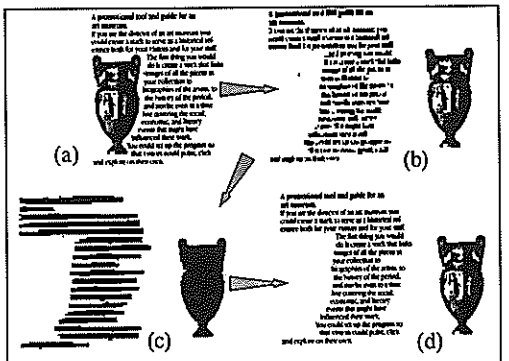


Figure 4 Segmentation of non block structured page

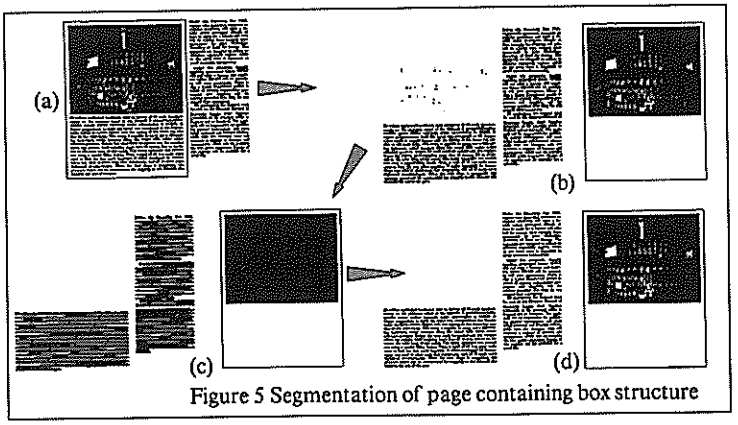


Figure 5 Segmentation of page containing box structure

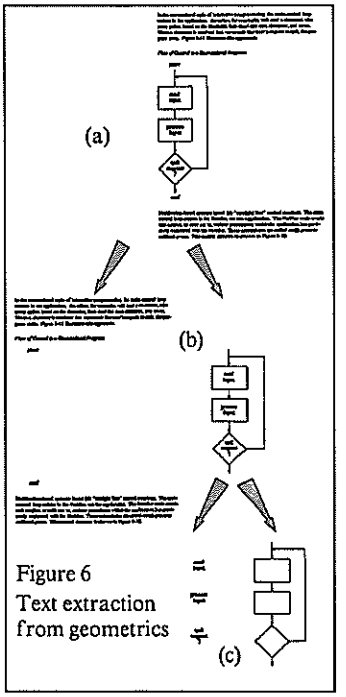


Figure 6
Text extraction
from geometrics

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ARABIC TYPESET : AN OCR APPROACH

H.Y. Abdelazim, & M.A. Hashish

IBM Cairo Scientific Center
56, Gamiaat Al Dowal Al Arabia St.,
El Mohandissen, Giza,
EGYPT

In the present work, the Arabic typeset Optical Character Recognition (OCR) is considered. The problem here differs significantly than other OCR problems for European languages, due to the technical difficulties associated with the nature of Arabic writing. A recognition system has been developed, and the experiments showed a 96% success. The key module in the recognition system is the segmentation module.

1. INTRODUCTION

Optical Character Recognition (OCR) has been a subject of intensive research for quite a long time, to provide practical means for processing large volumes of data automatically. It is believed that practical text readers will make a break-through in office automation [1].

Arabic typewritten OCR faces a main technical problem not encountered in pure latin text, which is the "cursiveness", characterizing the Arabic writing style (characters are not isolated).

Arabic typeset, however, faces three more technical problems, namely : character overhanging, character overriding, and multiple shapes per character.

Intensive R&D work has been done on Latin OCR [1-3,12,14], while the technology of Arabic OCR is still in the phase of research and experimentation. Few research has been done on Arabic typewritten OCR [5-10], while no efforts have been reported on the OCR of Arabic typeset, due to relative complexity. This is a fact even though the majority of Arabic books and government records are in typeset form, and the automation of many applications requires a fast and efficient way for invoking these data into the machine.

In the present work, a recognition system has been developed consisting of four stages, namely : preprocessing, segmentation, feature extraction, and finally the learning/recognition stage.

In the preprocessing stage, simple image processing tasks are performed for line detection and Arabic word isolation. The segmentation stage, which is in fact the key for Arabic OCR in general, performs three main functions : resolving overhanging characters, vertical segmentation, followed by horizontal segmentation.

In the feature extraction phase, two functions are performed, first a two dimensional feature vector (v_size), is generated, simply by computing the length and the width of the primitive at the input to this stage. The second is repositioning (centering) the primitive in a matrix of fixed dimension.

In the learning/recognition phase, the system first learns the input text, and generates a coded information to be used later in the recognition. The learning phase is subdivided into two successive phases, an unsupervised followed by a supervised learning phase.

In the unsupervised phase, all the primitives are gathered and representative prototypes are generated for each primitive class. An unsupervised clustering[11] algorithm is used to group the primitives depending on v_size into cluster centers.

In the supervised phase, the membership of each primitive class is registered, using nearest neighbor approach and a simple absolute distance measure. Structural rules are also learned in this phase, to assign different attributes to the primitives.

The recognition phase is composed of three successive modules : preclassifier, template matching, and reconstruction.

In the following sections the characteristic of Arabic typeset will be presented as well as some details of the recognition system.

2. CHARACTERISTICS OF ARABIC TYPESET

The nature and characteristics of Arabic writing style in general differs significantly than the writing style for most European languages [6]. The Arabic writing can be divided (from an OCR perspective) into Arabic typewritten and Arabic typeset. For both types, the Arabic writing is basically cursive in nature. Concatenation of isolated Arabic characters, is an unacceptable way of Arabic writing, unlike Latin writing.

The Arabic typewritten script is constructed by linear "welding" of characters in an adjacent manner. The Arabic typeset, however allows the overriding (solid overlapping) of a particular combinations of characters, as well as character overhanging (non-solid overlapping). These properties have a direct impact on the complexity of the OCR system due to the burden imposed on the segmentation.

Figure 1. shows the form of a Latin word, versus an Arabic typewritten and an Arabic typeset words. The figure also demonstrates the overriding and overhanging properties.

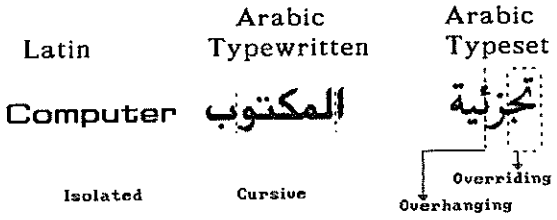


Figure 1

3. RECOGNITION SYSTEM

The overall block diagram of the recognition system is shown in figure 3. Each component will be addressed in the following,

3.1 Scanning and Digitization

The input device used is a feed-through binary digital scanner, which generates an image file after scanning the document. The resolution employed is 240 dpi, which is suitable for most OCR applications[13].

3.2 Line Detection

In this stage the binary image obtained from the previous phase is being processed to isolate lines of text. Defining a scan line as a horizontal line of pixels within the image, the line detection is accomplished by simply segmenting a histogram made up of the profile of the scan lines[8]. In some situations, this approach may fail, in particular for the case of overhanging lines, and a more sophisticated line detection algorithm is needed[8,13].

3.3 Word Isolation

This is accomplished by vertically isolating "black" patterns of the image (within the boundaries of the detected line). For Arabic text, these isolated black patterns correspond to Arabic words (or part of a word), and corresponds to characters in Latin text.

3.4 Segmentation

A segmentation algorithm, called pitch segmentation was used in early OCR systems for the recognition of Latin text[12]. Using this approach for segmentation of proportionally spaced Latin fonts as well as Arabic text produced significant errors.

For Arabic typewritten, a segmentation algorithm based on traversing an energy-like curve[6-9], with a predefined threshold, has been successfully applied. This approach cannot be directly applied to the case of Arabic typeset, due to the overhanging property as demonstrated in figure 1. The segmentation algorithm developed in the present work, consists of three successive steps,

* Resolving overhanging characters, and this is done by tracing the contour of the word[14].
* Vertical segmentation : Using the approach used for Arabic typewritten text [8].
* Horizontal segmentation : Isolating dots[6].

The result of the segmentation is a set of primtives as depicted in figure 2.

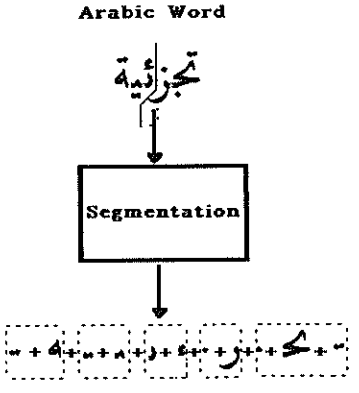


Figure 2 Segmentation Algorithm

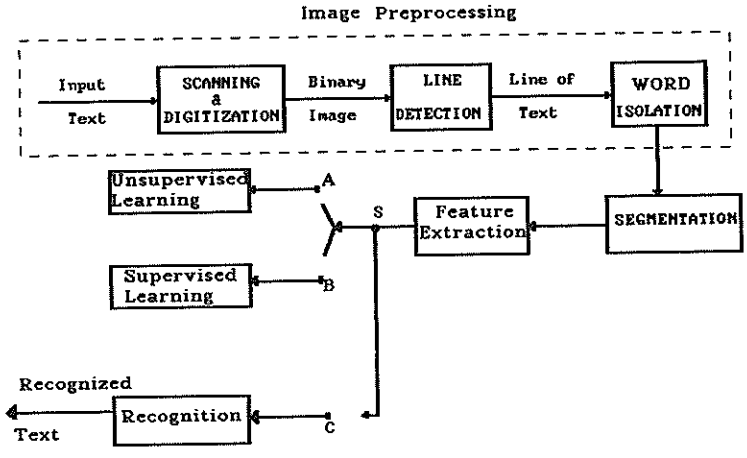


FIGURE 3 RECOGNITION SYSTEM

3.5 Feature Extraction

For developing a fast and efficient classifier, a simple two dimensional feature vector "v_size" comprising the width and length of each primitive is computed. Also in this phase, each primitive is centered in a matrix of fixed dimension to preserve the positional information of the pixel distribution

3.6 Learning/Recognition

As shown in figure 3., the system can be instructed to switch "S" between the learning and recognition modes.

3.6.1 Unsupervised Learning

In this phase a library of prototype templates, is generated for all the primitives, by learning a given text.The learning algorithm is described as follows,

1. IF K=1 INSERT THE PRIMITIVE IN THE LIBRARY
2. COMPUTE θ(k) [8,9].
3. GET θm=MAX{θ(k), k=1,..K}.
4. IF θm<T THEN NEXT PRIMITIVE, GOTO 2
   ELSE APPEND PRIMITIVE TO LIBRARY

K : is the current number of primitives in the library. θ is the correlation index between the input primitive and the k th primitive class.θm is the maximum correlation value between the input primitive and all the K primitive classes.T is a preselected threshold controlling the total number of primitives in the library(K is inversely proportional to T).

The second function performed, in this phase is deriving certain cluster centers using a K-means clustering algorithm[14]. The features used are those obtained in the feature extraction phase, namely depending on the length and width of the primitive. The learning in this phase is said to be unsupervised, due to the fact that the primitives are gathered and cluster centers are determined without external labeling or association between the generated primitives and the original characters

3.6.2 Supervised learning

In this stage, exactly two functions are performed, the first function is grouping each of the K primitive classes into the nearest neighbor cluster center using an absolute distance measure.

The second function performed is done interactively, where the user is prompted to dictate certain structural rules to be used in the recognition phase. The method adopted is assigning an attribute code to each of the primitives. The code depends on the nature of the primitives as follows,

- The primitive is a character.
- The primitive is part of a character.
- The primitive is a group of characters.
- The primitive is an upper dot.
- The primitive is a lower dot.

These are the possible features that could be associated with each of the primitives.

3.6.3 Recognition

The recognition of the input primitive is accomplished through three main steps,

- Preclassifier : In which the input primitive is grouped to one of the cluster centers which considerably reduces the search. This grouping (assignment) is based on an absolute distance measure between the v_size of the input primitive computed in the feature extraction phase and each of the cluster centers.

- Main Classifier : Within a particular group, the correlation index $\theta[8,9]$ is computed between the unknown input primitive and each of the primitive classes within the group (as stored in the library of primitives). The recognition rule is as follows,

$k(\text{recognized}) = \arg \{\text{Max}[\theta(k)]\}$, $k=1.....C$.

C: number of primitive classes in the group.

- Reconstruction : In this phase the system utilizes the information obtained in the supervised learning phase, to reconstruct the characters from the recognized primitives[9].

4. RESULTS

Recognition experiments have been done on a sample of a typeset font, using an IBM PS/2 (286 processor) and IBM 3118 general purpose scanner. The resolution used is 240 dpi. The total number of primitives evolved from the segmentation and the unsupervised learning phase are 210. The number of cluster centers is 20 with an average of 13 primitive class per cluster. The cluster are overlapping to permit accurate recognition. Preliminary results showed a recognition rate of 96%. The recognition speed is about 30 wpm(words per minute).

5. CONCLUSION

In the present work, a recognition system for OCR of Arabic typeset is presented. The technical difficulties associated with the nature of Arabic typeset are explained, particularily the segmentation problem. The segmentation algorithm discussed hereinbefore is the key for Arabic OCR. The number of primitives to be searched is large as compared to Latin OCR, and accordingly a multistage classifier is used to reduce the search. The recognition rate reached is 96% which is promising, while the recognition speed (30 wpm) needs further improvements, which can be achieved by using faster hardware classifiers.

REFERENCES

[1] P.L. Anderson, "OCR enters the Practical Stage," Datamation, vol. 17, pp. 22-27, Dec. 1971.

[2] V.A. Kovalsky, "Character Readers and Pattern Recognition," New York : Spartan Books, 1968.

[3] M.E. Stevens, Guest Ed., "Special Issues on Optical Character Recognition, Pattern Recognition, vol. 2, pp. 145-239., Sep., 1970.

[4] C.Y. Suen, "Advances in Optical Character Recognition," in Proc. Canadian Computer Conf., pp. 262-268, May 1978.

[5] H.Y. Abdelazim, and M.A. Hashish, "Automatic Recognition of Arabic Text," 10 th Image/ITL conference in IBM Toronto Lab., August 1987.

[6] H.Y. Abdelazim, and M.A. Hashish, "Arabic Reading Machine " 10 th NCC Computer Conference, King Abdulaziz University, Saudi Arabia, March 1988.

[7] H.Y. Abdelazim, and M.A. Hashish, "Arabic Reading Machine " 10 th NCC Computer Conference, King Abdulaziz University, Saudi Arabia, pp.733-744 March 1988.

[8] H.Y. Abdelazim, "Text Recognition : Theory and Implementation," PhD Thesis, March 1989,.

[9] H.Y. Abdelazim, and M.A. Hashish, "Automatic Reading of Bilingual Typewritten Text, " Proc. of COMPEURO 89' VLSI & Computer Peripherals IEEE Conference in Hamburg, vol.2 pp. 140-144 May 8-12, 1989.

[10] M. Khemakhem, and M.C. Fehri, "Arabic Typewritten Character Recognition Using Dynamic Comparison," Proc., 1st. Kuwait Computer Conference, pp. 455-462, March, 1989.

[11] J. Hartigan, "Clustering Algorithms," John Wiley, 1962.

[12] C.R. Jih, "Segmentation Method for Fixed Pitch Machine Printed Documents," IBM Technical Disclosure Bulletine, 23, 1194, August 1980.

[13] R.G. Casey and C.R. Jih, " A Processor-Based OCR System," IBM Journal of R&D, vol. 27, Number 4. July, 1983.

[14] R.O. Duda and P.E. Hart, "Pattern Classication and Scene Analysis," John Wiley & Sons, 1973

SIGNAL PROCESSING V: Theories and Applications.
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Noise Removal in Forward-Looking Infrared Images

María José Pérez-Luque, Carlos Muñoz and Narciso García

Grupo de Tratamiento de Imágenes, E.T.S. Ingenieros Telecomunicación
Universidad Politécnica de Madrid, E-28040 Madrid, Spain

The technology of FLIR image acquisition systems is currently unable to provide the desirable signal to noise ratio. So, it is necesary to process these images to improve their quality. After analyzing the typical FLIR noise, a 2-D non linear filter was chosen for quality upgrading. Several filters have been implemented and evaluated (under statistical and subjective criteria) on real and synthetic images to look for the best one. The selected filter will be that one having the least noise and the best step and line responses. After performing the comparison study, for a general purpose operation, the adaptive mean filter has been identified as the best choice.

1. Introduction

FLIR (Forward-Looking Infra-Red) image sequences have very low quality, due to the high frequency noise generated during the acquisition process in the image tube. So, the usefulness of their direct application for visualization or ulterior processing is seriously reduced.

As technology for infrared sensor devices is currently unable to provide low noise outputs, it is required to design a real-time system capable to minimize this noise. This quality improvement (increase of signal to noise ratio) must be achieved keeping all signal features, as there is no knowledge on the use of the output improved signal (visualization, object location, target tracking, ...).

The noise consists of random and uncorrelated variations of pixel amplitudes. Since they are uncorrelated, the noise high frequency components are bigger than those of the signal; so, low pass filtering can be used for noise reduction, having the following alternatives:

1 – *Linear low pass filtering*, but, unfortunately, it smears the edges and attenuates the narrow lines. This negative effect avoids its use for some applications.

2 – *Non linear low pass filtering*, where a noise reduction may be ideally obtained without altering the edges and the fine details. This kind of filtering strongly depends on the statistical noise properties and on the features of the images, ...

As we are considering FLIR sequences, this filtering should be performed in 3-D space-time domain, but for implementation reasons simpler filters are usually considered. So, bidimensional space filters operate on two of the three dimensions, not requiring big amounts of memory. On the other hand, temporal filtering requires, at least, one full field or frame memory, depending on the type of the sequence. Therefore, only 2-D space non linear filters will be considered.

This work is structured as follows: initially the study carried out is presented including the used image set and the selection criteria, after the applied filters are described being classified in four types, then the results are discussed performing a comparative study among the filters, and finally the conclusions are presented.

2. Description of the Study

MSE and local statistics have been used in order to evaluate the efficiency of the implemented filters. To help this procedure, synthetic noisy images have been built and the filters have been tested on them as well as on the raw FLIR images. The filtered images have been also subjectively evaluated by trained operator.

2.1. Images

As working with real FLIR images obtained in field test does not allow to perform error estimation, it has been required to design a set of synthetic images to carry out a statistical study. As is desirable to have as much similarity as possible between synthetic and real images, the next points have been considered for it:

1 – The object–background size ratio.

2 – The object–background intensity discontinuity.

3 – The noise (from FLIR images of uniform areas).

According to these points and considering the goal of the study (noise removal and details preserving) the next images, sized 128×128 pixels, have been created:

Image I. Consisting of two solid objects (one square and one triangular) on a darker background. Gray levels are 150 for background and 200 for objects. It serves to study how the filter smears solid edges.

Image II. Consisting of several variable width an direction lines on a darker background. Gray levels are 150 for background and 180 and 190 for lines. It serves to study how the filter preserves thin lines and fine details.

2.2. Selection Criteria

Being X_o the noise free image, X_f the filtered image, X_n the noisy image, and X any of the former, the statistical study is based on the following measures:

- MSE $= \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (X_o(i,j) - X_f(i,j))^2$

- MSE Quotient. The ratio between the MSE of the filtered image and that of the noisy one.

- Local Mean $(M_l) = \sum_{i=1}^{3} \frac{X(i)}{3}$

- Local Variance $(V_l) = \sum_{i=1}^{3} (X(i) - M_l)^2/3$

- Local RMS $= \frac{1}{F} \times \sqrt{\sum_{i=1}^{3} (X(i) - X_o(i))^2}$

3. Applied Filters

Several studies on different non linear filtering schemes have been presented during the last years [1-9], but there is a lack of comparative evaluation studies on images obtained from real enviroments, as only comparisons on artificial images have been conducted [10]. Here, a new comparison study is introduced to find the best non linear filter for FLIR images. Therefore, this study has been carried out on all the previous proposed filters, implementing the most appropriate and choosing their optimal parameters for the considered environment. They are listed below, being F.P. the optimal filter parameters and W the side of the square analysis/filtering window):

1 – L-type filters. Its output may be expressed as a fixed linear combination of the order statistics.

- α-TM filter [2]. F.P. are (W=3), and α (0.3, 0.4).

- Modified trimmed mean filter (MTM) [6]. F.P. are W=3 and the amplitude range q (12, 20).

- Double-Window modified trimmed mean filter (DWMTM) [9]. F.P. are W_1=3, W_2=5, and the amplitude range q (12, 20).

- K-NN filter [4]. F.P. are W = 3, and the number of elements K (3, 5).

- MK-NN filter [9]. F.P. are W = 3, and the number of elements K (3, 5).

2 – R-type filters. Its output may be expressed as an order statistics of a linear function of the input data.

- Wilconson filter [3]. F.P. is W = 3.

- Limited degree Wilconson filter (LDW) [10]. F.P. are W=3, the divisor D (2, 3), and a controller of the number of elements M (2, 3).

- FIR-Median hybrid filters (FMH) [8]. The best two filters have been used for the statistical study: UFMH, unidirectional FMH with a horizontal mask, and BFMH, bidirectional FMH with a diagonal cross mask.

- Multilevel filtering [8]. F.P. are the level of filtering (3) and the filter type of the previous subfilter outputs (UFMH with four masks).

3 – M-type filters. Its output is an estimator that can be described as a generalized form of maximum-likelihood estimates.

- Adaptive mean filter[9]. F.P. are W=5 and the intensity range C=12.

- Adaptive median filter. F.P. are W=5 and the intensity range C=12.

4 – Median class filters.

- Conventional median filter. F.P. is W=3.

- Separate median filter. F.P. is W=3.

- Max/median filter. F.P. is W=3.

4. Comparison Study on Experimental Results.

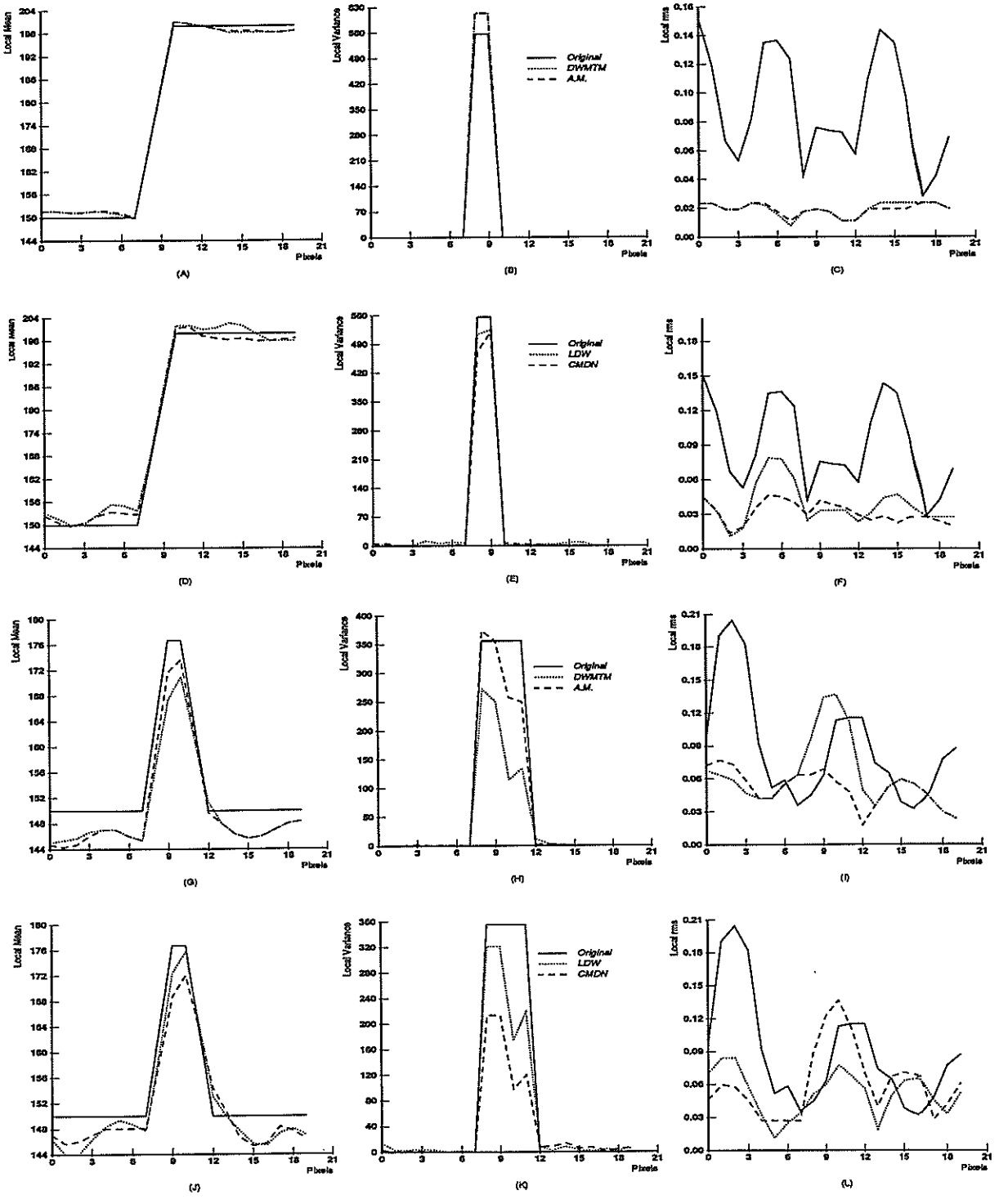
Statistical and subjective studies have been performed for all the previously introduced filters on all the test images. As is impossible here to discuss all the results, four filters (the best ones for each filter class) have been selected, based on the MSE criterion.

4.1. Noise Reduction. MSE Analysis

Table 1 presents for both test images the MSE and the MSE Quotient (R). These statistics are a measure of the deviation of the filtered image from the original

	MSE1	MSE2	R1	R2
L-type filters				
α-TM (α=0.3)	8.89	12.20	0.36	0.49
α-TM (α=0.4)	9.45	12.84	0.38	0.52
MTM (q=12)	9.19	13.17	0.37	0.53
MTM (q=20)	7.91	12.12	0.32	0.49
K-NN (K=3)	18.43	18.59	0.75	0.79
K-NN (K=5)	13.37	14.23	0.54	0.58
M-KNN (K=3)	10.66	14.41	0.43	0.58
M-KNN (K=5)	10.62	14.78	0.43	0.60
DWMTM (q=12)	4.23	8.66	0.17	0.35
DWMTM (q=20)	6.03	12.26	0.24	0.50
R-type Filters				
Wilcoxon Filter	11.91	12.78	0.48	0.52
LDW (M=2,D=2)	9.54	10.53	0.39	0.43
LDW (M=3,D=3)	12.27	13.43	0.50	0.54
LFMH	12.65	16.48	0.51	0.66
RFMH	7.64	15.31	0.31	0.62
MFMH	15.77	15.69	0.64	0.64
M-type filters				
A.Mean. (C=12,w=5)	3.90	4.38	0.16	0.17
A.Median. (C=12,w=5)	8.49	8.51	0.34	0.34
Median class				
Conv. Median	9.45	12.84	0.38	0.52
Separate Median	11.61	14.97	0.47	0.61
Max/Median	16.75	16.80	0.68	0.68

Table 1. Filter Results



Figures (A)-(L): Local statistics for the comparison study of the four filters

one. As this deviation can be caused either by the noise or by the bad filter response, the best filter, following this criterion, will be that one having the least MSE. Accordingly, the selected filters are:

- L-type. DWMTM (q=12).
- R-type. LDW (M=2, D=2).
- M-type. Adaptive Mean (AM) (C=12, W=5).
- Median Class. Conventional Median (CMDN).

4.2. Step Response. Local Statistics

The *Local Mean* evaluates the difference between the filtered image and the original one along the edge and the two sides regions. Figures (A) and (D) represent this function for the outputs of the four selected filters and the original image. DWMTM and AM behave very good, both for edge and sides. On the other hand, LDW and CMDN present worse responses, mainly in the sides, where oscilations from the mean exist.

The *Local Variance* evaluates the degree of edge blurring. Figures (B) and (E) represent the four local variances. DWMTM and AM give the best performance, whereas the other two present non-zero variance in the side-regions.

The *Local RMS* evaluates both the noise reduction and the step response. Figures (C) and (F) show the results of the four filters. The RMS of DWMTM and AM are very similar, and both very small refered to the original one. LDW give the worst result.

4.3. Fine Details Response. Local Statistics

The line response of the considered filters is always worse than that of the step one. Four lines of different width (one to four pixels) have been examined. Logically, as the width increases, the response upgrades. A one-pixel wide line is totally degraded with a $W = 3$ CMDN filter, therefore, only two-pixels wide lines are considered for the study, as wider lines should give better results.

Local Means are presented in figures (G) and (J) and *Local Variances* in figures (H) and (K). It is shown that the best filter response is given by AM filter. Here, *Local RMS* shows that the error can be caused by noise or a bad response, being represented in figures (C) and (F). The error increases (at the line) for the DWMTM and for the CMDN. The other two present a better line response, being the AM the best.

5. Conclusions

According to the experimental results, the following conclusions can be stated:

- The two best filters are the AM and the DWMTM, being the MSE small for both, and the step response similar, but the line response is better for AM.
- Subjective results confirm these ideas.
- AM is the best choice for general purpose.

References

[1] J.B.Bednar and T.L.Watt, *Alpha-Trimmed Means and Their Relationship to Median Filters*, IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-32, No. 1, pp. 145-153, February 1984.

[2] R.J.Crinon, *The Wilconson Filter: a Robust Filtering Scheme*, Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc., pp. 668-671, 1985.

[3] L.S.Davis and A.Rosendfeld, *Noise Cleaning by Iterative Local Averaging*, IEEE Trans. Syst. Man Cybern. SMC-8(9), pp. 705-710, 1978.

[5] Y.H.Lee and S.A.Kassam, *Generalized Median Filtering and Related Nonlinear Filtering Techniques*, IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-33, No. 3, pp. 672-683, June 1985.

[6] A.Nieminen, P.Heinonen and Y.Neuvo, *A new class of detail preserving filters for image processing*, IEEE. Trans. Pattern Anal. Mach. Intell., Vol PAMI-9, No.1, pp. 74-90, January 1987.

[7] S.R.Peterson and S.A.Kassam, *Edge Preserving Signal Enhancement Using Generalizations of Order Statistic Filtering*, Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc., pp. 672-675, 1985.

[8] I.Song and S.A.Kassam, *Nonlinear Filter Based on Generalized Ranks for Edge Preserving Smoothing*, in Proc. IEEE Int. Symposium on Circuits and Systems, pp. 401-404, 1986.

[9] C.A.Pomalaza-Ráez and C.D.McGillen, *An Adaptive, Non Linear Edge-preserving Filter*, IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-32(3), pp. 571-576, 1984.

[10] Y.Fong, C.A.Pomalaza-Ráez and X.Wang, *Comparision Study of Nonlinear Filters in Image Applications*, Optical Engineering, Vol. 28, No. 7, pp. 749-760, July 1989.

[11] Z.Mao and R.N.Strickland, *Image Sequence Processing for Target Estimation in Forward-Looking Infrared Imagery*, Optical Engineering, Vol.27 No. 7, pp. 541-548, July 1988.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

SEGMENTATION OF SPOT IMAGES BY CONTEXTUAL SEM

MASSON P., PIECZYNSKI W.

Groupe Traitement d'Images
Département Mathématiques et Systèmes de Communication
E.N.S.T de Bretagne BP 832 29285 Brest Cedex France.

We present a non-supervised method of bayesian contextual image segmentation. The problem of estimating the components of a mixture of distributions is treated by using a recent variation on the EM algorithm, called SEM. The algorithm obtained is tested on synthetic images and applied to the segmentation of SPOT image.

1. INTRODUCTION

The segmentation of SPOT images consists in attributing a $\omega \in \Omega = \{\omega_1, \ldots, \omega_k\}$ class to each $X_s \in R^3$ observation. The rectangular image I is divided into a finite number of elementary rectangles which as a set will be denoted S. It can be supposed therefore that on each elementary rectangle $s \in S$, the X_s measurement is unique. It is supposed furthermore that each $s \in S$ is asso-ciated with $\omega \in \Omega$ representing its "nature". For example, Ω could represent {water,forest,urban area}. Consequently we have two applications: $\xi: S \to \Omega$; and $X: S \to R^3$; where X is known and ξ unknown. Fixing a segmentation rule amounts to defining a Φ application which associates an application $\xi \in \Omega^S$ with each application $X \in (R^3)^S$. The very nature of this modeling excludes any determinist link between ξ and X : if ξ_s="forest", this forest may be more or less dense, which will give several possible values to X_s.. It is therefore necessary to model this "natural variability" by a probability distribution which would be dependent on the $\xi_s = \omega_m$ realization. This defines the conditional distributions which convey the "stochastic" link between ξ and X. For each $s \in S$, ξ_s can in turn be considered as the realization of a random variable: $s \in S$ may be seen as a place within the satellite's acquisition system, as it moves, the different "natures" or types of terrain "flash" through s.

The realization of ξ_s (type of terrain which "is to be found in s" when the picture is taken) is consequently random.

2. BAYESIAN APPROACH

The problem of the segmentation of SPOT images (statistical approach) can be expressed in the following terms: two collections of random variables $(\xi_s)_{s \in S}$, $(X_s)_{s \in S}$ (random fields) are considered. From an observed realization $X=x$ we wish to estimate the unobservable realization of ξ . Let us denote $P_\varepsilon = P(\xi = \varepsilon)$ and f^ε the density of the distribution of X conditional to $\xi = \varepsilon$. This defines the distribution of (ξ, X) and therefore the conditional distribution of ξ knowing that $X=x$ (*a posteriori* distribution) which we shall denote P^x . We can then make the estimate we wanted by using the Bayesian rule r (this rule is shown to minimise the probability of making a poor estimate) defined by :

$$r(x) = \hat{\xi} \Leftrightarrow P^x\left[\hat{\xi}\right] = \sup_\varepsilon P^x[\varepsilon] \qquad (1)$$

This rule can also be expressed as follows :

$$r(x) = \hat{\xi} \Leftrightarrow P_{\hat{\xi}} f^{\hat{\xi}}(x) = \sup_\varepsilon P_\varepsilon f^\varepsilon(x) \qquad (2)$$

The functions $g_\varepsilon = P_\varepsilon f^\varepsilon$ are called "discriminating". From the basis of (2) it is impossible to

find $\hat{\xi}$ directly , due to the excessively high number of possible ε realizations (equal to k^n, n=card (S)). This difficulty can be overcome by assuming the ξ field markovian and the random variables $(X_s)_{s\in S}$ to be independent conditionally to any realization of ξ . This modeling makes it possible to simulate the *a posteriori* distribution (with the Gibbs sampler for example) which forms the basis of the algorithm "simulated annealing" of D.Geman which makes possible the construction of a sequence (ε_n) converging to $\hat{\xi}$.

Let us denote ,for each $s\in S$, $P_m^{x,s} = P^x[\xi_s = \omega_m]$ The MPM algorithm [12] consists in estimating each ξ_s separately by :

$$\hat{\xi}_s = \omega_m \Leftrightarrow P_m^{x,s} = \sup_{1\le q\le k} P_q^{x,s} \qquad (3)$$

The *a posteriori* marginal distributions $P^{x,s}$ being estimated by frequencies obtained in simulations. These two methods are called "global". The contextual method consists in estimating the realization of each ξ_s from $X_V = (X_t)_{t\in V}$ with $s\in V$. The bayesian rule is expressed as follows:

$$\hat{\xi}_s = \omega_m \Leftrightarrow g_m(x_V) = \sup_{1\le q\le k} g_q(x_V) \qquad (4)$$

Let n = card (V) , $V^* = V - \{s\}$. For each $\varepsilon^* \in \Omega^{n-1}$ let us put $P_{m,\varepsilon^*} = P[\xi_s = \omega_m, \xi_{V^*} = \varepsilon^*]$, f^{ε} the distribution of X_V knowing $\xi_V = \varepsilon$. The discriminating functions g_q above are expressed as follows :

$$g_q(x_v) = \sum_{\varepsilon^*} P_{q,\varepsilon^*} f^{(\omega_q,\varepsilon^*)} \qquad (5)$$

They can be determined, therefore, as soon as the P_ε distribution of ξ_V and the conditional distributions f^{ε} knowing $\xi_V = \varepsilon$ are determinated. Use of global methods requires knowledge of the distribution of ξ (Gibbs distribution) and conditional distributions, the estimate of the parameters defining these distributions presents a difficult preliminary statistical problem ([6], [4]). The contextual methods seems to us to be better adapted to the segmentation of SPOT images. The preliminary statistical problem is that of estimating the components of a mixture of distributions. Our study proposes the use of a recent algorithm,the SEM,the good results of which are expounded in ([2], [3]).

3 . ALGORITHM SEM

Let $V_1,\ldots,V_n$ be a sequence of contexts in S , let $x_i = x_{Vi}$. Let us denote $E = (\varepsilon_1,\ldots,\varepsilon_q)$, with n = card(V) and $q = k^n$, the set of possible realizations of ξ_V and f_i conditional (knowing $\xi_V = \varepsilon_i$) distributions of X_V, which will be supposed gaussian. Each X_s takes its values in R^3 , f_i is then a function from R^{3n} to R defined by the mean m_i and the covariance matrix Γ_i . Let Π be the distribution of ξ_V :

$$\Pi_j = P[\xi_v = \varepsilon_j]$$

The SEM estimates the parameters:

$\Pi = (\Pi_1,\ldots,\Pi_q)$, $m = (m_1,\ldots,m_q)$, $\Gamma = (\Gamma_1,\ldots,\Gamma_q)$

by an iterative proceding from the sample $x_1,\ldots,x_n$. The process is :

Initialisation:

One defines , for each x_i , a distribution on E :

$$\Pi^o(x_i) = (\Pi_1^o(x_i),\ldots,\Pi_q^o(x_i))$$

Step S

For each x_i we pick up a realization ε of ξ_V in the set E with the probability $\Pi^o(x_i)$. This gives a partitioning $Q_1,\ldots,Q_q$ of the sample $x_1,\ldots,x_n$ with $Q_j = \{x_i / \varepsilon_j$ is the result of the picking from the distribution $\Pi^o(x_i)\}$.

Step M

We estimate each (m_j,Γ_j) on Q_j via classical estimators (empirical means and covariances). We define :

$(m_j^o, \Gamma_j^o) = (\hat{m}_j, \hat{\Gamma}_j)$

In the same time Π^o is estimated as follows:

$$\Pi_j^o = \frac{card(Q_j)}{n}$$

Step E

Once the parameter (Π^o, m^o, Γ^o) obtained ,we defined ,for all x_i, $\Pi^1(x_i)$ as the *a posteriori* distribution. Then we return to step S.

4. RESULTS

4.1 Simulations

We consider a two class (Image 1) Markovian field corrupted with a real gaussian noise spatially independent (Image 2) and spatially correlated (Image 3). The corresponding means are $m_1=1$, $m_2=3$, and dispersion $\sigma^2= 1$. The correlated noise was simulated by the moving average (M.A), the covariance between two neighbouring pixels being 0.4. Image 4 (respectively Image 5) is a restoration of Image 2 by the blind method (respectively contextual). Similarly, Image 6 (respectively Image 7) is a restoration of Image 3 by the blind method (respectively contextual). Images 8 and 9 are restorations of Images 2 and 3 by Koontz's algorithm. T designates the rate of correctly classified pixels.

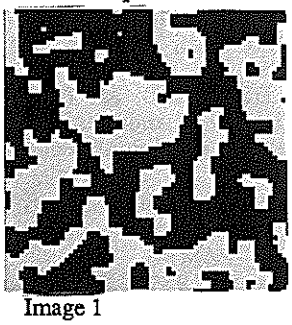


Image 1

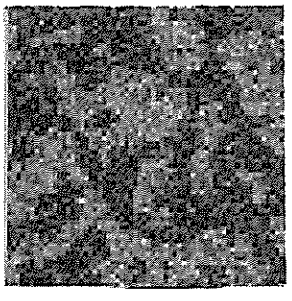


Image 2

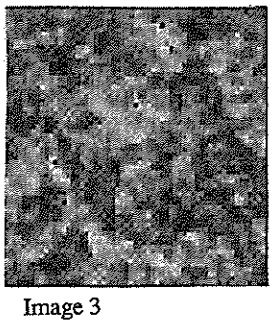


Image 3

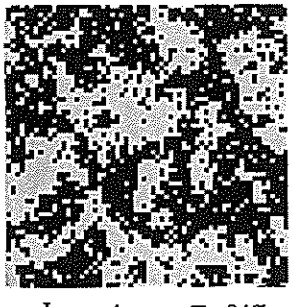


Image 4 T= 84%

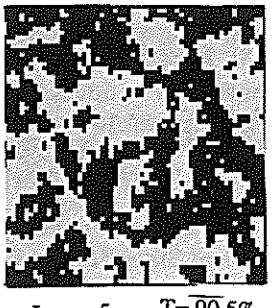


Image 5 T= 90.5%

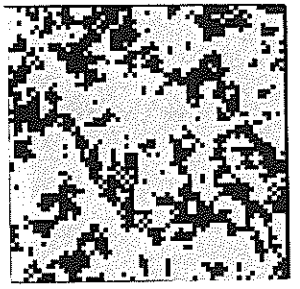


Image 6 T= 76%

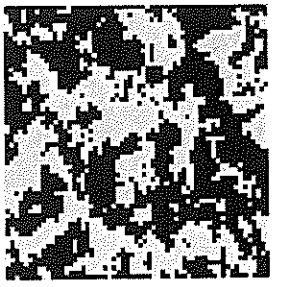


Image 7 T= 82%

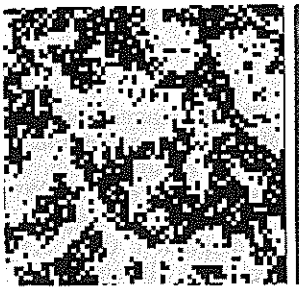


Image 8 T= 82.5%

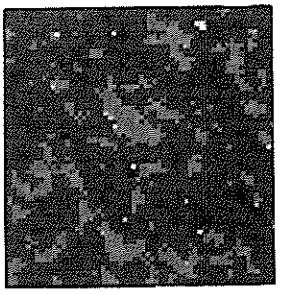


Image 9 3 classes

4.2 Real image

Image 10

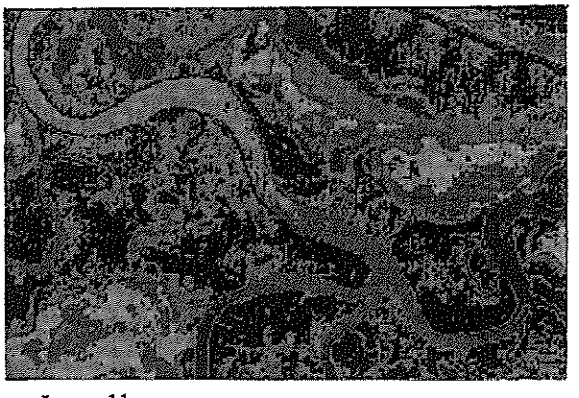


Image 11

Image 12

Image 10 represents a real SPOT image, Image 11 its segmentation by KOONTZ's algorithm and Image 12 the segmentation by the blind method. The result of the segmentation of the real SPOT image (Image 10) by the blind method (Image 11) seems to contain more information that its segmentation by Koontz's algorithm (Image 12).

5. CONCLUSION

Non supervised segmentation appears as a key problem in image segmentation because, in practice, the parameters determining the distribution of the couple (ξ,X) are not known. Compared with existing techniques ([4], [6], [10]), our method seems relatively simple to implement; however, it cannot be justified theoritically.
Numerous simulations make the following conclusions possible :
-In both cases (independent or correlated noise) the contextuel method is more efficient than the blind method.
-The efficiency of the contextual method deteriorates as the spatial correlation of the noise increases.
-Authors generally consider spatially independent noise, because taking spatial correlation into account is difficult when considering "global" methods. This simplification seems rather strong, for on the one hand, the human eye is sensitive to this correlation (Image 2 and Image 3), and on the other hand the real noise ("natural variability") is manifestly correlated.
-Spatial correlation of the noise seriously lowers the efficiency of Koontz's algorithm (appearance of false class, Image 9).
-The contextual method is less efficient, in terms of rate of correct classification, when the noise becomes spatially correlated (Image 5,Image 7). However, visually, the difference is not so important.

REFERENCES

[1] BESAG J.-1986- On the Statistical Analysis of Dirty Pictures , JRSS b - nº 48.
[2] CELEUX G.- DIEBOLT J.-1986- L'algorithme SEM: un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités-Rev.Stat.Appli. Vol 34 nº 2.
[3] CELEUX G. - DIEBOLT J. - 1986 - Comportement asymptotique d'un algorithme d'apprentissage probabiliste pour les mélanges de lois de probabilités. Rapport de recherche INRIA nº 563.
[4] CHALMOND B.-1988- An iterative Gibbsian technique for simultaneous structure estimation and reconstruction of M.ary images. Preprint , Université de Paris-Sud, Mathématiques, Batiment 425, 91405 ORSAY, France.
[5] DEMPSTER-LAIRD-RUBIN-1977-Maximum-likelihood from incomplete data via the EM algorithm - JRSS.B - Vol.39.
[6] DEVIJVER P.A.,DEKESEL M.-1988- Champs aléatoires de Pickard et modélisation d'images digitales. Traitement du Signal. Vol. 5 nº5.
[7] GEMAN S. and GEMAN D. -1984- Stochastic relaxation - Gibbs distributions and the bayesian restoration of images -IEEE -PAMI-6.
[8] GUYON X., YAO J.-1987- Analyse discriminante contextuelle, Fifth International Symposium of Data Analysis and Informatics (BP 105 , 78153 Le Chesnay Cedex , INRIA).
[9] KOONTZ, NARANDRA, FUKUNAGA -1976- A graph-theoretic approach to non parametric cluster analysis - IEEE Trans.- Comput. Vol C-25,no 9,pp. 936-944.
[10] LAKSHMANAN S., DERIN H.-1989-Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing . IEEE Trans. PAMI vol.11 nº 8.
[11] MARDIA K.V.- 1989- Markov models and Bayesian methods in image analysis . Journal of Applied Statistics , vol . 16 nº 2.
[12] MARROQUIN,MITTER,POGGIO-1987 Proba-bilistic Solution of Ill- Posed problems in computing vision - JASA nº 82 .
[13] PIECZYNSKI W. -1989 -Estimation of context in random fields-Journal of Applied Statistics, vol.16 nº 2 .
[14] QIAN W. , TITTERINGTON D.M.-1989- On the use of Gibbs Markov chain models in the analysis of images based on second-order pairwise interactive distributions . Journal of Applied Statistics, vol.16 nº 2 .

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

SPOT IMAGE MOSAIC AND DYNAMIC PROGRAMMING

Pascale POUSSET - MS2i - LTIS -
BP 235 - 78052 St Quentin en Yvelines Cedex

Marie-Lise DUPLAQUET - ONERA-DES
BP 72 - 92322 Chatillon Cedex - FRANCE

Seam-Point searching algorithms intend to eliminate spurious edges that may occur on mosaic, especially when the images have been acquired at different dates. This paper presents a dynamic programming method for seeking the optimum seam-line on the overlap area of SPOT images. Bellman's algorithm is used to seek the least-cost path on the overlap area. The cost-function to minimize takes into account intensity differences and common edges.

1-Introduction.

Building up image mosaic, i.e. gathering several images into a larger one, is frequently used in remote sensing. Once manual <WOLF-83> or interactive <BERN-75>, this operation has become automatic thanks to improvements in picture processing techniques. These automatic methods deal with LANDSAT or SPOT images (use of multispectral data possible <LIST-86>) as well as radar <LEWI-70> or aerial images <CHEN-86>. However, the quality of the mosaic image always depends upon the geometrical superposition of the original images and their relative intensity variations.

The mosaic problem can be set as follow: after first geometric and intensity corrections, find the best seam-line across the overlap area. Try to avoid bad choices: spurious edges will alter further interpretations. Panchromatic SPOT data are very sensible to this problem because of their high resolution (10m x 10m per pixel). Two images taken at very different dates show intensity differences in several local areas the seam-line must avoid.

After a quick description of the previous calibration steps, this paper presents an overview of existing methods and two major improvements: a complex criterion image and an application of dynamic programming. Results are shown on SPOT panchromatic images.

2-The calibration between the images.

Geometric and intensity coherence on the overlap region of the mosaic image is necessary for further treatments and interpretations, either manual or interactive.

- Geometric superposition can be achieved by transforming the geometry of an image into the geometry of the other one (this is called "level S1" for SPOT images). However, calibration of both images in order to allow superposition with maps is better ("level 2"), and high differencies in ground elevation induces stereo effects that must be corrected by introducing a Digital Terrain Model ("level 3").

- Intensity coherence is achieved by global adjustment of the images (equalization of average and variance). Local adjustment on each line may be used, but it is not compatible with our specifications (see <POUS-89>).

3-Some mosaicking methods.

Several methods have been proposed for seeking the seam-line :

-Milgram <MILG-75> seeks the best point at each line near the last chosen point. The quality of the seam-line depends on the first chosen point, while backtracking is not possible.
In <MILG-77>, he introduces a Dynamic Programming algorithm: at each point of row r, the best choice between 2u+1 previous points of row r-1 is made. Quality of the path is measured by the sum of intensity differences along the path. Despite better results, two problems appear: if u is small, the path is too "rigid", if u is large, vertical discontinuities are visible.

-Shiren & al. <SHIR-89> propose a two-dimensional seam-point searching to avoid vertical edges. But the amount of operations at each point is not compatible with a dynamic programming method on large images.

-A completely different method is used by Tianxi <TIAN-85> who modelises the seam-line as a polynomial curve, using a least squares optimization. Good global results are obtained but the constraint could be locally too strong.

In all cases, smoothing the intensity values near the seam-line is necessary. As we do not want to implement this last step, we chose a dynamic programming method, with a more complex *criterion* value to measure the quality of the path, and a particular set of *allowed directions* from a point to its successors.

4-The criterion image.

The criterion image represents the local incompatibilities that remain between the two images after the calibration step: a value measuring the interest to take a point in the seam-line is calculated on the overlap area. In most automatic applications, only intensity differences are computed. But for manual mosaicking, operators try to follow edges so that the seam-line does not appear as a spurious edge.

Consequently, our criterion image combines two images:
-a *disparity* image D: average of the absolute differences in a neighborhood of the point (2u+1 x 2u+1 square) :

$$d_{kl} = \sum_{i=-u}^{u} \sum_{j=-u}^{u} \left| f(k+i,l+j) - g(k+i,l+j) \right|$$

f and g are the left and right images.

-a *common edges* image S: a Sobel mask is computed simultaneously on both images. We take the lowest value when transition signs are identical and force to 0 when opposite. The image is complemented (black edges on a white background).

After normalization, the two images D and S are linearly combined. Proportion between disparity and common edges depends on the quality of the images: in case of blurred images, disparity must be preponderant while the edges detected are thick and could lead to false common edges.

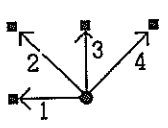


fig 1: point k,l and its 4 possible predecessors.

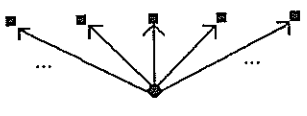


fig 2 : Milgram's allowed predecessors.

5-Seeking the seam-line.

5.1. Dynamic Programming.

The mosaicking problem is one of optimization under constraint: how to minimize the sum of the criterion values along the seam-line while keeping the curve continuous. Since the points are selected one by one, the problem can be broken down into several stages, each stage dealing with one point depending on previous choices. This division of the problem leads us to use a dynamic programming algorithm.

Introduced by Bellman in the last 50's <BELL-65>, dynamic programming was first used to solve classic optimization problems. Later, it gave good results in speech recognition and in picture processing <WU-87>,<OHTA-85>.

Here, dynamic programming corresponds to a least-cost path search in the graph deduced from the criterion image, using Bellman's algorithm (see <SAKA-84> for mathematical formulation). Termination points are fixed by the relative position of the images. Each point, from the starting point (top left) to the endpoint (bottom right), is treated in the video raster order (symmetric order and allowed directions must be chosen to go from top right to bottom left).

5.2. The algorithm.

The best choice at point k,l is made between 4 possible moves (see fig 1). Compared to Milgram's approach, the path is always continuous and an horizontal move is permitted which takes into account each by-passed point. No simple direction can be added: Bellman's algorithm needs graphs without cycle.

The potential value z_{kl} measures the quality of the path, n_{kl} its length and x_{kl} indicates the direction selected. The algorithm is described below:

Initial step: potential of the begin point is set to 0.

Step k,l: the new values are computed

$$z_{kl} = \min \left\{ \begin{array}{c} (z_{k-1,l} \cdot n_{k-1,l} + c_{k-1,l}) / (n_{k-1,l} + 1) \\ (z_{k-1,l-1} \cdot n_{k-1,l-1} + c_{k-1,l-1}) / (n_{k-1,l-1} + 1) \\ (z_{k,l-1} \cdot n_{k,l-1} + c_{k,l-1}) / (n_{k,l-1} + 1) \\ (z_{k+1,l-1} \cdot n_{k+1,l-1} + c_{k+1,l-1}) / (n_{k+1,l-1} + 1) \end{array} \right\}$$

x_{kl} = 1, 2, 3 or 4 and n_{kl} is incremented.

All x_{kl} values must be held in memory, while only two lines (l and l-1) of potential and length are currently necessary.

Last step: when all points have been treated, the path is determined from the end point to the begin point by following the x_{kl} directions (fig 3).

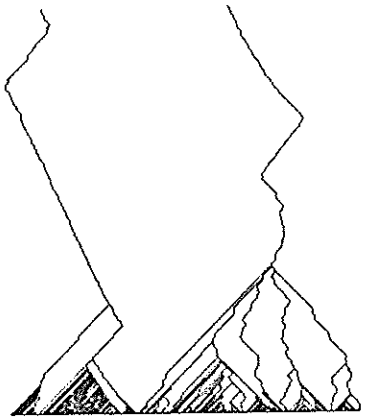


fig 3: all best paths from the current line. The starting point is not fixed and the horizontal move is not allowed.

Now the seam-line is determined. The mosaic image is constructed simply by copying left image pixels until the seam-point and right image pixels after. No smoothing is made: it was one of our constraints, but it would also degrade intensity information near the real edges chosen by the seam-line.

6-The results.

A complete mosaic between two SPOT panchromatic images (6000x6000 pixels each one) was computed. The overlap region was about 300 pixels wide. Figure 4 shows part of the result on the overlap region, at full resolution. High intensity differences between left and right images are visible, but the seam-line is fairly good.

7-Conclusion.

The results obtained with SPOT scenes having high intensity differences prove the usefulness of introducing common edges in the criterion images. Formalizing the search for the seam-line as a least-cost path problem allows a global and optimum solution to be reached. Since no smoothing is computed, the mosaic image has the same intensity properties as the source images.

Improvements could be made in the formulation of the criterion, like introduction of others informations (presence of clouds...). The method, giving good results for panchromatic SPOT images, could be extended to multispectral or aerial images.

References:

<BELL-65> R.Bellman, R.Kalaba: "Dynamic Programming and Modern Control Theory", *Academic Press Inc.* 1965.

<BERN-75> R.Bernstein, D.G.Ferneyhough Jr: "Digital Image Processing",*Photogrammetric engineering & Remote Sensing*, Vol 41, N° 12, Dec 1975, pp 1465-1476

<CHEN-86> Yin-Pao Chen & B.Krishna Mudunuri: "An Anti-Vigneting Technique for Super Wide Field of View Mosaicked Images", *Journal of Imaging Technology* , Vol 12, N° 5, Oct 86, pp 293-295.

<LEWI-70> A.J.Lewis, H.C.Macdonald: "Interpretive and Mosaicking Problems of SLAR Imagery". *Remote sensing of environment*, N° 1, 1970, pp231-236.

<LIST-86> List FK, Meissner B, Pohlman G: "Landsat-MSS Remote Sensing and satellite cartography". *Proc. of IGARSS'86 Symposium* Zurich 8-11 Sep. 86, pp 1503-1510.

<MILG-75> David L Milgram: "Computer Methods for Creating Photomosaics" *IEEE Trans. on Computers*, Vol C-24, Nov 1975, pp 1113-1119

<MILG-77> D.L.Milgram: "Adaptative Techniques for Photomosaicking". *IEEE Trans. on Computers,* Vol C-26, N° 11, Nov 1977, pp 1175-1180.

<OHTA-85> Yuichi Ohta, Takeo Kanade: "Stereo by Intra- and Inter-Scanline Search Using Dynamic Programming", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol PAMI-7, N° 2, March 1985.

<POUS-89> P.Pousset, M.L.Duplaquet "Mosaique d'images SPOT et Programmation Dynamique", *Proc. of AFCET*, Paris 29 Nov.-1st Dec. 1989, pp 1099-1106.

<SAKA-84> M.Sakarovitch :"Optimisation Combinatoire", *Hermann, Edition des sciences et des arts.* 1984.

<SHIR-89> Yang Shiren, Li Li, Gao Peng :"Two Dimensional Seam-Point Searching in Digital Mosaicking", *Phot. Eng. & Remote Sensing*, Vol.55, N° 1, Jan 89, pp 49-53.

<TIAN-85> Wang Tianxi :"A New Mosaicking Method for Landsat Remote Sensing Images", *Kexue Tongpao (Sci Bulletin)*, Vol 32, N° 12, Jun. 87, pp 854-859.

<WOLF-83> P.R.Wolf :"Photomaps and Mosaics", *Elements of Photogrammetry, Mc Graw-Hill Book Company*, 1983, pp 211-224.

<WU-87> Yifeng WU:"Application de la Programmation Dynamique au Recalage d'images", *Rapport de these ENST-Paris* Oct.87.

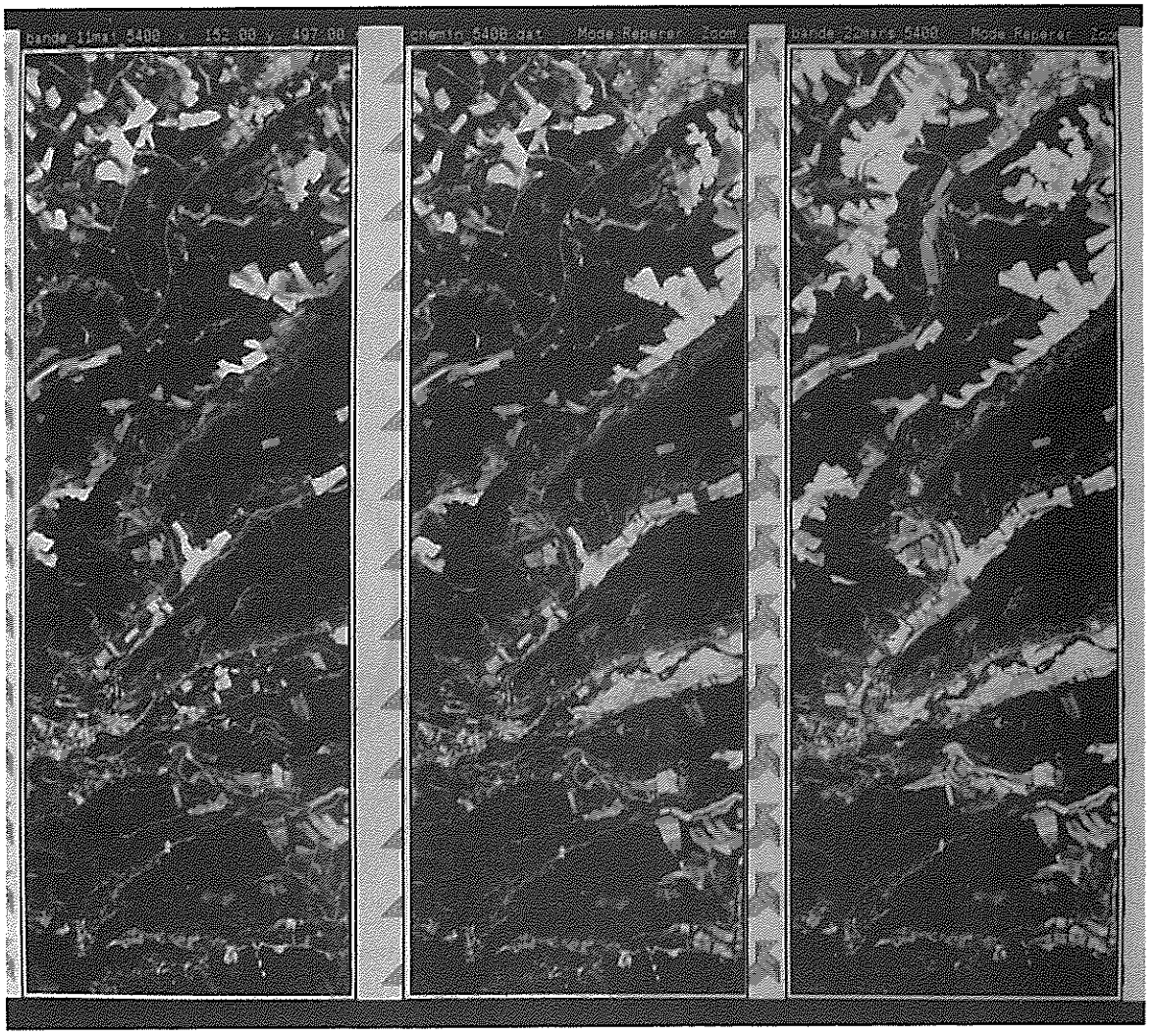


fig 4: left image, mosaic and right image.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

MATCHING MULTI-SOURCE IMAGES : SPOT IMAGE - GEOGRAPHIC MAP

Michel ROUX, Jaime LOPEZ-KRAHE, Henri MAÎTRE

Département IMAGES, TÉLÉCOM PARIS
46 Rue Barrault - 75634 PARIS Cedex 13 - FRANCE

ABSTRACT : In this paper we present a method for matching a SPOT image and a geographic map using high level structures, like crossroads. We first assume the detection of linear networks and crossroads on both images. A geometric transform between the two networks is calculated using matching crossroads as control-points. This transform is then refined by a local analysis on each control-point. The method is tested on real images and the results are given.

I. INTRODUCTION

The resolution of satellite images (10m/pixel for panchromatic SPOT images) allowed to tackle the problem of road extraction in those images. However the different researches in this field show the difficulties to extract a reliable road-network from the only SPOT images with classical image processing methods. Results are usually noisy and incomplete. This kind of problem can not be solved only with the information contained in the image, we have to introduce some external information [4].

The method proposed in this paper uses a fundamental operation of computer vision which is the matching of an image with a model of the scene. The model for high resolution satellite images is the corresponding geographic map and the matching operation is then useful for many applications : cartographic data acquisition, map up-dating or image-based navigation [6,11].

The image-to-model matching results from the matching of features extracted from the image to objects present in the model [2,5,8]. Rather than use low-level features like pixels or line segments, we did prefer to match high-level structures like crossroads, whose detection is more suitable, even if they have not a precise localization.

We first proposed a technique to extract the crossroads from SPOT images and from digitized maps. Then the image to map matching process is decomposed in two steps : 1.) global coarse matching of the two images using the crossroads as control-points, 2.) matching refinement by a local analysis on each control-point. Finally some experimental results will be given.

II. CROSSROAD DETECTION ON A SPOT IMAGE AND ON A DIGITIZED MAP

This detection requires first the extraction of the linear network from the SPOT image and the road-network from the map. Two different processings assume those extractions, then the crossroads are obtained in the same way in both networks.

II.1. Lines detection on SPOT images.

Among all the methods proposed in the past for the extraction of linear networks in aerial or satellite imagery, mathematical morphology provides a set of transforms well suited for this problem and of easy implementation [3].

Morphological transforms combination with 3x3 structuring elements allows the detection of lines on grey level images :

The "top-hat" transform for light lines :

IMAGE - dilation (erosion (IMAGE))

Dual transform of "top-hat" for dark lines :

erosion (dilation (IMAGE)) - IMAGE

Then a thresholding selects the most contrasted lines of the transformed image. A thinning step is needed to obtain a thin network, and the removal of short lines efficiently clears this skeleton. Nevertheless the final linear network obtained after this process is still incomplete and noisy, especially in very textured regions, like urban areas.

II.2. Road-network extraction from a map

The map we use at this stage is a commercially available map at the 1/50,000 scale , which is scanned in 256 grey levels with a pixel equivalent to approximately 10 meters on the ground (digitization at 50 pixels/cm).

The extraction of a complete road-network from a such digitized map is a quite difficult problem, because of all printed characters and other symbols which perturb the detection. All the solutions proposed to this problem are dependant from the kind of map used and from the representation of roads in the map [10,11]. The method presented here is well suited for maps where roads are bounded by two parallel dark lines, as in our case (fig. 4.a)[5].

The extraction is done by filtering the image with the structuring element : ■□■, in the 4 directions.

Then the same process is applied to this road-network as to the one extracted from the SPOT image : thresholding, thinning, removal of short lines.

The main problem met during this detection is due to the presence of printed characters and other symbols, which would require a specific pre-processing, in order to get a complete and less noisy network [1]. However the removal of short lines allows to suppress a large amount of noise and the fact that lines are often broken will not disturb too much the detection of crossroads like it is presented in §II.3.

For colored maps, a digitization with RGB filters is also possible and allows to obtain a more reliable and more complete network of the main roads.

II.3. Crossroad detection in both networks

The two networks, coming from the SPOT image and from the map, are then processed in the same way. A linear approximation of the networks gives a vectorial representation of the networks, and the local orientation of lines is calculated.

Then the aligned segments are grouped, the junctions (X or T shaped) are detected, and the crossroads determined : a crossroad is defined as at least one junction, but usually gathers several junctions which are close enough :

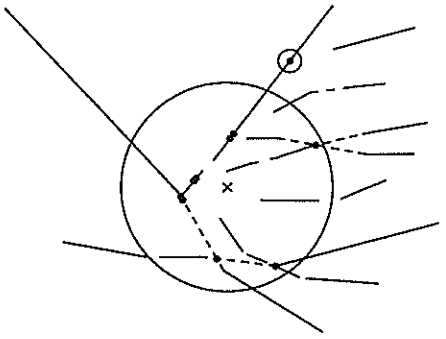


Fig. 1 : Crossroads detection. Junctions are represented by points, crossroads by circles.

The results of these detections on both networks are presented figure 3.b and 4.b.

III. MATCHING OF THE SPOT IMAGE ON THE MAP

Our purpose is now to determine the transform that allows to pass from the SPOT image to the digitized map. Despite the spatial distortion due to the image acquisition system of the SPOT satellite, we restrict our research to linear transforms (composition of one rotation, one homothety and one translation). More sophisticated methods would use polynomials or B-splines.

A linear transform Φ is determined by 4 parameters :

- rotation angle : θ
- homothety coefficient : α
- translation parameters : dx, dy

To get these parameters, the matching process uses two kinds of calculations :

- Two control-points define exactly one linear transform (a control-point is the association of a point of one image with a point of the second image).
- A linear transform can be determined with a set of control-points using a least squares approximation.

III.1. Crossroads matching

At that point, each crossroad is reduced to the mass centre of the junctions of which it is composed (see §II.3).

A control-point is then a couple of points like : (A_i , B_j), with A_i a crossroad of SPOT image and B_j a crossroad of the map. From two such pairs, a transform Φ is computed matching SPOT image on to the map. This transform provides an image $\Phi(A_k)$ of any crossroad A_k. If there exists one B_l so that *distance* $(\Phi(A_k) , B_l)$ is less than a threshold δ we decide that the couple (A_k , B_l) supports the transform Φ. With all the couples that supports the transform Φ, we compute by a least squares approximation a new value of Φ. This process is iterated until the control-points set is stable.

For each pair of control-points, a final transform Φ is obtained, associated to a set of control-points :

$$\Gamma_\Phi = \left\{ (A_i,B_j) \,/\, distance\ (\Phi(A_i) , B_j) \leq \delta \right\}$$

To each control-point (A_i,B_j) of Γ_Φ corresponds an elementary cost $c_\Phi\ (A_i,B_j)$, which is the residual error for this control-point of the least squares approximation of the transform Φ :

$$c_\Phi\ (A_i,B_j) = distance\ (\Phi\ (A_i) , B_j)$$

The global cost C_Φ of the transform Φ is the mean of the control-points of Γ_Φ elementary costs :

$$C_\Phi = \frac{\sum_{(A_i,B_j) \in \Gamma_\Phi} c_\Phi\ (A_i , B_j)}{Card\ (\Gamma_\Phi)}$$

Among the transforms obtained from all the possible pairs of control-points, we keep the transform Φ_f with the lowest cost. The matching of SPOT image on the map using this transform Φ_f is shown figure 5.

III.2. Networks matching refinement

The transform obtained by the coarse crossroads matching process gives an imprecise match, because of the lack of precision on the crossroad localization. A local analysis is needed in order to determine a precise control-point for each couple of associated crossroads.

For each couple of crossroads, we attempt to match the segments that contribute to the SPOT image crossroad and those to the map crossroad.

Using the parameter θ of the transform Φ_f, it is possible to do a first selection of couples of segments that can be matched.

Then, using compatibility criteria, a dynamic programming technique determines the best matching of these segments.

The control-point corresponding to the couple of crossroads (A,B) will be the couple of points (G_A, G_B), where G_A (respectively G_B), is the mass centre of the intersections of the SPOT image segments (respectively the map segments) kept after the dynamic programming process :

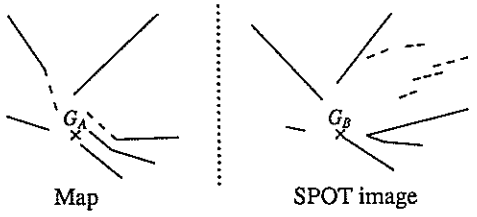


Fig. 2 : Segments matching by dynamic programming. The segments kept are drawn in full line. The pair (G_A,G_B) determines the new control-point.

The final transform is computed with all the new control-points and figure 6 shows the final matching.

IV. Conclusion

The efficiency of the method has been checked by matching an area in south east of France. The completely blind method presented here provides matching results as good as a manual registration made by a human observer.

The originality of this presentation is to show the possibility of using high-level structures for the matching of two multi-source images rather than using low-level features like pixels or straight lines.

References :

[1] T. J. Amin, R. Kasturi : *Map data processing : recognition of lines and symbols* Optical Engineering, April 1987, Vol. 26, No. 4, pp 354-358.

[2] J. K. Cheng, T. S. Huang : *Image registration by matching relational structures,* Pattern Recognition, Vol. 17, N° 1, pp 149-159, 1984.

[3] I. Destival : *Mathematical morpholgy applied to remote sensing,* Acta Auronautica, Vol. 13, N° 6/7, pp 371-385, 1986.

[4] M. A. Fischler, J. M. Tenenbaum, H. C. Wolf : *Detection of Roads and Linear Structures in Low-Resolution aerial Imagery Using a Multisource Knowledge Integration Technique,* CGIP, Vol. 15, pp 201-223, 1981.

[5] T. Matsuyama, H. Arita, M. Nagao : *Structural matching of line drawings using the geometric relationshop between line segments,* CVGIP, Vol. 27, pp 177-194, 1984.

[6] G. Medioni : *Matching Images Using Linear Features,* IEEE Trans. on PAMI, Vol. PAMI-6, No. 6, Nov. 1984, pp 675-685.

[7] T. Nagao, T. Agui, M. Nakajima : *An automatic road vector extraction method from maps,* 9th ICPR, Rome, Italie, Nov. 1988, pp 585-587.

[8] T. Skordas, R. Horaud : *Stereo Correspondence Through Feature Grouping and Maximal Cliques,* IEEE Trans. on PAMI, Vol. PAMI-11, No. 11, Nov. 89, pp 1168-1180.

[9] R. Solberg, M. Robb : *Satellite Imagery for semi-automatic map revision,* Proc. of IGARSS '88 Symposium, Edinburgh, Scotland, 13-16 Sept. 1988, pp 1179-1183.

[10] S. Suzuki, M. Kosugi, T. Hoshino : *Automatic line drawing recognition of large-scale maps,* Optical Engineering, July 1987, Vol. 26, N° 7, pp 642-649.

[11] Z. Zhu, Y. Kim : *Algorithm for automatic road recognition on digitized map images,* Optical Engineering, Vol. 28, N° 9, pp 949-954, Sept. 1989.

Fig. 3.a : panchromatic SPOT image (512x512 pixels). (© SPOT IMAGE)

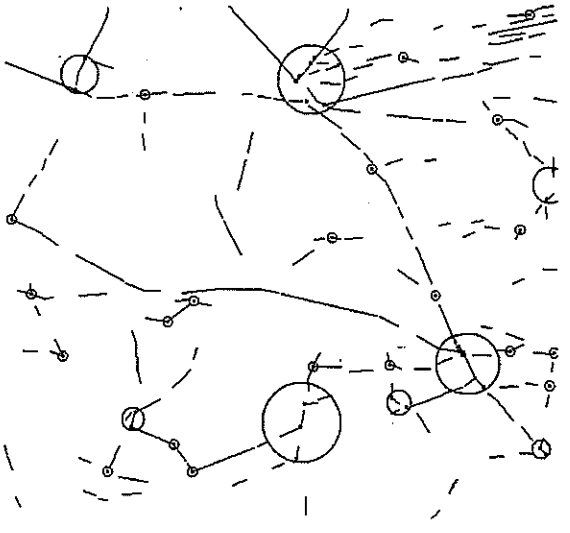


Fig. 3.b : Crossroads detected on the SPOT image.

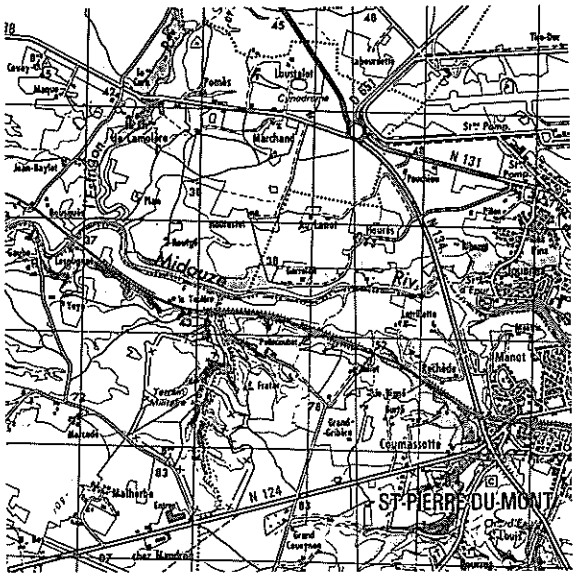


Fig. 4.a : Map at the 1/50,000 scale. (© IGN, nº 1542, série orange)

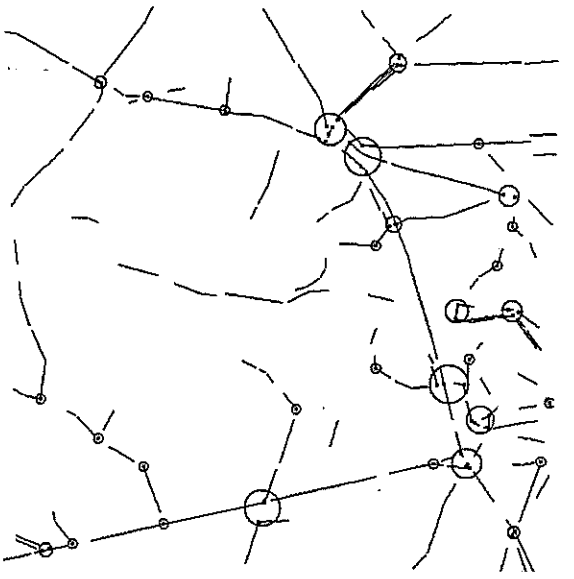


Fig. 4.b : Crossroads detected on the scanned map.

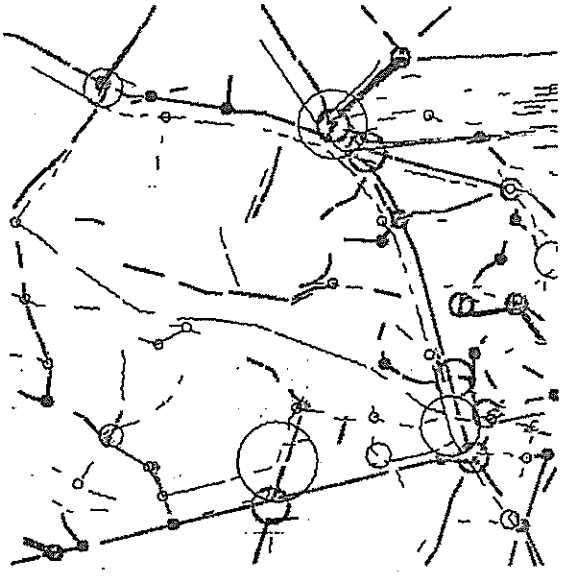


Fig. 5 : Global coarse matching using the mass center of crossroads as control-points.

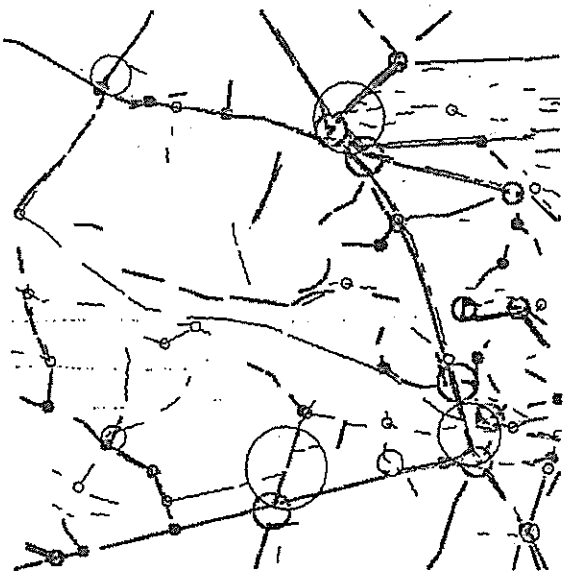


Fig. 6 : Matching refinement after local analysis on each control-point.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

AN AR BASED ALGORITHM FOR IMAGE REGISTRATION

P. Concetti, G. Orlandi(*), F. Piazza

Dept. of Electronics and Automatics, Univ. of Ancona, Via Brecce Bianche, 60131 Ancona, Italy
(*) Dept. INFOCOM, University of Roma "La Sapienza", Via Eudossiana 18, Roma, Italy

The phase correlation algorithm provides high performance in presence of rigid plane translation, but has high computational cost since it always requires complex arithmetic. The use of direct correlation between the edges of the images greatly reduces this computational cost since it is possible to use the residual Fermat Number Transform (FNT); neverthless its performance is low. This paper presents an improved registration algorithm which uses the well-known Burg filter to whiten the images before the contour extraction in order to approach the performance of the phase correlation method but applying the more computational efficient FNT algorithm.

1. INTRODUCTION

An important problem in many applications such as remote sensing, medical instrumentation, etc., is the accurate registration of relatively long sequences of images whose observation is made difficult by more or less unpredictable shifts of the sensitive camera relative to the observed object.
A well known method to solve this problem is the phase correlation algorithm [1] which provides high performance in presence of rigid plane translation, but requires high computational cost since it uses complex arithmetic also if boundary-only binary images are registered. However, the application of the method results in a sharp peak at the point of registration which allows to easily find the correct displacement.
The use of direct correlation between the edges of the images greatly reduces this computational cost since it is possible to use the Fermat Number Transform (FNT) [2] which requires only integer residual arithmetic. Nevertheless its performance is low.
The better performance of the algorithm [1] is related to the fact that it uses "phase-only" [3] images which emphasize contour and edge information, while ignore low frequency disturbances. In fact, it was shown [4] that the use of boundary-only binary images do not impair the overall performance of the method.

This work was supported in part by the Consiglio Nazionale delle Ricerche of Italy under the project "Materials and Devices for Solid State Electronics" and in part by the Ministero Pubblica Istruzione of Italy.

Since the phase images can be considered "whitened" versions of the original images, it is important to find a way to "whiten" the original images before the contour extraction, in order to approach the performance of the phase correlation method but applying the more computational efficient FNT algorithm.
The well-known Burg inverse filter [5] can be used to obtain "whitened" versions of the images. In fact a 2D registration problem can be transformed into a 1D one by rearranging the pixels in a vector following the order of the raster scan [6]. Moreover the filter coefficients could be computed only for the reference image without significantly reducing the performance.
The proposed method, joining some advantages of the algorithm [1] and the algorithm [2], has much less computational and memory requirements than the first, and performs better than the second.

2. THE PROPOSED ALGORITHM

Let I_0 be the reference image and $\{I_k, k=1,2,...\}$ be a sequence of randomly displaced version of the reference image. The proposed algorithm consists of the following steps:

1) Computation of the AR model of the reference image I_0 by the Burg algorithm. This algorithm is applied the monodimensional sequence obtained by successively adjoining the rows of the image I_0. The order of the model necessary in this case is usually low due to the characteristics of real-world images.

2) Use of the FIR inverse filter to whiten the reference I_0 and displaced images I_k. This filter is obtained by considering the parameters of the previously computed AR model as the samples of its finite impulse response. The filtering operation is performed on the monodimensional version of the images as in step (1).

3) Detection of boundary maps from the whitened images. Most of the current edge detectors can be used in this step to obtain purely binary images.

4) Computation, using the FNT, of the cross-correlation functions between the binary reference image I_0 and the binary displaced images I_k obtained in the previous step. Also this operation is performed on the monodimensional versions of the images.

5) Detection of the registration peaks in the not normalized cross-correlation surfaces. The coordinates of every peak provide the displacement of each image I_k with respect to the reference one I_0.

The additional computational cost required by the steps (1) and (2) with reference to the algorithm proposed in [2], is not high with respect to the obtained performance. In fact the Burg algorithm in step (1) is applied only to the reference image, and could be compute off-line. Moreover the filtering operation in step (2) requires only a small amount of hardware due to the low order of the used AR model. On the other hand the cross-correlation surface in step (5) presents a registration peak sharper then that obtained in [2]. In fact the detection of the boundary maps in step (3) provides very noisy binary images, but with the noise still highly correlated.
The advantages of using the FNT at step (4) are the absence of rounding errors, a small word length and a reduced number of operations.In fact, in order to represent all the numbers in the ring module a Fermat number ($2^{2^t}+1$; t=0,1,...), 2^t+1 bits are actually required. By choosing the diminished-1 representation of numbers, it is possible to perform all calculations using standard binary adders and multipliers $b=2^t$ bits wide, treating the (2^t+1)-*th* bit just as a flag to be tested. On the other hand, in many practical cases an equivalent FFT, computed using $b/2$ bits for the real and $b/2$ bits for the imaginary part, cannot give satisfactory results in term of rounding errors, indeed to obtain comparable results an FFT will require about $2b$ bits.
However, the FNT has a limited choice of possible transform length N and, moreover, the result of the convolution computed by the FNT is only congruent, that is 'equal module F', with the result which would be computed by ordinary arithmetics, and it is necessary to keep the result below F in order to avoid ambiguities. This limits the dynamic range of the input data, more sensibly for large N. The available transform sizes for t=3 (wordlength=8 bits) and t=4 (wordlength=16 bits) are of great interest for image registration purposes. In particular a wordlenght of 16 bits is appropriate for the correlation of binary images up to 64k pixels (256x256), while wordlenght of 8 bits could be used for matching binary icones up to 16x16 pixels. There is no concern with the dynamic range, since the maximum N can always be exploited with binary data.

Table 1. Some possible combinations of word length (WL), transform basis α, transform length N and permitted data range for the Fermat numbers (F) with t=3 and t=4.

WL	F	α	N	range
8+(1)	257	9	128	0÷1
	257	3	256	0÷1
16+(1)	65537	26	8192	0÷2
	65537	81	16384	0÷2
	65537	9	32768	0÷1
	65537	3	65536	0÷1

3. EXPERIMENTAL RESULTS

A comparison of the proposed method with respect to the Kuglin algorithm [1] and the algorithm reported in [2] was carried out in order to prove the effectiveness of the method. Several real-world images were processed; in this section only a single example will be reported.
The example image, called 'Angela', was digitized by a TV camera in a 256x256 pixels format with 256 grey levers per pixel. From a normalized version of this image, several subimages of 128x128 pixels were cut. The first was considered as the reference image, the others, respectively displaced of 2,4,8,16,32,48, and 64 pixels along the main diagonal, were considered as displaced images to be registered.
Fig. 1 shows the reference image (a) and the same image displaced of 32 pixels (b). Fig. 2 shows the phase image (a) of Fig. 1b and the image (b) obtained by the Fig. 1b filtered through the FIR filter inverse to the AR model of order 6 computed on the image of Fig. 1a. The algorithm reported in [7] was used to generate the contour maps in these experiments.
Fig. 3 shows the performances obtained by the three methods applied to the above example images in terms of Signal-to-Noise ratio (S/N).

The S/N was defined as follows:

$$S/N = 20\,Log_{10}\frac{PeakValue\text{-}NoiseMean}{\sqrt{NoiseVariance}}$$

where PeakValue is the value of the registration peak in the cross-correlation surface, NoiseMean is the mean of the noise outside the peak area, and NoiseVariance is the variance of the same noise. The peak area was considered to be a circle of radius 10 pixels centered at the peak coordinates.

The curves of Fig.3 show that the performances of the proposed method are superior to the algorithm in [2] while are lower than those relative to the Kuglin method. In this last method complex arithmetics are involved, which requires much more costly hardware.

To evaluate the robustness of the proposed algorithm when noise is present, the same experiments were carried out adding uniform noise to the images. The noise was added independently to each image to be registered as a percentage of the white level (255). The resulting S/N's, relative to the 'Angela' image, are reported in Fig. 4 (5% of noise) and in Fig. 5 (10% of noise); they confirm the performances obtained in absence of noise.

REFERENCES

[1] C.D. KUGLIN, D.C. HINES, "The phase correlation image alignment method", Proc. of IEEE Int. Conf. Cybern. and Soc., 1975, pp. 163-165.

[2] C. MORANDI, F. PIAZZA, A. DOLCETTI, "Image registration using Fermat Transforms", Electronics Letter, Vol. 24, No 11, 1988, pp. 678-679.

[3] A.V. OPPENHEIM, S.J. LIM, "The importance of phase in signals", IEEE Proceedings, Vol. 69, No. 5, 1981, pp. 529-541.

[4] C. MORANDI, F. PIAZZA, R. CAPANCIONI, "Digital image registration by phase correlation between boundary maps", IEE Proc.s, Vol. 134, Pt. E, No 2, 1987, pp. 101-104.

[5] S.M. KAY, S.L. MARPLE, "Spectrum analysis - A modern perspective", Proc. of the IEEE, Vol. 69, No. 11, 1981, pp. 1380-1419.

[6] C. MORANDI, F. PIAZZA, R. CAPANCIONI, "Robust 1-D equivalent of phase correlation image registration algorithm", Electronics Letters, Vol. 22, No. 7, 1986, pp. 386-388.

[7] V. CANTONI, I. DE LOTTO, M. FERRETTI, "A template matching operator for edge-points detection in digital pictures"' Signal Processing, 4, 1982, pp. 349-360.

Fig. 1 (a) the reference image, (b) the same image displaced of 32 pixels along the main diagonal.

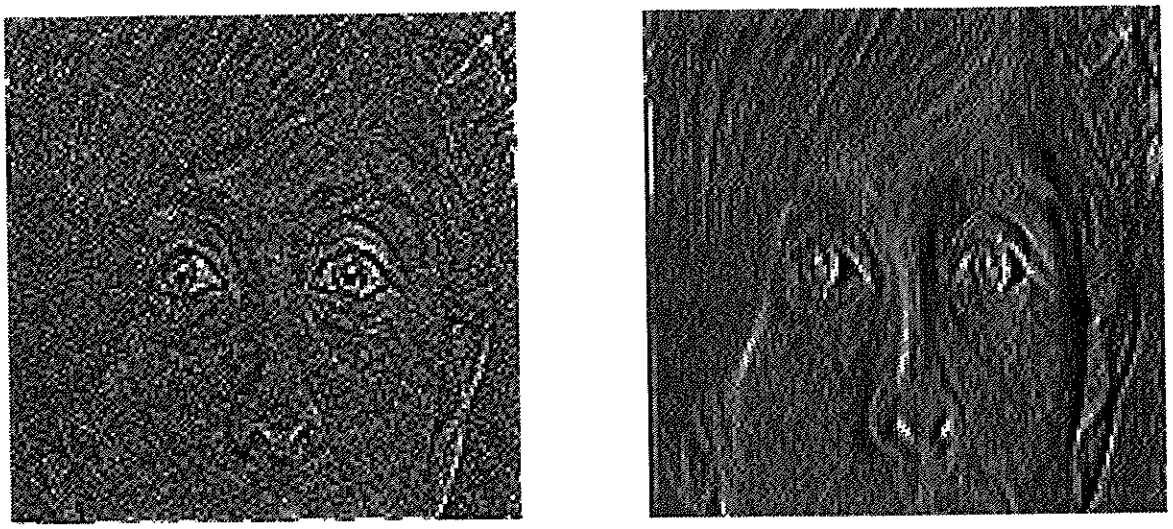


Fig. 2 (a) the phase image corresponding to Fig. 1b, (b) the image obtained by the Fig. 1b filtered through the FIR filter inverse to the six order AR model computed on the image of Fig. 1a.

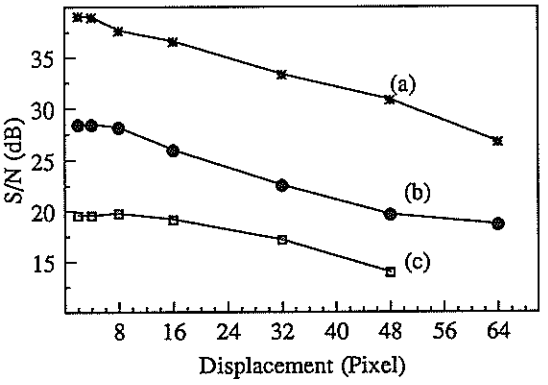


Fig. 3 Resulting S/N in the case of 'Angela' image sequence registration: (a) the Kuglin algorithm [1], (b) the proposed method, (c) the algorithm reported in [2].

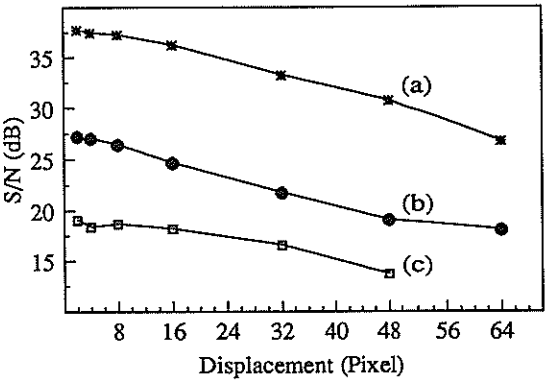


Fig. 4 As in Fig. 3 but with 5% of uniform noise added to the images.

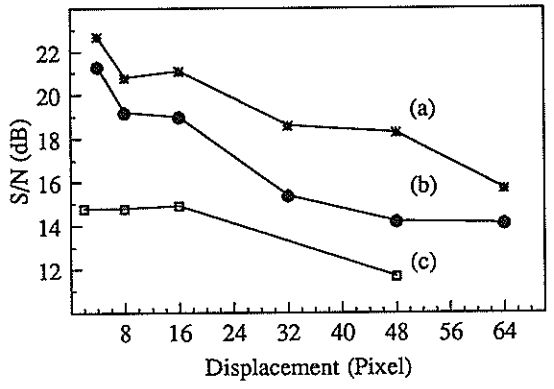


Fig. 5 As in Fig. 3 but with 10% of uniform noise added to the images.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

SUM OF ABSOLUTE DIFFERENCE VALUES SMOOTHING: COMPARISON TO NEW ALGORITHMS AND APPLICATION TO REMOTE SENSING

A. de ALBUQUERQUE ARAÚJO

Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Cx. Postal 702, 30161 Belo Horizonte - MG, Brasil

M. A. de BARROS and J. E. R. de QUEIROZ

Lab. Associado de Sensoriamento Remoto, Universidade Federal da Paraíba, Cx. Postal 10105, 58100 Campina Grande - PB, Brasil

In this work, an edge-preserving smoothing method presented in [1] has its performance compared to recently published smoothing methods. The application of these smoothing methods as preprocessing schemes in multispectral imagery classification tasks is also reported and illustrated.

1. INTRODUCTION

The purpose of image enhancement (contrast stretching, smoothing, sharpening, and highlighting of specific features) is to improve picture quality. Spatial-domain noise-smoothing techniques have been widely used in image enhancement tasks. Their application is expected to facilitate further processing and to increase analysis results.

Because of their efficacy and relative easy of implementation, filters performing in the spatial domain have been continuously studied. These algorithms perform some type of local operation, such as averaging through a mask to the picture. The gray level of the pixel at the center of this mask is replaced by the gray level average of the pixels inside the mask.

This work deals with spatial filters and is an extension of the paper presented in [1]. There, the sum of absolute difference values smoothing algorithm had its performance compared to seven stablished algorithms. This comparison is now repeated with the introduction of some recently published adaptive working filters.

The algorithms tested and compared in this work are:

sum of abs. dif. values smooth. 5x5 - SADVS [1],
unweighted neighbor averaging 3x3 [2],
median filtering 3x3 [2],
k-nearest neighbor averag. (k=6) 3x3 - KNN6 [3],
most homog. neighborhood smooth. 5x5 - MHNS [4],
slope facet model smoothing 5x5 [5],
sigma filter 5x5 [6],
adapt. order statist. filt. (j=1) 5x5 - AOSF [7],
adapt. sigma filt. ($k=3, \alpha=2, w_i=1/3$) 3x3 - ASF [8],
ad. weigh. median filt. ($c=5, \bar{w}=20$) 5x5 - AWMF [9].

2. EVALUATION

A synthetical image of size 128x128 pixels, consisting of a circle of radius 50 is used to evaluate the performance of the algorithms. The digital image processing system (SITIM - Sistema de Tratamento de IMagens - ENGESPAÇO, Brazil) resolution allows 256 gray levels. Gaussian noise with null mean and standard deviation 20.0 is added to the original image, which is then processed by the algorithms (four iterations). The methods were implemented in C and no effort was made to speed up the programs.

Two performance criteria were considered: a) noise smoothing efficiency, expressed by the reduction of noise standard deviation (SD), and b) fidelity to the original noise free image, expressed by the mean square error (MSE).

Table 1 shows the results obtained for the noise standard deviations of the filtered images computed from a flat area of size 20x20. It can be seen that SIGMA, AVERAGE, SADVS, and MEDIAN are the most efficient ones in terms of noise smoothing. MHNS and FACET converge rapidly (up iteration 2). Between the adaptive filters, AWMF worked the best.

The results obtained for the mean square error are shown in Table 2. The SIGMA filter is again the best performer, followed by SADVS, MEDIAN, and AWMF. The obtained reduction within the MSE reflects the algorithms' capacity of noise smoothing with edge retention. Because it affects the original information, generally in the form of blurring, AVERAGE presents the worst results, although it reduces well the noise.

In addition to the above performance criteria, scan line profiles of the filtered images (iteration 0) were used to visualize the algorithms' capacity of edge retention.

Table 1 CIRCLE: NOISE STANDARD DEVIATION IN A FLAT AREA

Methods	It.0	It.1	It.2	It.3	It.4
AVERAGE	7.09	5.01	3.89	3.45	3.11
MEDIAN	8.12	5.92	5.04	4.40	4.09
KNN6	8.98	6.90	5.77	5.01	4.33
MHNS	7.90	5.69	5.07	5.06	5.06
FACET	7.19	5.92	5.73	5.59	5.58
SIGMA	8.81	4.94	3.01	2.72	2.39
AOSF	9.21	7.91	7.09	6.90	6.43
ASF	8.25	6.61	5.70	5.09	4.79
AWMF	7.99	6.10	5.11	4.59	4.52
SADVS	7.89	5.60	4.69	4.11	3.81

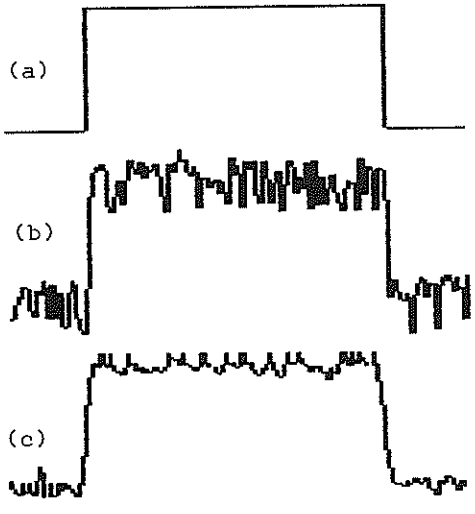
Original: SD = 0.0
With Gaussian noise: SD = 20.0

Table 2 CIRCLE: MEAN SQUARE ERROR

Methods	It.0	It.1	It. 2	It.3	It.4
AVERAGE	82.09	73.61	79.85	86.06	93.96
MEDIAN	70.21	50.06	41.11	35.36	31.38
KNN6	81.96	54.02	45.49	40.91	39.71
MHNS	71.17	51.02	47.12	46.30	46.09
FACET	59.01	44.20	41.05	40.91	40.79
SIGMA	73.41	29.06	17.12	13.99	12.98
AOSF	72.12	60.92	56.17	54.08	50.99
ASF	73.05	57.02	49.93	40.49	35.08
AWMF	70.19	52.97	40.11	36.14	34.01
SADVS	71.70	42.98	38.01	33.16	30.11

Original: MSE = 0.0
With Gaussian noise: MSE = 391.06

Figure 1 presents the profiles of scan line 40 of the original, noise corrupted and filtered images. It can be observed that all algorithms except averaging are very good in preserving edges.



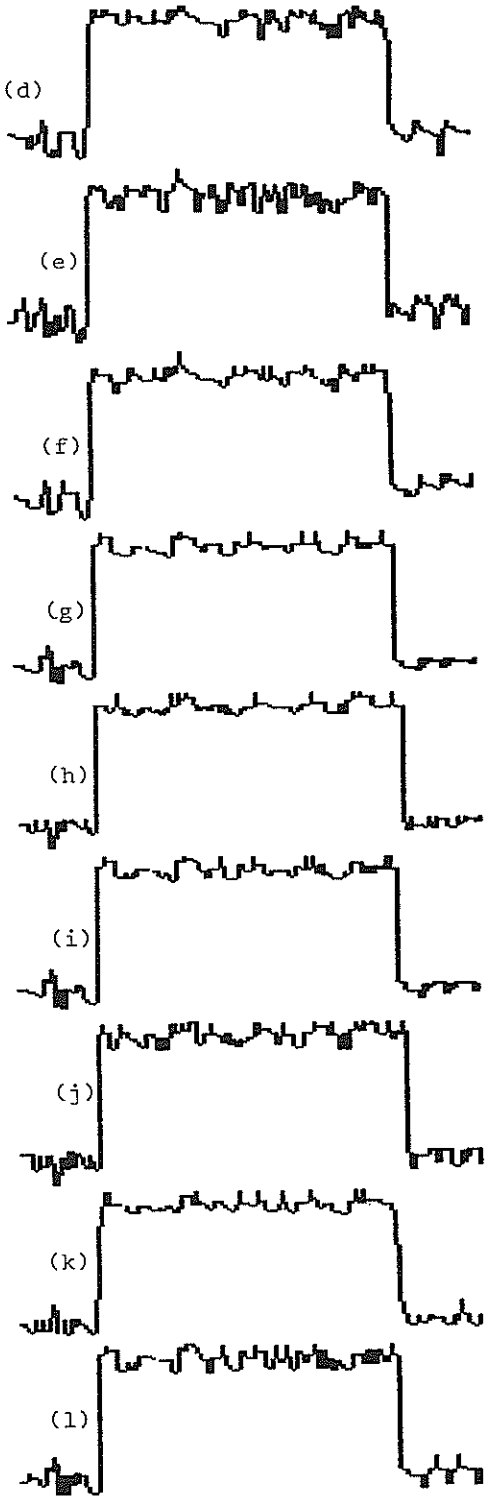


Figure 1. Profiles of scan line 40: (a) original, (b) with noise, (c)-(l) images filtered (it. 0) by AVERAGE, MEDIAN, KNN6, MHNS, FACET, SIGMA, AOSF, ASF, AWMF, and SADVS.

3. APPLICATION

Classification tasks consist basically in assigning targets to one class of a set of previously specified classes, according to an appropriate classification rule. Generally, the developed methods are based on image statistical features. They classify each pixel of a multispectral image individually, considering the information about spectral correlation within bands or other available features. They do not take into account the information about spatial correlation of the data.

In an attempt to enhance the spatial correlation within the original data, the adaptive-order statistical filtering (AOSF) and the sum of absolute difference values smoothing (SADVS) algorithms were applied to a LANDSAT/TM-5 multispectral image (WRS 223.76). The purpose of this procedure is to increase the accuracy of the classification process.

In this approach, bands 5, 4, and 3 (original bands of the image), bands i, j, and k (bands 5, 4, and 3 preprocessed by SADVS) and bands l and m (bands 5 and 4 preprocessed by AOSF) have been used.

A feature extraction algorithm based on the Jeffrey-Matusita (JM) maximum and mimimum distances' criteria is used to choose an optimal three-channel subset. The selected bands have been l, m, and k. During the feature extraction process, it was observed a tendency to select at least one preprocessed band among other bands of the selected subset.

Tables 3 and 4 present the results of the classification process for the original data and for the preprocessed data (selected bands l, m, and k), respectively. In Table 4 it can be noticed an increasing of the performance mean and a decreasing of the abstention and confusion means of the classification results.

These initial results have demonstrated that spatial domain edge-preserving smoothing techniques contribute to an increasing of the accuracy in multispectral image classification tasks.

4. CONCLUSIONS

SADVS, a selective-neighborhood-working noise smoothing algorithm, had its performance evaluated in the work presented in [1]. This comparison was repeated with the introduction of adaptive working algorithms.

In terms of the two criteria (noise smoothing efficiency and fidelity to the original image) used in this study, SADVS worked better than the

Table 3 - CLASSIFICATION MATRIX: ORIGINAL DATA

%		COMPOSITION 543				THRESHOLD 6,20			
CLASSES		nclas	1	2	3	4	5	6	7
1	water	4,8	95,2	-	-	-	-	-	-
2	cane 1	9,6	-	90,4	-	-	-	-	-
3	cane 2	9,4	-	0,5	90,1	-	-	-	-
4	corn	5,6	-	-	-	94,4	-	-	-
5	sand	4,1	-	-	-	-	90,6	5,4	-
6	soya	5,1	-	-	-	-	0,1	94,2	0,6
7	forest	10,4	-	-	-	-	-	-	89,6
Performance: 92,24		Abstention: 6,61				Confusion: 1,15			

Table 4 - CLASSIFICATION MATRIX: PREPROCESSED DATA

%		COMPOSITION lmk				THRESHOLD 6.20			
CLASSES		nclas	1	2	3	4	5	6	7
1	water	4,8	95,2	-	-	-	-	-	-
2	cane 1	8,8	-	91,2	-	-	-	-	-
3	cane 2	10,5	-	-	89,5	-	-	-	-
4	corn	5,0	-	-	-	95,0	-	-	-
5	sand	3,4	-	-	-	-	91,2	5,4	-
6	soya	4,2	-	-	-	-	0,1	95,6	0,1
7	forest	9,8	-	-	-	-	-	-	90,2
Performance: 93,21		Abstention: 5,90				Confusion: 0,89			

adaptive spatial filters. Generally, the adaptive filters require the definition of some parameters. The selection of parameters may not be a trivial task, although it can work as a fine tunning control in interactive applications.

In this study, the SIGMA filter was the best performer. It is worth to observe that in such a simulated study, with the noise conditions well known, the selection of the paramenters required by the SIGMA filter could be done without great problems.

The preprocessing of LANDSAT TM/5 images by spatial filters brouhght out a precision increasing of the classification. The application of spatial filters as preprocessing schemes in an ultrasound tissue characterization task and in edge-detection and segmentation tasks were reported in [10] and [11], respectively.

ACKNOWLEDGMENTS

The first author would like to thank the Conselho Nacional de Desenvolvimento e Pesquisa - CNPq, Brazil, and the Deutscher Akademischer Austauschdienst - DAAD, West Germany, for the financial support of this work.

REFERENCES

[1] Araújo, A. de A., Sum of Absolute Grey Level Differences: an Edge-Preserving Smoothing Approach", Electronics Letters, 1985, 21, pp 1219-1220.

[2] Pratt, W., Digital Image Processing, (Wiley & Sons, N.Y., 1978).

[3] Davis, L.S. and Rosenfeld, A., Noise Cleaning by Iterated Local Averaging, IEEE Trans., 1978, SMC-8, pp 705-710.

[4] Nagao, M. and Matsuyama, T., Edge Preserving Smoothing, Computer Graphics & Image Processing, 1979, 9, pp 394-407.

[5] Haralick, R.M. and Watson, L., A Facet Model for Image Data, ibid., 1981, 15, pp 113-129.

[6] Lee, J.S., Digital Image Smoothing and the Sigma Filter, ibid., 1983, 24, pp 255-269.

[7] Lee, Y.H. and Fam, A.T., An Edge Gradient Enhancing Adaptive Order Statistic Filter, IEEE Trans. 1987, ASSP-35, pp 680-695.

[8] Jung, S.-H. and Kim, N.-C., Adaptive Image Restoration of Sigma Filter Using Local Statistics and Human Visual Characteristics, Electronics Letters, 1988, 24, pp 201-202.

[9] Loupas, T., McDicken, W.N., and Allan, P.L., An Adaptive Weighted Median Filter for Speckle Suppression in Medical Ultrasonic Images, IEEE Trans., 1989, CAS-36, pp 129-135.

[10] Araújo, A. de A., Kubalski, W., Jensch, P., and Ameling, W., Einflüsse von Moving-Window-Verfahren auf Texturdiskriminanz-eigenschaften in Echokardiogrammen, in: Niemann, H. (ed.), Mustererkennung 1985 (Springer-Verlag, Heidelberg, 1985) pp 213-217.

[11] Araújo, A. de A., Sum of Absolute Difference Values Smoothing: Evaluation and Application, in: Young, I.T., Duin, R.P.W., Biemond, J., and Gerbrands, J.J., Signal Processing III: Theories and Applications (North-Holland, Amsterdam, 1986) pp 773-776.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ON COMPUTING THE LENGTH OF DIGITAL LINES

Takayasu ITO and Hitoshi INO

Department of Information Engineering
School of Engineering,Tohoku University
Sendai,JAPAN 980

Extraction and analysis of digital lines have been the most basic and important in theories and applications of digital image processing and understanding.The length of digital lines is one of the most basic properties of digital lines.However this basic property of image processing and digital computational geometry has not been studied well.This paper is the first systematic and theoretical attempts to discuss this problem of computing the length of digital lines in the 2-dimensional and 3-dimensinal images. After discussing the nature of the problem an algorithm to compute the average length of digital lines will be given. This discussion will be extended to compute the area and volume surrounded by digital lines and surfaces,respectively.

1.INTRODUCTION

Extraction and analysis of digital lines have been the most basic and important in theories and applications of digital image processing and understanding.Progresses in these areas have explored a new demand of finer quantitative analysis of digital images.As is seen in Huffman-Waltz labelling ,fine qualitative and structural analyses of digital lines have explored better analysis and understanding of scenes and line drawings. Once detailed qualitative understandings of digital lines are obtained,it is natural to ask better quantitative understandings of digital lines, such as the length of digital lines and the area surrounded by digital lines.But only a little attention has been paid on these aspects of analysis and understanding of line drawings.The length of digital lines is one of the most basic quantities of digital lines, and it appears in analyses of maps, remote sensing data,biomedical image data,etc. However this basic problem of computing the length of digital lines has not been studied well in image processing and computational geometry.

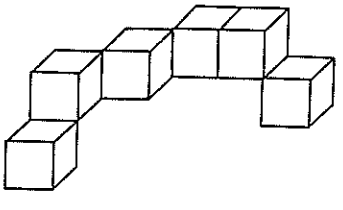
In this paper we attempts to discuss this problem of computing the length of digital lines in the 2 dimensional and 3 dimensional cases.After discussing the nature of the problem and explaining why the existing methods are unsatisfactory we will give an algorithm to compute the average length of digital lines.This discussion will be extended to compute the area and volume surrounded by digital lines and surfaces,respectively.

2.Problem of Computing The Length of Digital Lines

The length of a curve "C" can be given by the following line integral:

$$L = \int_C ds$$

Assume that we divide a line into the connected sequence of meshes in the 3 dimensional spaces as in the figure below.



Let
$(x_0, y_0, z_0), (x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$
be the set of points that a curve crosses the pixels or the boxels. {In case of the two dimensional images the "z"-components may be neglected.This comments may be applied whenever the formulae of the three dimensional case are given.}
Then "L" will be approximated as follows:

$$L \sim L_0 = \sum_{i=1}^{n} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}$$

However in case of digitized images we do not know the curve precisely.The only information that we know is a sequence (or,a set) of boxels,which is called as the digitized curve.After applying some thinning algorithms we obtain a digitized line for this digitized curve. Then the problem of computing the length of the digitized curve will be reduced to the problem of computing the length of the digitized line.
If we can estimate the average length of each pixel or boxel "l_i" we can estimate the length of the digital line as follows:

$$L \sim l_1 + l_2 + \ldots + l_n$$

Thus the estimation of the average length of each boxel is the central issue of computing the length of digital lines.

2.1 Approaches to Compute the Length of Digital Lines

There are several approaches to compute the length of digital lines. The existing and well-known approaches are
1) Method of curve fitting by polynomials
An optimal polynomial f(x)

$$f(x) = c_0 + c_1 x + c_2 x^2 + \ldots + c_m x^m$$

for a two-dimensional digital line will be found by the minimum mean square error criteria.In the three dimensional case we denote it in terms of two polynomials f(x) and g(x) in the following form:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ f(x) \\ g(x) \end{pmatrix}$$

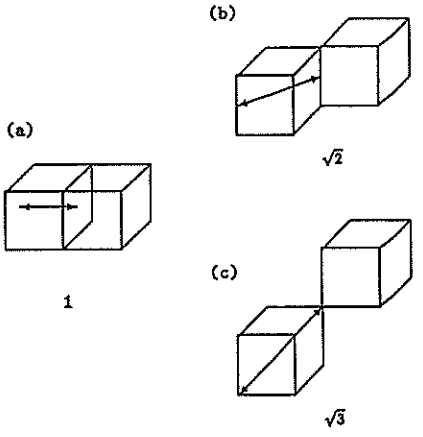
Then the problem of computing the length of digital lines is the problem to compute the following integral in a given interval.

$$\int \sqrt{1 + (f'(x))^2 + (g'(x))^2}\, dx$$

This approach has the following problems:
(a)The order of computational cost is $O(n^3)$ even if we compute the above integral in advance for the given polynomials, where "n" is the number of pixels.
(b)How to determine the degree of polynomials and how to carry an polynomial approximation are not easy because of the nature of digitization.
2) Method of giving certain constants in terms of the neighbourhood connectivity
Condider a digital line with the 8-neighbourhood connectivity.
In case of the 2 dimensional image we can consider the following two cases:

1 $\qquad$ $\sqrt{2}$

In case of the three dimensional image we can consider the following three cases:



1 $\qquad$ $\sqrt{2}$ $\qquad$ $\sqrt{3}$

Then the problem of computing the length of a digital line will become
to count the pixels and boxels for each case;
to multiply the corresponding weights;
then we sum up the results to obtain the total length of a digital line.
This method is good in its computational cost,but it is not good in its accuracy. That is,the computational cost is in O(n)

in terms of the number of pixels,but the error induced by this method will be too big in actual applications.

In order to resolve the problems of these methods we propose an algorithm to compute the length of digital lines using the idea of estimating the length of pixels by neighbourhood dependency and statistical estimation.It will be shown that the computational cost of this method is in O(n),and the statistical estimation approach may be more comfortable from the nature of digitized curves and digital lines.

2.2 Estimation of Length of Digital Lines

The information that we have on digital lines is the connected sequence of pixels or boxels,so that their real nature is not known; it may be a result of digitization of straight lines,smooth lines,high-frequency curves or fractals. From the standpoint of digital image processing the pixel information may be considered to be the most basic in terms of resolution of image devices. It may be natural to assume that the digitization of images is sufficiently fine,unless any other information is supplied.Thus our fundamental assumption is:

"Each pixel may be approximated by a line segment."

Under this assumption the length of a digtal line with n connected pixels $\{x_1, x_2, \ldots, x_n\}$ can be given by

$$L \sim l_1+l_2+\ldots+l_n$$

where l_i is an estimate of length(x_i).

There are many possibilities that a line segment pass through and cross a pixel. In analysis of a complicated DNA electron micrograph the directions of DNA strings may be considered randomly distributed.In this case a line segment passes through a pixel randomly;that is,we can estimate its length statistically.

In analysis of a road map some parts of a road are more straight and other parts of it are more winding;in this case it will be better to estimate the length of each line segment on the basis of dependency of of each pixel with its neighbourhood.

In case of two dimensional digitized images we have two kinds of the neighbourhood definitions:

- 4-neighbourhood:a pixel is connected with another,sharing its line edge;
- 8-neighbourhood:a pixel is connected with another,sharing its edge and corner point.

We can extend this neighbourhood relation into the three dimensional case.For a boxel $x=(i,j,k)$ we can define three kinds of its neighbourhood,depending on "plane-","line-" and "point-" connectivity with other boxels. T_{plane},T_{line} and T_{point} denote the sets of the plane-,line- and point-connected boxels, respectively.Let N_{plane},N_{line} and N_{point} be the plane-, line- and point- connected neighbours of x.They are expressed as follows:

$$\begin{aligned}
T_{plane}(x) &= \{(p,q,r)\,;|p-i|+|q-j|+|r-k| = 1\}\\
T_{line}(x) &= \{(p,q,r)\,;|p-i|+|q-j|+|r-k| = 2,\\
&\qquad |p-i|\leq 1,|q-j|\leq 1,|r-k|\leq 1\}\\
T_{point}(x) &= \{(p,q,r)\,;|p-i|=|q-j|=|r-k| = 1\}
\end{aligned}$$

$$\begin{aligned}
N_{plane}(x) &= T_{plane}(x)\\
N_{line}(x) &= T_{plane}(x)\cup T_{line}(x)\\
N_{point}(x) &= T_{plane}(x)\cup T_{line}(x)\cup T_{point}(x)
\end{aligned}$$

2.2.1 Connectivity and Crossing Segments

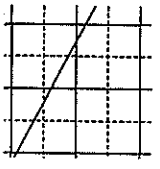
There is the important relation between the neighbourhood connectivity and the segments to cross a pixel. It is easy to have an intuitive understanding in case of two dimensional case.

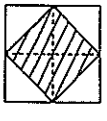
<4 Neighbourhood Connectivity and Segments>

In case of 4 neighbourhood connectivity a pixel is connected with its neighbours by its edges so that any segment to cross an edge must be considered as a segment to cross the pixel.

<8 Neighbourhood Connectivity and Segment>

In case of 8 neighbourhood connectivity a pixel is connected with its 8 neighbours by its edges or its corner points.That is, the 8 neighbourhood connectivity is weaker than the 4 neighbourhood connectivity. In this case we allow only a segment which crosses the shadowed area of the pixel as below.





In case of the three dimensional case we must consider the cases of
the plane connectivity,
the line connectivity
and the point connectivity. The arguments similar to the two dimensional case are left for the readers as an exercise.

2.2.2 Average Length of Line Segments to Cross Pixels/Boxels with No Dependency

Since this case is a basis of our method we explain in some details.

<Average Length of Plane-Connected Boxels>

Consider a boxel with the edge-length T. Let a line p cross this boxel with the angels θ and ϕ. This line can be obtained by rotating $p_0=(0,0,r)$ with the angels θ and ϕ. This rotation can be expressed by

$$R = \begin{pmatrix} \cos\theta\cos\phi & -\sin\phi & \sin\theta\cos\phi \\ \cos\theta\sin\phi & \cos\phi & \sin\theta\sin\phi \\ -\sin\theta & 0 & \cos\theta \end{pmatrix}$$

For the ease of computing the length of lines we transform the co-ordinate from xyz to x'y'z' so that p becomes parallel to z'. This transformation can be expressed by the inverse matrix R^{-1}.

Let $D(x',y',\theta,\phi)$ be the boundary surface (in x'y'z') which will be crossed by the line segments. Taking into account of the geometrical explanations of the following figure, we can give a formula for the average length "l" of lines to cross the pixels as follows:

$$l_3 = \frac{\int_0^{\pi/2}\int_0^{\pi/2}\sin\theta d\theta d\phi \int_{D(x',y',\theta,\phi)} dx'dy'L(x',y',\theta,\phi)}{\int_0^{\pi/2}\int_0^{\pi/2}\sin\theta d\theta d\phi \int_{D(x',y',\theta,\phi)} dx'dy'}$$

If we fix $\theta=\pi/2$ and y'=0, then we obtain the following formula which corresponds to the two dimensional case:

$$l_2 = \frac{\int_0^{\pi/2} d\theta \int_{D(x',\theta)} dx'L(x',\theta)}{\int_0^{\pi/2} d\theta \int_{D(x',\theta)} dx'}$$

<Average Length of Line-Connected Case and Point-Connected Case>

The formulae can be obtained using
$D_{line}(x',y',\theta,\phi)$
and $D_{point}(x',y',\theta,\phi)$
instead of $D(x',y',\theta,\phi)$.

2.2.3 Average Length of Line Segments with Neighbourhood Dependency Considerations

The line segments to be computed with neibourhood dependency of boxels should satisfy the condition that they cross the boxels. Consider a neibourhood dependency of n boxels with the crossing regions $D_1,\cdots,D_n$; then we have:

$$l_{c3} = \frac{\int_0^{\pi/2}\int_0^{\pi/2}\sin\theta d\theta d\phi \int_{D(x',y',\theta,\phi)} dx'dy'L(x',y',\theta,\phi)}{\int_0^{\pi/2}\int_0^{\pi/2}\sin\theta d\theta d\phi \int_{D(x',y',\theta,\phi)} dx'dy'}$$

$$D = D_1 \cap \cdots \cap D_n$$

In case of the line- and point- connectivities thre are two cases:

1)<conservative estimate> We consider only the lines in the connected regions

2)<non-consrvative estimate> We consider all the line segments inside the parallelooidal cones formed from the neighbouring boxels.

2.2.4 Table to give Average Length of Line Segments to Cross Boxels

The table below gives the actual numerical values of the average length of line segments to cross boxels, where "*" indicates the non-conservative estimate mentioned above. Also, "1-Relation" and "3-Relation" mean the cases with "No Neighbourhood Dependency" and "3 Neighbourhood Dependency", respectively.

The tables for "2-Relation", "PLANE-LINE", "PLANE-POINT", "LINE-LINE", "LINE-POINT" and "POINT-POINT" are not included here because of the limitation of space.

[Computational Cost]

The computational cost of computing the length of digital lines using the tables is obviously O(n) in terms of the number of pixels.

1 RELATION

		plane	line	point
(000)		0.67	0.83	0.93
		—	0.89*	1.12*

3 RELATION

PLANE-PLANE		plane	line	point
(011)		0.49	0.64	0.86
(022)		1.22	1.16	1.09

REFERENCES

T.Ito & K.Sato: Computer analysis of electron micrographs of DNA, in "Digital Processing of Biomedical Images", ed.by K.Preston and M. Onoe, 89-100(1976)

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

STATISTICAL ANALYSIS OF RESOLUTION IN IMAGES

José MARTINEZ-AROZA*, José J. QUESADA-MOLINA* and Ramón ROMAN-ROLDAN**

*Dept. of Matemática Aplicada **Dept. of Física Aplicada
Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain

The goal of this paper is to develop a theoretical background for analyzing classes of images by using the entropy-resolution diagrams. We study some properties of these diagrams, and, in particular we consider the diagrams obtained for Gibbs distributions. The Principle of Maximum Entropy plays an important role and it is considered with different restrictions. Some open problems and questions are presented.

1.INTRODUCTION

There is an increasing interest on image analysis, and the research on this subject has been growing along the last years. The various applications of image analysis go from machine intelligence or robotics to many other scientific areas.

This paper is devoted to develop a theoretical background which turns out to be useful for the analysis of either classes of images or particular images. Our approachis mainly based on the entropy-resolution diagrams associated to classes of images. Some properties for these diagrams were already conjectured by the authors [6]. Those and other properties are proved in section 2. The main idea is to use these diagrams for clasifying and analyzing images. Several interesting open problems related with these entropy-resolution diagrams are presented in section 3 of this paper. In any case, we should point out that various applications, such as restoration or superresolution, can be derived from the study considered in the present work.

First of all, we consider the neccesary notation. An *image at resolution level* 0 is an array of pixels, each one of them being black or white. Therefore, at the minimum size of observation, we consider images to be binary. An m-region is any rectangular cluster of pixels containing a determined number of them and having a fixed shape. The m-regions are obtained after partitioning the image. The total number of pixels in an m-region will be denoted by R_m (2^m) and the *grey level* of an m-region will mean the number of black pixels in this particular m-region.

We will suppose that for any m considered in the study, the image can be partitioned in an integer number of m-regions, and also that any m-region contains an exact number of m'-regions, for every m'<m. N_m will stand for the number of m-regions in the image and $N_{m,k}$ the number of those m-regions with grey level equals k. Therefore $N_m = \sum_{k=0}^{R_m} N_{m,k}$.

We can easily associate a probability distribution $\{ P_{m,k} ; 0 \le k \le R_m \}$ to any image given at resolution level m. Here, $P_{m,k}$ is the probability of grey level equals k for the next m-region to be observed. It has been explained in [5] that Shannon's entropy for $\{ P_{m,k} ; 0 \le k \le R_m \}$ measures the uncertainty about the next m-region to appear [2]. In any case the interesting problem is about the uncertainty of the next image to appear. Then, we are led to consider N_m times the 'entropy of the histogram', i.e.,

$$N_m \cdot H_m = -N_m \cdot \sum_{k=0}^{R_m} P_{m,k} \log P_{m,k} .$$

And then H_m/R_m measures the uncertainty of the whole image, redardless of its size, i.e., it is the entropy *per pixel*. Now, for any image, we can compute the number of m-regions in the image for each grey level, and then, consider the frequencies $f_{m,k}=N_{m,k}/N_m$ as the probabilities $P_{m,k}$. When this is done for every m, we can draw a graph with the points $(m , H_m/R_m)$, i.e., the *entropy per pixel* vs. *resolution* diagrams.

In our study we make constraints in order to avoid several problems, such as overlapping of regions or bounding effects.

Let us consider an image, which is given at resolution level 0, and we know either $\{ P_{0,0} , P_{0,1} \}$ or H_0 . What is the class of images (i.e. the probability distribution $\{ P_{m,k} ; 0 \le k \le R_m \}$) at m resolution level, either being compatible with $\{ P_{0,0} , P_{0,1} \}$ or having entropy H_0 , and such that it has maximum entropy at resolution level m?. The answer is the well-known Gibbs distribution at level m (see Guiasu & Shenitzer [1]), which is given by

$$P_{m,k} = e^{-k\beta}/\Phi , \quad 0 \le k \le R_m , \quad \Phi = \sum_{k=0}^{R_m} e^{-k\beta}$$

and β is the unique solution of the equation

$$\sum_{k=0}^{R_m} [k - P_{0,1}R_m] \cdot e^{-\beta[k-P_{0,1}R_m]} = 0. \qquad (1)$$

Moreover, the above posed question is the statement of the Principle of Maximum Entropy with the constraint given by $R_m \cdot P_{0,1} = \sum_{k=0}^{R_m} k \cdot P_{m,k}$.

2. PROPERTIES OF THE ENTROPY-RESOLUTION DIAGRAMS

Before studying some interesting properties of the entropy-resolution diagrams associated to either classes of images or particular images, we recall some results for the Gibbs distributions.

Theorem 1: Let $\beta_m(P_{0,1})$ be the function assigning to each $P_{0,1}$ $(0<P_{0,1}<1)$, for a given m, the unique solution β of equation (1). Then β_m is a continuous and strictly decreasing function from $(0,1)$ into $\mathbb{R}$ such that

i) $\beta_m(\frac{1}{2}) = 0$

ii) $\lim_{P_{0,1}\to 0} \beta_m(P_{0,1}) = +\infty$

ii) $\lim_{P_{0,1}\to 1} \beta_m(P_{0,1}) = -\infty$

We will stay that the probability distribution $\{P_i\}_{i=1}^n$ is majorized by the probability distribution $\{Q_i\}_{i=1}^n$ if $\sum_{i=1}^{j} P_{[i]} \le \sum_{i=1}^{j} Q_{[i]}$, j=1,2,...,n-1, where $P_{[i]}$ (respectively $Q_{[i]}$) represents the i^{th} smallest probability in the probability distribution $\{P_i\}_{i=1}^n$ (respectively $\{Q_i\}_{i=1}^n$), i.e., $P_{[1]} \le P_{[2]} \le \ldots \le P_{[n]}$ (and analogously for $\{Q_i\}_{i=1}^n$). A complete study of theory of majorization of probability distributions can be found in Marshall & Olkin [3].

Lema 2: If two probability distributions $\{ P_{0,0} , P_{0,1} \}$ and $\{ P'_{0,0} , P'_{0,1} \}$ are such that either $P'_{0,1} > P_{0,1} > 1/2$ or $P'_{0,1} < P_{0,1} < 1/2$, then the Gibbs distribution associated with $\beta_m(P_{0,1})$ majorizes the Gibbs distributions associated with $\beta_m(P'_{0,1})$.

Proofs of Theorem 1 and Lemma 2 can be found in [7].

By using these previous results we can state and prove the following theorems:

Theorem 3: Let H_m and H'_m be the entropies of Gibbs distributions associated to the probability distributions $\{ P_{0,0} , P_{0,1} \}$ and $\{ P'_{0,0} , P'_{0,1} \}$ with entropies H_0 and H'_0 respectively. If $H'_0 < H_0$, then $H'_m < H_m$.

Proof: If $H'_0 < H_0$, then we have either $P'_{0,1} > P_{0,1} > 1/2$ or $P'_{0,1} < P_{0,1} < 1/2$, or any other case which can be reduced to one of these cases by exchanging the role of black and white pixels. Then, by Lemma 2, we have that the Gibbs distribution at level m compatible with H_0 always majorizes the Gibbs distribution at the same level compatible with H'_0 $(H'_0<H_0)$, and the theory of majorization [3] gives us the desired result. ■

Theorem 4: Among all probability distributions at level m with the same entropy H_m, Gibbs distribution is the one having minimum entropy at level 0.

Proof. Let us suppose that there exists a probability distribution $\{P^*_{m,k}\}$ such that its corresponding entropy at level 0 is H^*_0, and for Gibbs distribution $\{P_{m,k}\}$ we have entropy H_0 at level 0, and suppose that both probability distributions have the same entropy H_m and $H_0 > H^*_0$. Therefore, by theorem 3 there is a Gibbs distribution for H^*_0 such that its entropy H^*_m is strictly smaller than H_m, but this is a contradiction with the existence of $\{P^*_{m,k}\}$, and the result is proved. ■

In our study we have considered the entropy H_m that represents the uncertainty about the next m-region, then $N_m \cdot H_m$ is the uncertainty about the whole image; and if we divide by N_0, we obtain $N_m \cdot H_m / N_0 = H_m / R_m$, i.e., the uncertainty of provided information per pixel. Next, we state and proof two results for the entropy per pixel-resolution diagrams that we obtain by drawing the points $\{(m , H_m/R_m); m \ge 0\}$ for particular images and some classes of images. The following theorems give a positive answers to the conjectures that were made in [6].

Theorem 5: If $H_{m'}$ and H_m are the entropies associated with a particular image when it is observed respectively at resolution levels m' and m (m'<m), then $H_{m'}/R_{m'} > H_m/R_m$.

A proof for this theorem can be seen in [4] and [7].

Theorem 6: The sequences $\{ H_m/R_m ; m>0 \}$ obtained for Gibbs distributions (see fig.1) are strictly decreasing with respect to m.

Proof: It is always possible to find an image whose histograms at level m' and m (m'<m) can be as approximate to Gibbs distributions as desired. A pass-to-limit process on this argument provides us with the non-strict decreasing of Gibbs sequences, and the strictness follows from the strict convexity of Shannon's entropy. For more details, see [7]. ■

3. SOME OPEN PROBLEMS

Several open problems and questions show up in this context. One of them, concerning the restrictions in the Maximum Entropy Principle, is briefly presented as follows:

When a class of images is given with a certain resolution level, we can go backwards (more resolution, smaller m), and determine, via some conditional probability distributions, the class of images given at resolution level m' (m'<m) that are compatible with the original class. Now, if we start with the class of images given at 0 resolution level and we apply the Principle of Maximum Entropy step by step, that is, considering as restrictions in each step the compatibility with the entropy at the previous step, then we obtain the same entropy-resolution diagram as we got backwards by using the conditional probabilities. But these do not match with the entropy-resolution diagram for Gibbs distributions, due to a non-transitivity property of the M.E.P.

REFERENCES

[1] GUIASU, S., SHENITZER, A. "The Principle of Maximum Entropy", **The Mathematical Intelligencer** 7 (1985), 1, 42-48.

[2] GULL, S.F., SKILLING, J. *The entropy of an image*, **Maximum-entropy and bayesian methods in inverse problems**, C. Ray Smith and W.T. Grandy,Jr. (Eds.), D.Reidel Pub.Co., 1985.

[3] MARSHALL,A.W., OLKIN,I. **Inequatilies: Theory of Majorization and Applications**, Academic Press, New York 1979.

[4] MARTINEZ-AROZA,J., QUESADA-MOLINA,J.J., ROMAN-ROLDAN,R. *A characterization of images through entropy-resolution diagrams*, This volume.

[5] QUESADA-MOLINA,J.J., ROMAN-ROLDAN,R. *On Some Entropies For Images*, **Computing and Information**, R.Janicki & W.W.Koczkodaj (Eds.), North Holland, New York 1989, 263-265.

[6] QUESADA-MOLINA,J.J., ROMAN-ROLDAN,R. *Resolution-Dependent Entropy For Images*, **Proc. ICCI'89**, Canadian Scholars' Press, Toronto, 1989, 200-201.

[7] ROMAN-ROLDAN,R., MARTINEZ-AROZA,J., QUESADA-MOLINA, J.J. *A new Approach for Image Analysis*, to be published in Signal Processing.

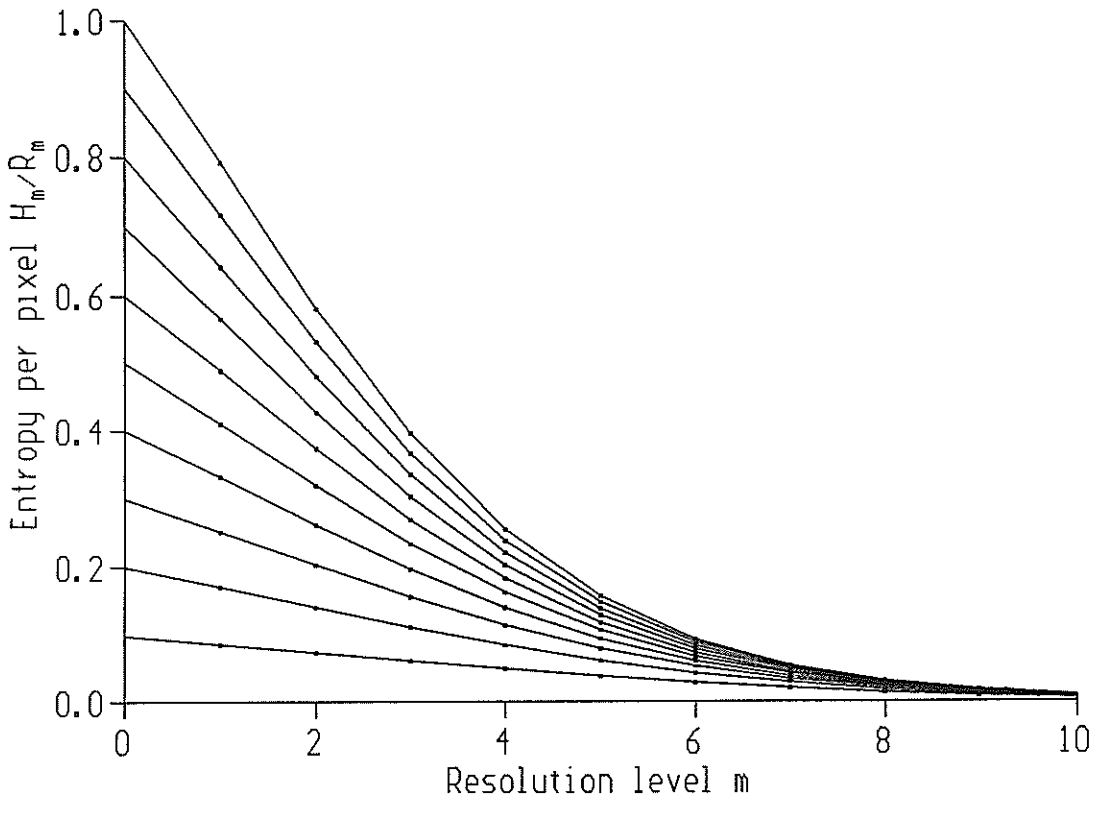


FIGURE 1

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A CHARACTERIZATION OF IMAGES THROUGH ENTROPY-RESOLUTION DIAGRAMS

José MARTINEZ-AROZA*, José J. QUESADA-MOLINA* and Ramón ROMAN-ROLDAN**

*Dept. of Matemática Aplicada **Dept. of Física Aplicada
Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain

In this paper we study the relationships between the entropy of the grey-level histogram and the resolution in images. As a consequence, the entropy-resolution diagrams are built, including the Gibbs net as a reference. Some theorems and other results concerning to these concepts are also presented. Finally, it is shown how to characterize some features of images through these diagrams.

1. INTRODUCTION

Entropic methods have been applied for image processing with various results. For instance, thresholding segmentation has been carried out [5]; and, maximum entropy methods are currently applied to restoration in images [10]. Different entropies have been used [2], even though digital image processing does not need to take into account the physical process of generation of the image.

Therefore, we work on the entropy of the grey-scale histogram, for it is a measure of both the uncertainty about the next image to appear from the proper source, and also the mean information provided by it [7]. We mean that the source is any abstract device which outcomes an image, according to a certain probability distribution defined on the grey scale of their regions, but regardless of its physical support.

In addition to the entropy, we mainly deal with the *resolution*, for both of these concepts refer to the information associated to images, and are related to each other. In fact, we investigate about the nature of these relationships [8] [11], as they reveal the interdependence.

We define the resolution as the size of the regions being observed, the finest one corresponding to black or white pixels, it is to say, we deal with binary images.

It must be noted that, when dealing with the entropy of the histogram, any geometric information in the image is lost. However, information and resolution are general enough to be able to apply when dealing with some signals other than images, such as crystals.

2. ENTROPY-RESOLUTION DIAGRAMS

For a resolution level m, we define: R_m: size of any m-region (observing cluster); $N_{m,k}$: # of m-regions with k black pixels (grey level equals k); $N_m = \sum_{k=0}^{R_m} N_{m,k}$; $P_{m,k} = N_{m,k}/N_m$. Then Shannon's entropy $H_m = -\sum_{k=0}^{R_m} P_{m,k} \cdot \log P_{m,k}$ gives a measure of the amount of information provided by the image at resolution level m, and H_m/R_m gives the entropy per pixel. If we plot the points $(m, H_m/R_m)$ in a coordinate system, the we get a entropy-resolution diagram (see fig.1) whose study constitutes the goal of this paper.

The entropy-resolution diagram of a particular image can be considered as a kind of spectrum of the image, containing certain characteristics of it. For example, $H_0=0$ means that the image is entirely black or white, while $H_0=1$ means that there are as much black pixels as white ones. In general, a high entropy at level m means that the image offers many different grey levels at this resolution level, and a lower entropy means a poor grey-scale being observed.

Analysis of images through their entropy-resolution diagrams can be useful in order to estimate some characteristics of them, but since we neglect any geometric consideration about images, several other characteristics may be dropped. For this reason, however, this study can be of application in any other kind of binary signal or system.

The main property of the entropy-resolution diagrams is their strict decreasing with respect to the resolution level m:

Theorem 1: *For any particular image, we have* $H_{m'}/R_{m'} > H_m/R_m$ *when* $m'<m$.

Proof: If $m'=m-1$ then $R_{m'}=R_m/2$ and each m-region is made of two m'-regions; let P_{m,k_i,k_j} be the probability of grey levels k_i and k_j for the respective m'-regions belonging to a m-region randomly chosen from the image, and let H_m^* the entropy for this probability distribution. From Theory of Information we have $H_m^* \le 2H_{m'}$ with equality if and only if there is independence. If we call $H_{m,k}$ to the entropy of the distribution $\{P_{k_i,k_j}/P_{m,k} \; ; \; k_i+k_j=k\}$ then, by the branching property [1], we have $H_m = H_m^* - \sum_{k=0}^{R_m} P_{m,k} \cdot H_{m,k}$ and, therefore, $H_m \le H_m^*$ with equality if and only if there is no branching. Then $H_m \le H_m^* \le 2H_{m'}$ and both equalities cannot be

satisfied simultaneously, except when $H_m=0$.■

3. THE GIBBS NET

It is possible for an entropy-resolution diagram to fall down to zero at any resolution level, when the image is uniform at this resolution level (all m-regions have the same common grey level). But the growing of a diagram is bounded. In first time, we have Theorem 1, and the Principle of Maximum Entropy gives us another bound, as follows: Gibbs distribution at level m [9] is the one having maximum entropy H_m^G and being compatible with a fixed H_0. Then, any diagram for a particular image beginning at H_0, cannot have entropy at level m higher than the one of the Gibbs distribution. This is the most expected probability distribution at resolution level m, when H_0 is given.

If we join the entropies H_m^G/R_m for all Gibbs distributions being compatible with the same H_0, then we get a special entropy-resolution diagram (Gibbs sequence) which cannot be exceeded by any particular diagram beginning at the same point H_0. Thus, the net obtained with all the Gibbs sequences is useful as a reference when drawing particular diagrams. This fact is reinforced by the following two important properties:

Theorem 2: H_m^G *is a strictly increasing continuous function of* H_0. This means that the more information provided by an image at resolution level 0, the more upper bound of information can be reached by the image at any other resolution level.

Theorem 3: *Gibbs sequences* $\{H_m^G/R_m\}_{m\geq 0}$ *are strictly decreasing.* This means that the upper bound for the amount of information that can be obtained when analyzing an image decreases with the resolution level used for viewing it.

Proofs for theorems 2 and 3 can be seen in [6] and [9].

The bounding property of Gibbs sequences does not mean that sometimes a particular diagram could not cross one of them down to up when m increases, as can be seen in fig.1. One interesting problem arises when asking for sequences that avoid this circumstance. The answer comes when considering a variant of Gibbs point of view in the application of Principle of Maximum Entropy: if we seek the distribution at level m which maximizes entropy H_m constrained only by a given entropy H_{m-1} at previous resolution level, then a new set of sequences could be obtained by mean of a step-by-step Principle of Maximum Entropy. This sequences have the required property: no particular diagram can cross any of them down to up. By the moment, the construction, study and properties of such a set of sequences is an open problem.

4. SOME APPLICATIONS

4.1. Periodic Images

Fig. 2 shows the diagrams for a series of patterns whose cell sizes increase as 2^m. As it can be observed, the entropies decay to zero at a resolution matching the cell size. An image is no longer informative from this specific resolution.

When $H_m=0$ for an image, (and then $H_{m+j}=0$ too for all $j>0$), information is not provided: uniformity is not informative. In fact, this is true for *typical* images, so that their *quasi-entropy* is not significantly different from the entropy of the source of the images [4].

This feature of the diagrams appears when fitting the region size sequences (2^m) to the characteristic image cell sizes. Real images, of course, do not fulfill, in general, this condition. The trouble could be bypassed by adjusting the device which provides the image, or by working with different region size scales (other than 2^m) and choosing the best one.

4.2. Damaged Images

Fig.3 presents a grid with several noise levels added. Noise has been implemented by randomly changing all pixels throughout the image, according to a given probability (noise level). The general effect of adding noise is an increase in entropy for all resolution levels. However, the entropy decreases for noise levels close to the maximum, which loses the original image entirely.

For a net noise, the diagrams are monotonically increasing. Each histogram $\{P_{m,k}\}$ is a binomial with R_m tries and k successes. As it is well-known, these distributions go towards a gaussian one as R_m goes to infinity. The associated entropy is maximum among all distributions with the same standard deviation. Thus, the maximum value for H_m in the above diagrams (when they exist) should be justified by a composition between the original image and the noise. It is suggested that there is no possibility of reliable recovering of the original image when it has been damaged by noise beyond that maximum.

Fig.4 shows an object blurred differently. But now, we present the results in a modified diagram, $H_m/\log(R_m+1)$, having 1 as the maximum value for each m. These diagrams are useful for emphasizing details that are difficult to distinguish when m is high, because of the compression under the Gibbs bound. We point out how sensitive the diagrams are to a little blurring, and this fact may be used in restoration.

4.3. Backgrounds

Fig.5 shows the effect of having several textures with the same superimposed object on the diagrams at resolution less than a certain m_0, namely the particular resolution at which H without any texture would decay to zero. When the 'period' of the texture is much smaller than m_0, it should be easy to filter out the texture by different methods, including our own [9].

4.4. Contrast

Since Shannon's entropy is symmetric with respect to its arguments, it is possible to have different histograms with the same entropy by exchanging indexes in $\{P_{m,k}\}$. For instance, we have at first (fig.6) a checkerboard with black and white squares 8x8-sized. Then, the grey level of white squares is modified. It turns out that H_6 does not change, but H_m does it at $m<6$ because the 6-regions are made of different m-regions ($m<6$). F2 has been obtained by adding a fine-grained texture to white squares, so its diagram is only different from F1 at $m\leq 2$. In the other hand, F3 and F4 have been generated by adding random noise to white squares. This method allows more grey levels for the 6-regions, and then its diagrams are more different from that from F1. This example tells us that the contrast, a tool frequently used to make the information provided by images more reliable, does not always reveal itself in the diagrams.

REFERENCES

[1] ACZEL,J. *Measuring Information Beyond Communication Theory*, Information Processing and Management 20 (1984),3, 383-395.

[2] GULL,S.F., SKILLING,J. *The entropy of an image*, Maximum-entropy and bayesian methods in inverse problems, C. Ray Smith and W.T. Grandy Jr. (Eds.), D.Reidel Pub. Co., 1985.

[3] MARSHALL,A.W., OLKIN,I. **Inequalities: Theory of Majorization and Applications**, Academic Press, New York 1979.

[4] MASURIPUR,M. **Introduction to Information Theory**, Prentice Hall, Englewood Cliffs, N.J. 1987.

[5] PAL,N.R., PAL,S.K. *Entropic thresholding*. **Signal Processing 16** (1989), 97-108.

[6] MARTINEZ-AROZA,J., QUESADA-MOLINA,J.J., ROMAN-ROLDAN,R. *Statistical analysis of resolution in images*, this volume.

[7] QUESADA-MOLINA,J.J., ROMAN-ROLDAN,R. *On Some Entropies For Images*, **Proc. ICCI'89**, North Holland, New York 1989, 263-265.

[8] QUESADA-MOLINA,J.J., ROMAN-ROLDAN,R. *Resolution-Dependent Entropy For Images*, **Proc. ICCI'89**, Canadian Scholars' Press, Toronto 1989, 200-201.

[9] ROMAN-ROLDAN,R., MARTINEZ-AROZA,J., QUESADA-MOLINA, J.J. *A new Approach for Image Analysis*, to be published in Signal Proc.

[10] SKILLING,J. *Theory of Maximum Entropy Image Reconstruction*, paper presented at the 4th Workshop on Maximum Entropy and Bayesian Methods in Appl. Statistics, Calgary, 1984.

[11] WONG, A.K.C., VOGEL, M.A. *Resolution-dependent information measures for images analysis*, IEEE on Systems, Man and Cybernetics, Vol SMC-7, 1, Jan. 1977.

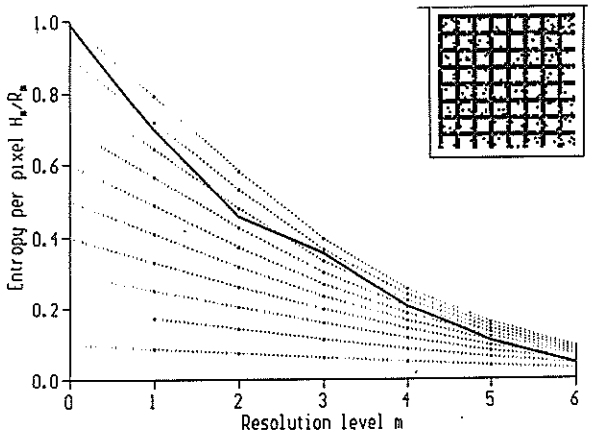


FIGURE 1

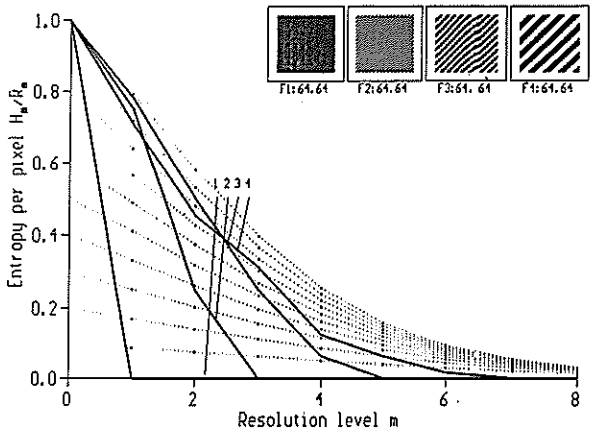


FIGURE 2

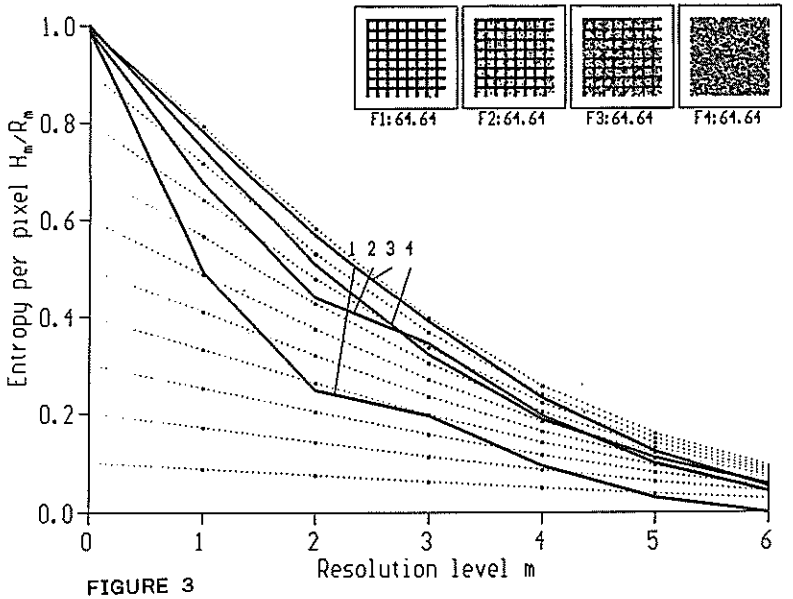


FIGURE 3

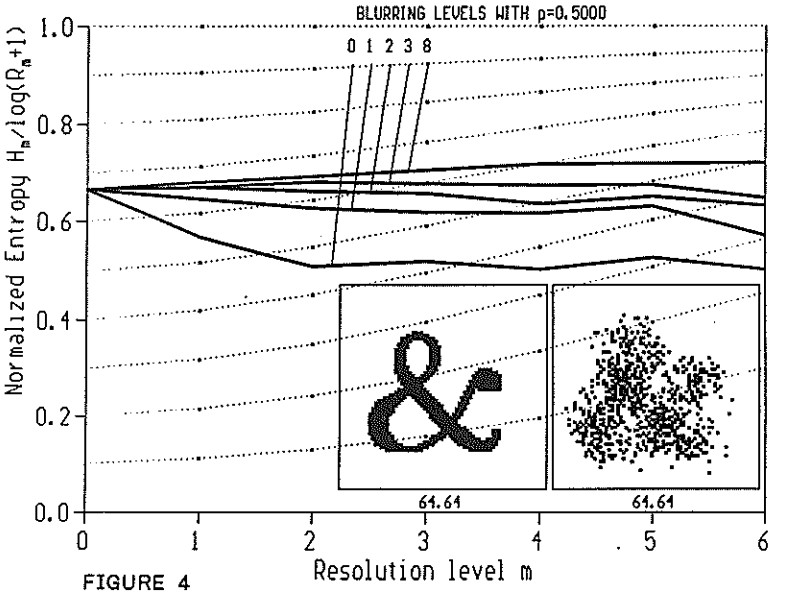


FIGURE 4

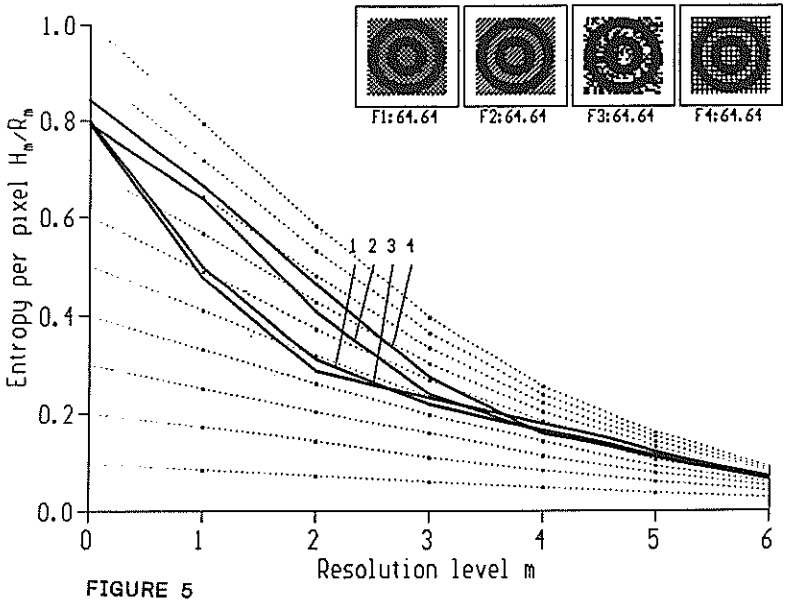


FIGURE 5

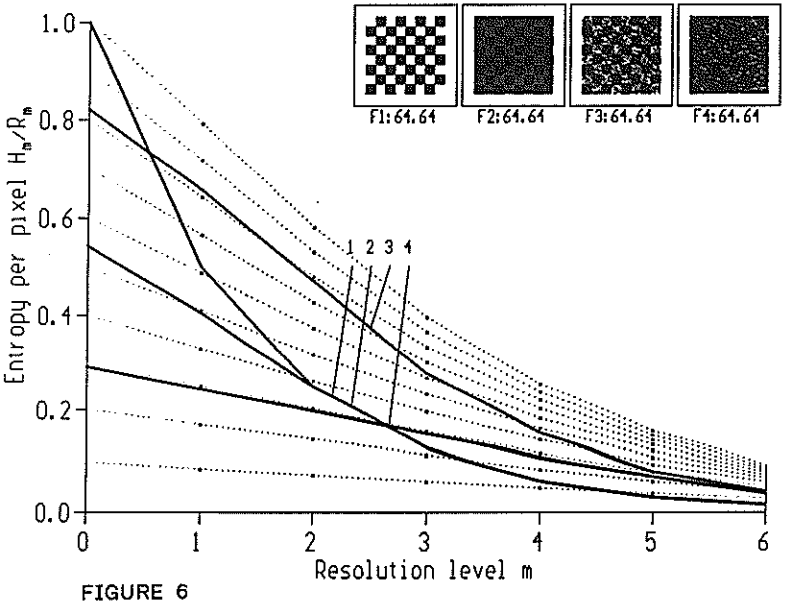


FIGURE 6

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

A STRUCTURAL APPROACH TO TOPOGRAPHIC LABELING OF DIGITAL IMAGES

G.Bordogna (*), D.Delfini (°), P.Mussio (°), A.Rampini (*)

(*) SIAM-IFCTR CNR via Ampere 56 20131 Milano, ITALY
(°) Physics Dep., Univ. degli Studi di Milano, via Viotti 7, 20135, MILANO, ITALY

A structural method to identify significant topographic features in digital images is presented. The image topographic characterization is accomplished by a two phase procedure based on successive labeling of the significant structures. First the digital image is labeled by associating to each pixel a digital code which summarizes the local topographic properties, then by means of Bidimensional L-Systems, a family of parallel recognition systems, contextual topographic properties of the image surface are identified.
An application in the astronomical field will be used to discuss the proposed method.

1. INTRODUCTION

In the approaches to image description based on surface characterization [1], a coloured digital image is studied as a surface in a 3-D space, in which at each point, identified by the pixel coordinates, the pixel grey level represents the surface height. This surface is described by denoting each structure in it by the name of the topographic entity it resembles, such as slope, valley, ridge and plateu, etc. (see for examples [2], [3], [4], [5], [6], [7]). In other words, a structure in the image is denoted by the name of a topographic entity (t-entity) if the grey levels of the pixels composing it satisfy a set of well defined topological and geometrical relations .

Topographic labeling is the activity by which these relations are evaluated so that the t-entities in the image are recognized and each pixel of the digital image can be mapped into a label identifying the name of the t-entity to which it belongs.

In our approach the pixels of a structure are first identified by the study of a local approximation of the surface gradient, the Generalized Slope Code (GSC), a generalization of the Slope Code defined in [8], which summarizes the local grey level variations of the digital surface in its 8-neighbourhood . GSC can be suited to the image model and to the application environment in order to neglect the grey level variations which an expert judges unimportant for the experiment. As it only describes local relations among pixel grey levels and is unable to recognize t-entities extended on variable neighborhoods larger that the local 3x3 cell, and characterized by contextual relations, an image is studied as a sentence of a finite bidimensional language. This language is defined via a generative device , a Bidimensional L-System (BLS) [9], [10], which defines the t-entities taking into account the variable 2-D contexts in the image in which each GSC may appear. In analysing an image by a BLS, each rewriting step maps a labeled image into another labeled image; each label denotes the structure to which the labeled pixel is associated at that stage of the description process.

The procedure for the recognition of t-entities in an image is therefore in two steps: the local topographic labeling by the GSC and the contextual labeling by BLSs.

The procedure is characterized by two main features: it is completely digital as all notions are

defined in the discrete space rather than in the continuous one, i.e. any approximation in the real space is avoided and , due to the local coding technique and to the use of L-systems, it is based on a parallel computational model.

2. LOCAL TOPOGRAPHIC LABELING: THE GENERALIZED SLOPE CODE

The local topographic labeling consists in summarizing the local surface trend around each pixel by a number derived by the comparison of its colour value with those of its 8-neighbours .

This number is called the Generalized Slope Code and is defined as follows :

$$(1)\quad GSC = \sum_{j=0}^{7} ((C_j - C) > h) * 2^{j}\ C_j, C \in N$$

where C is the grey level of pixel P, C_j is the grey level of the j-th neighbour of P numbered according to a clockwise sequence around P, and h is a positive threshold. GSC takes values in **GSC**={0,....,255}, i.e. it can be codified in a 8-bit variable in which the j-th binary element is set to 1 if the j-th neighbour's grey level C_j is greater than C+h, to 0 otherwise.

The value of the threshold h determines which grey level changes are considered important in the experiment, on the basis of the knowledge about the application environment, the adopted image model and the expected structural appearance of the sought objects.

For example, a threshold h=0 is adopted to study images in which no edges exist and/or image textures or different kinds of structures are to be selected. This can be useful for example when the experimenter is interested in detecting weak sources, like in [10], where tracks of faint astronomical sources were to be recognized. In this case the GSC with h=0 reduces to the SC discussed in [8].

By choosing a positive threshold h, all local grey level changes (cj-c) which are below h are neglected, while the greater ones are still taken into account: in this case the GSC operates as an edge detector. This is the case of the restoration of digitized technical drawings [11] in which the interesting tracks are characterised by abrupt variations of the grey level and therefore can be identified by detecting the corresponding sharp edges.

In the same way in the digital image of figure 1, a grey level representation of an astronomical field containing tracks of stars and a spiral galaxy acquired by a Charged Coupled Device, steep excesses from the background corresponding to intense signals can be detected.

In fact here the overall aim of the experiment is to identify and to describe galaxy structures like arms. To this end first the star and galaxy bulge tracks, which correspond to the most intense signals and are characterized by high grey level variations, are detected so as to subtract them from the image, thus permitting the weaker structures to be studied. Thereafter, an analysis based on a different threshold h points out pixels characterized by more gradual slopes and belonging to weaker tracks of interest. In any case the pixels labeled by the different classes of **GSC** can be selected and represented for visual inspection.

Figures 2 and 3 show those pixels of figure 1 whose GSCs equal zero for two different values of h. They are labeled by black points and identify surface points in the image lying either on plateaux or on relative maxima.

In particular, figure 2 is the result of the codification with the GSC using h=0. It can be noticed that almost all the black pixels are single points spread all over the image surface . They can either be tracks of structured excesses or background noise

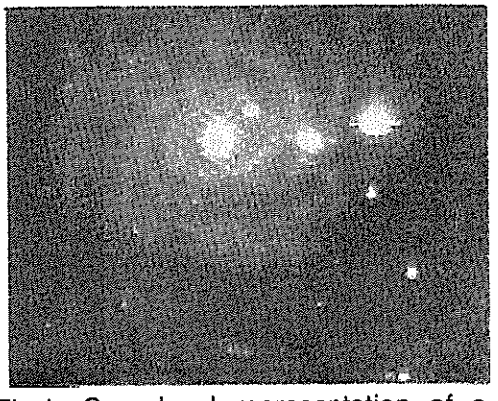


Fig.1. Grey level representation of a CCD Astronomical field.

Fig.2 Codified image of Fig.1 using GSC with h=0. Black points mark pixels with GSC=0, i.e. plateaux or maxima.

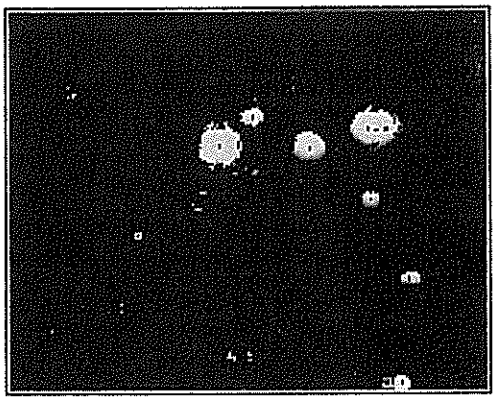


Fig.3 Codified image of Fig1 using h=20 The black points correspond to the flattened regions and maxima. The white round regions are the stars and the galaxy bulge.

spikes.

A threshold h>0 has been empirically determined which flattens the grey level variations corresponding to noise spikes or to weak galaxy structures. In this second case, the plateaux corresponding to the flattened regions and the relative maxima shown in figure 3 are marked by black points; here the white round regions with a central maximum point denote the presence of the most intense signals corresponding to the stars and the galaxy bulge.

3. CONTEXTUAL TOPOGRAPHIC LABELING BY BIDIMENSIONAL L-SYSTEMS

On the basis of the local labeling process, more complex topographic structures such as ridges, valleys, etc. can be identified by grouping together connected sets of pixels sharing common topographic properties.

In the proposal discussed in [10], this grouping is accomplished by propagating the information associated to certain pixels, for example pixels denoting a plateau whose GSC=0.

This propagation occurs in the eight directions across those pixels which are labeled by suitable codes. The process ends when the propagating signal meets a stopping label. All the pixels touched by the propagating signals are labeled by the same symbols. In this way , t-entities of any extension are detected within the image.

This process of propagation is described by a Bidimensional L-System, a recognition system defined only on one (bidimensional) alphabet, in which the rewriting occurs in parallel.

More formally a Bidimensional L-System (BLS) is a quadruple

$$BLS: <V, Rb, Ax, Mr>;$$

where V is an alphabet of bidimensional words. A bidimensional word on V is a matrix MxN (with $M,N \in \mathcal{N}+$) whose entries are in V:

$$\omega = \| wij \|, wij \in V, i=1,M, j=1,N$$

To describe the evolution of a bidimensional word, a set of bidimensional rules is defined.

The set Rb of bidimensional rules on V is a collection of couples $<\alpha,\beta>$ (denoted as $\alpha->\beta$) where α,β are bidimenional words on V of the same dimension. Ax is a set of bidimensional words over V named the set of axioms. Mr is a set of metarules which specifies how to use the rules in Rb to derive an axiom from a given string P on V.

The parallel bidimensional direct generation process is stated as follows: a bidimensional word η over V directly generates a bidimensional word ω if each subword η_k of η is rewritten into a subword ω_k of ω, and $<\eta_r, \omega_r> \in$ Rb.

In the examples mentioned above, the BLSs are used to perform the recognition of structures like "ridges", "valleys" and

"maximum and minimum plateaux".

The first step is the identification of directional t-entities: for each pair of opposite directions across the pixel, called side direction, four directional t-entities can occur: directional ridges, directional valleys, directional growths and directional decreases. In table 1 the BLS for directional ridges recognition (along the first side direction) is shown: similarly the BLSs for the other directional ridge recognition can be defined.

From this step it descends that four labels are associated to each pixel, one for each side direction; the label identifies one of the four directional t-entities.

A final BLS integrates this information associated to the pixel to infer the global topographic labeling. For example if a pixel has been labeled "directional ridge" in all four side directions it is recognized as part of a maximum plateau; if it has been labeled "directional ridge" ("directional valley") in three side directions and "directional valley" ("directional ridge") in the last side direction it is recognized as a saddle point part of a ridge (valley).

Figure 4 shows the results of the contextual analysis aimed at identifying ridges and valleys in the image in figure 1 It can be seen how the spiral ridges protruding from the galaxy bulge identified by a circular valley constitute the skeleton of the galaxy arms.

Table 1

BLS for Directional Ridge recognition: D-Ridge (1)

Alphabet V:= { **GSC** , 301, 401 }

Axioms:= matrix on V

Rules:=

$$\begin{pmatrix} a5 & b \\ c & F1 \end{pmatrix} \rightarrow \begin{pmatrix} a5 & b \\ c & 401 \end{pmatrix}$$ starting rule

$$\begin{pmatrix} 401 & b \\ c & F1 \end{pmatrix} \rightarrow \begin{pmatrix} 401 & b \\ c & 401 \end{pmatrix}$$

$$\begin{pmatrix} 401 & b \\ c & 301 \end{pmatrix} \rightarrow \begin{pmatrix} 301 & b \\ c & 301 \end{pmatrix}$$

$$\begin{pmatrix} 401 & b \\ c & a1 \end{pmatrix} \rightarrow \begin{pmatrix} 301 & b \\ c & a1 \end{pmatrix}$$

$b,c \in V$

$ai \in$ Growth(i):= set of **GSC** denoting pixels whose grey level is lower than the grey level of the i-th neighbour

$Fi \in$ Flat(i) := {**GSC** - (Growth (i)$\cup$ Growth (8|i+4) }

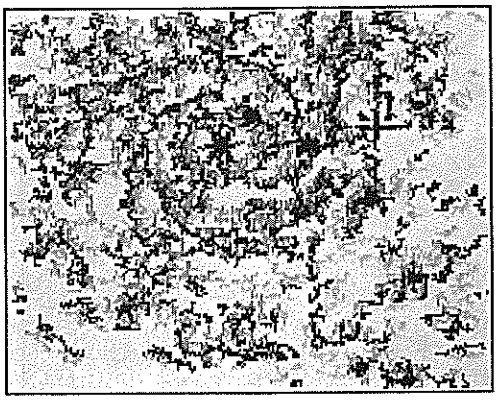


Fig.4 Result of the application of the contextual analysis for Ridge and valley recognition on the codified image in Fig.2 using h=0. The ridges are in black, the valleys in grey.

REFERENCES

[1] Chang, S.K., Principles of pictorial information systems design, Prentice-Hall, N.J., (1989).

[2] Meyer F., Signal Processing (1989) 16, pp.335-363.

[3] Peucher, T., Douglas , D. H., Computer Graphics and Image Processing (1975) 4, pp.375-387.

[4] Watson, L.T., Laffey, T.J., Haralik, R.M., Computer Vision Graphics and Image Processing (1985) 29, pp.143-167.

[5] De Floriani, L., Falcidieno, B., Pienovi, C., Computer Vision Graphics and Image Processing (1985) 32, pp.127-140.

[6] Nackman, L.R.,IEEE Trans. on Pattern Analysis and Machine Intelligence (1984) PAMI-6-4, pp.442-449.

[7] Besl, P., Jain, R.C., Computer Vision Graphics and Image Processing (1986) 6, pp.77-85.

[8] Accomazzi, A., Bordogna, G., Fresta, G., Mussio, P., Rampini, A., Signal Processing IV: Theories and applications, North Holland (1988), pp. 1601-1604.

[9] Salomaa, A., Formal languages, Academic Press, New York (1973).

[10] Accomazzi, A., Bordogna, G., Mussio, P., Rampini, A., Data Analysis in Astronomy III, Plenum Press, New York and London (1988), pp. 245-254.

[11] Cugini,U., Ferri,M., Mussio,P., Protti,M., Computer & Graphics (1984) 8, pp.337-350.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

AN HOMOMORPHIC METHOD FOR CRYSTAL QUALITY ESTIMATION

Juan P. Secilla (†) and Narciso García, (‡)

(†) IBM Madrid Scientific Centre - Paseo de la Castellana, 4 - 28046 Madrid - Spain
(‡) E.T.S. Ing. Telecomunicación - Universidad Politécnica - 28040 Madrid - Spain

Estimation of the quality of crystals is a problem that arises in several areas of Science. The most commonly used technique for this purpose is the evaluation of the two-dimensional autocorrelation function of the image. However, the results obtained are not always good. Here, we propose an alternate method based on evaluating the two-dimensional cepstrum of the image of the crystal, liftering it to retain the part accounting for the periodicity of the image, and computing the inverse transformation. The result is an image that provides a good qualitative estimator of the quality of the crystal under study. An example is presented in which the performance of the new method is compared to that of the autocorrelation function.

1. INTRODUCTION

The digital processing and analysis of pictures of crystals is a very useful way to obtain structural information in several areas of Science. So, in Molecular Biology[1] it has been used in the determination of the three-dimensional structure of several viruses. In Inorganic Chemistry[2] it has been applied to ascertain the spatial structure of certain compounds, etc.

The general problem that these techniques try to solve is that of obtaining information in those situations where the existence of very high noise tends to hide it. Well established filtering techniques usually fail in these situations. So, the usual strategy is to average several instances of the same sample or specimen in order to keep the common part of them, thus eliminating most effects due to noise and defects arisen in the preparation process. Using crystals for this purpose has the advantage that all the samples in the image are periodically arranged and have the same orientation. So, filtering them is quite straightforward, either in the spatial or the spectral domain[3].

Crystals are, however, subject to defects very often. This means that they are not strictly periodical. In these situations, those filtering methods that rely on the periodicity of the samples fail. For this reason, an important step before filtering a crystal through averaging is to estimate its periodicity. In this way, it is possible to determine in advance whether the sample is valid for further processing or not.

The usual method to determine the periodicity of a periodical image is the two-dimensional autocorrelation function[4]. Evidently, this function should exhibit sharp peaks that reflect the periodicity of the original image. However, its performance is not always very good. In this work, we propose liftering the cepstrum of the image to achieve the same goal. In this way, an image that reflects the periodicity of the original picture is achieved. This image provides an usually good estimation of the periodicity of the original sample.

2. DESCRIPTION OF THE PROBLEM

An image of a perfect crystal can be defined as

$$c_1(\bar{r}) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} m(\bar{r} - i\bar{a} - j\bar{b}) \qquad (1)$$

where $\bar{r} = (x,y)$, $m(\bar{r})$ is the basic motif that is periodically repeated along the crystal and $\bar{a}$ and $\bar{b}$ are the lattice vectors that define the periodicity of the picture.

When there exist geometric distortions, the situation changes. Let $\bar{r} = T(\bar{r})$ be the equations that describe the inverse geometric distortions of the crystal. A crystal with those distortions, c_2, would be defined by

$$c_2(\bar{r}) = c_1(T(\bar{r})) \tag{2}$$

In real situations, besides, noise is present, and the imaging system has also influence in the image available. If the point spread function (PSF) of the imaging system is $p(\bar{r})$, and there is only additive noise, $n(\bar{r})$, a real crystal c_3 can be expressed as

$$c_3(\bar{r}) = c_2(\bar{r}) \bigstar p(\bar{r}) + n(\bar{r}) \tag{3}$$

where the symbol $\bigstar$ represents the two-dimensional convolution.

Very often, these effects have a strong influence on the crystal. In these situations, determining the periodicity of the crystal becomes critical. The most widely used methods for this purpose are two:

1. In the Fourier domain: although the original image may have been strongly corrupted by the effects described previously, it usually keeps a strong periodical component. So, its two-dimensional Fourier transform should exhibit sharp peaks at integer multiples of the lattice vectors in the Fourier domain (Bragg's peaks). These peaks usually show much information about the crystal quality.

2. In the spatial domain: a method to study the crystal quality in the spatial domain is evaluating its ACF (usually using the FFT algorithm to speed up computations). For reasons that become evident, this ACF will show strong peaks at integer multiples of the lattice vectors $\bar{a}$ and $\bar{b}$. The stronger the periodicity of the image, the more sharp and intense the peaks are, so it is possible to estimate the crystal quality by means of a simple glance to its ACF. Herein on, we shall concentrate on this second kind of methods.

3. THE PROPOSED METHOD

The complex cepstrum[5] of a two-dimensional function $f(\bar{r})$ is defined as

$$\hat{f}_c(\bar{r}) = F^{-1}[\log(F[f(\bar{r})])] \tag{4}$$

where F denotes the Fourier transform.

The conditions for the existence of the two-dimensional complex cepstrum are rather strict[6]. Besides, it is necessary to previously compute the unwrapped phase[7] of the Fourier transform before taking the complex logarithm. Since we are interested, however, just in estimating qualities, it is not necessary to use the complex cepstrum. Using the two dimensional cepstrum is more convenient in this case. It is defined as

$$\hat{f}(\bar{r}) = F^{-1}[\log|F[f(\bar{r})]|] \tag{5}$$

In cases where there are zeroes in the Fourier transform, it is possible to extend the definition of the cepstrum to circumvent this problem by substituting it with $-M$, where M is a big positive number.

By computing the cepstrum of a picture, convolutions are mapped to additions. In our case, a perfect crystal can be expressed as

$$c_1(\bar{r}) = m(\bar{r}) \bigstar \left[\sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \delta(\bar{r} - i\bar{a} - j\bar{b}) \right] \tag{6}$$

So, the motif gets convoluted with a two-dimensional pulse lattice, $t(\bar{r})$. The relation between the cepstra is the following:

$$\hat{c}_1(\bar{r}) = \hat{m}(\bar{r}) + \hat{t}(\bar{r}) \tag{7}$$

The cepstra of the motif and of the pulse train are thus additively combined. Usually, it is straightforward to separate them (a circular, ideal high pass filter generally suffices). So, it is possible to keep only the cepstrum of the pulse lattice $\hat{t}(\bar{r})$. By performing the inverse operations (computing the Fourier transform, exponentiating and computing the inverse Fourier transform), it is possible to get an estimate of the pulse lattice (the cepstrum of a two-dimensional pulse train is another pulse train with the same lattice vectors, but with different, decaying amplitudes).

Once the cepstrum has been liftered, the image is recovered using the expression

$$c_f(\bar{r}) = F^{-1}[\exp(F[\hat{c}_f(\bar{r})])] \tag{8}$$

where $\hat{c}_f(\bar{r})$ is the liftered cepstrum. This image will show peaks at integer multiples of the lattice vectors $\bar{a}$ and $\bar{b}$. The stronger the periodicity of the initial picture, the higher the definition and intensity of these peaks. So, the crystal quality can be subjectively evaluated very quickly by simple visual inspection of the resulting image.

4. RESULTS

Figure 1a shows an hexagonal crystal of necks of the virus bacteriophage Φ29, obtained in an electron microscope. Figure 2a presents its 2D cepstrum. This cepstrum is composed of a part which corresponds to the crystal motif, which is located around the origin, and a 2D pulse lattice, that corresponds to the periodicity of the crystal. In Figure 2b we show the liftered cepstrum, in which only this latter part has been kept.

Figure 2c displays the image recovered using (8). It is possible to observe a strong 2D pulse lattice, which indicates that the crystal quality is relatively good. Figure 1b shows the ACF of the original image. Although there is also a 2D series of regularly arranged peaks, it is much weaker and presents less information.

In Figure 1c, an image of the motif of the crystal, after filtering through conventional ensemble averaging of all the crystal motifs, is shown. The quality of this image confirms that the original crystal was of good quality.

5. CONCLUSIONS

A new method for the evaluation of the quality of crystals has been presented. This method is based on computing the 2D cepstrum of the image, liftering it, and performing the inverse homomorphic transformation. This method has shown better performance than the traditional ACF method in a variety of test and real pictures of crystals.

This is a promising case of the use of homomorphic techniques in image processing of crystals. We feel that this sort of techniques can be widely used in this and related disciplines.

ACKNOWLEDGEMENT

We gratefully acknowledge Dr. José L. Carrascosa of the Centro de Biología Molecular for providing us with pictures of several biological crystals.

6. REFERENCES

1. Carazo, J.M., Donate, L. E., Herranz, L., Secilla, J. P., and Carrascosa, J.L., "Three-dimensional reconstruction of the connector of the phage Φ29 at 1.8 nm. resolution", Journal of Molecular Biology, vol. 192, pp. 853-867, 1986.
2. Hövmoller, S., Sgröjen, A., Farrants, G., Sundberg, M., and Marinder, B. O., "Accurate atomic positions from Electron Microscopy", Nature, Vol. 233, pp. 625-633, 1986.
3. Amos, L. A., Henderson, R., and Unwin, P. N. T., "Three-dimensional structure determination by Electron Microscopy of two-dimensional crystals", Prog. BioPhys. Mol. Bio., Vol. 39, pp. 183-231, 1982.
4. Frank, J. "The role of correlation techniques in computer image processing", in J. K. Koehler, editor, Advanced techniques in biological Electron Microscopy, pp. 215-269, Springer-Verlag, 1973.
5. Oppenheim, A. V., and Schaffer, R. W., Digital Signal Processing, pp. 480-531, Prentice-Hall, 1975.
6. Dudgeon, D. E., "The existence of cepstra for two-dimensional rational polinomials", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-33, pp. 242-243, 1975.
7. Tribolet, J. M., "A new phase unwrapping algorithm", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-25, pp. 170-177, 1977.

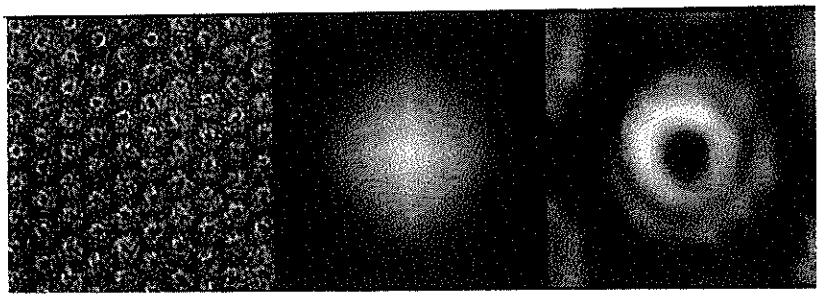


Figure 1. : a) Image of an hexagonal crystal of necks of virus bacteriophage Φ29 obtained in an electron microscope. b) Its autocorrelation function. c) Filtered motif obtained through ensemble averaging.

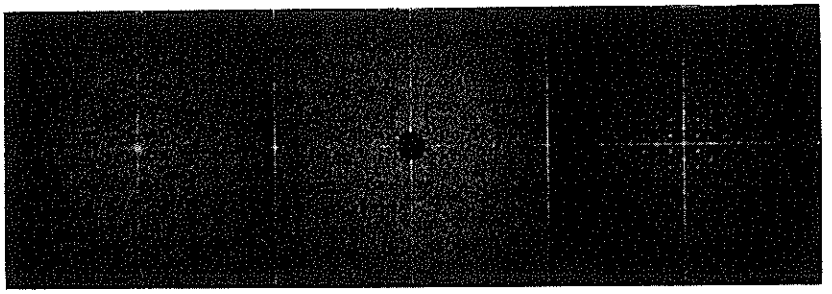


Figure 2. : a) 2D cepstrum of the image shown in Figure 1a. b) 2D cepstrum after being high pass liftered. c) Image recovered after inverse homomorphic transformation.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ANALYSIS AND MODELING OF FLAME IMAGES

L. Bordoni, A.G. Federico

ENEA C.R.E. Casaccia
Via Anguillarese km 1.300
ROME (ITALY)

The development of a monitor system for the control of the flame in firebox for thermoelectric plants is presented. The aim of the system is to improve the performance of the burners, the efficiency and the power of the plant.This is achieved by the syntesis of a thermic tridimensional map which is obtained using a methodology based on the image processing techniques. An improved algorithm for the thermic tridimensional map of a single flame is presented.

1. INTRODUCTION

In a stand-alone configuration, an IBM PC provides for the acquisition, retention and display of the data and a block diagram of the system is shown in figure 1. With this system several series of flame images produced by an experimental burner are acquired. An algorithm based on the tomographic techniques, appropriately modified has been developed.
The thermic map is represented as an image in false colours in which every colour is representative of a range of temperature. Profiles of temperature for different cross sections are also obtained.
The sperimental tests have been executed at the CSM Ansaldo of Genoa and the following actions on the acquired data have been performed :

- spatial differentiation;

- filterings.

A further step has concerned the implementation of an algorithm for temperature reconstruction. The characterization of scalar quantities of the flame is requested by this algorithm.

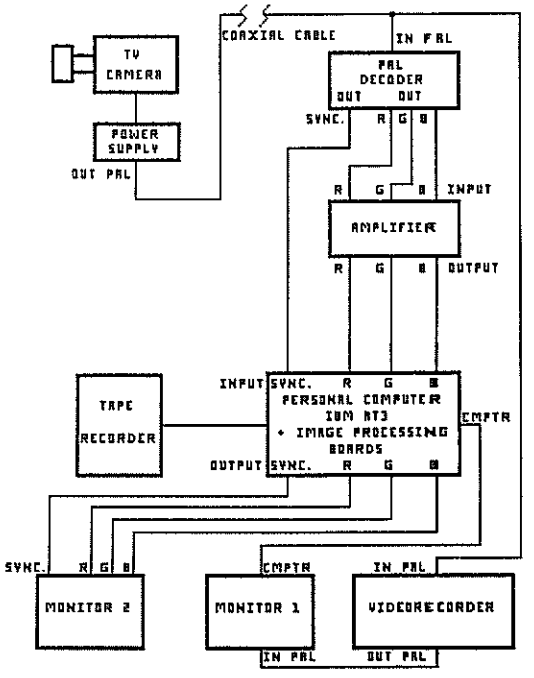


Figure 1

2. SYSTEM OVERVIEW

Figure 2 is a sequence of sixteen flame images acquired at time intervals of 40 milliseconds with a CCD TV camera. Because the blue light emission is negligible, the CCD camera uses only the red and the green channels. The first flame image on the upper left represents the temporal average of instantaneous flames.
This flame image allows to obtain some useful informations on the burner's symmetry, and also to apply the temperature algorithm reconstruction.

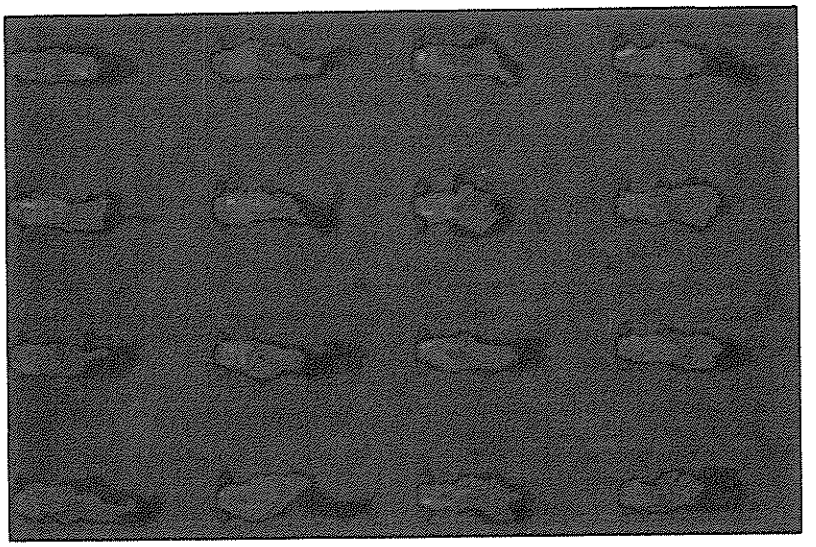


Figure 2

The two-colour pyrometry on this flame image is applied. This method consists in the detection of the luminous intensity at two distinct wavelengths. When the emissivity is indipendent from the wavelength it is possible to obtain univocally the temperature by the ratio of the luminous intensity in two wavelengths.
As shown in figure 3 the temperature is represented by the histogram. The vertical line intersects the flame in a section where the temperature is computed. The superimposed diagram on the flame image is the result. The temperature is represented in Kelvin degree.
Filtering operations are performed in order to remove the noise. Median filter is extremely effective, but the Laplacian filter allows to obtain an image in which the local contrast is emphasized.
The Sobel filter allows to display the gradients of temperature by a representation in false colours.
As shown in figure 4 the contours of flame images are emphasized because in those areas the gradients of temperatures are greatest.
Particularly, the blue-magenta colour represents the upper edges of the flame, whereas the lower edges are green-yellow.
The areas of flame with uniform luminosity are black. This type of elaborations can be useful to perform researches on the stability and pollution emission of a flame.

3.TEMPERATURE RECONSTRUCTION ALGORITHM

The pyrometry does not apply to the flame images since every point of image represents the contributions of light coming from areas with different temperatures. An approach based on tomographic techniques, in which the flame is assumed to be substantially symmetric, is presented.
A generic cross section of flame subdivided into concentric crowns is considered separating the varied emissivity's contributions in the valuation of the temperatures.
Two different techniques of tomographic reconstruction has been developed [1] :

- Onion Peeling (OP)
- Perturbative (P).

Both methods consider the flame as a set of circular concentric crowns and the reconstruction starts from the external edge and continues towards the centre. These methods are an improvement with respect to the two-colour analysis before described.
An algorithm MP (modified perturbative), based both on the perturbative method and on the two-colour technique, is developed in order to obtain an efficient tool of research. In this way the principal feature of the two-colour pyrometry, i.e. the robustness at the noise , is added to the advantages of the tomographic reconstruction technique.

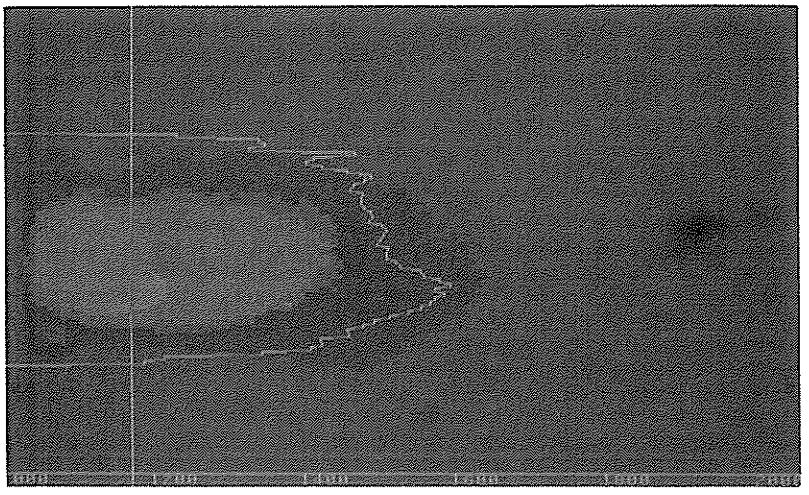


Figure 3

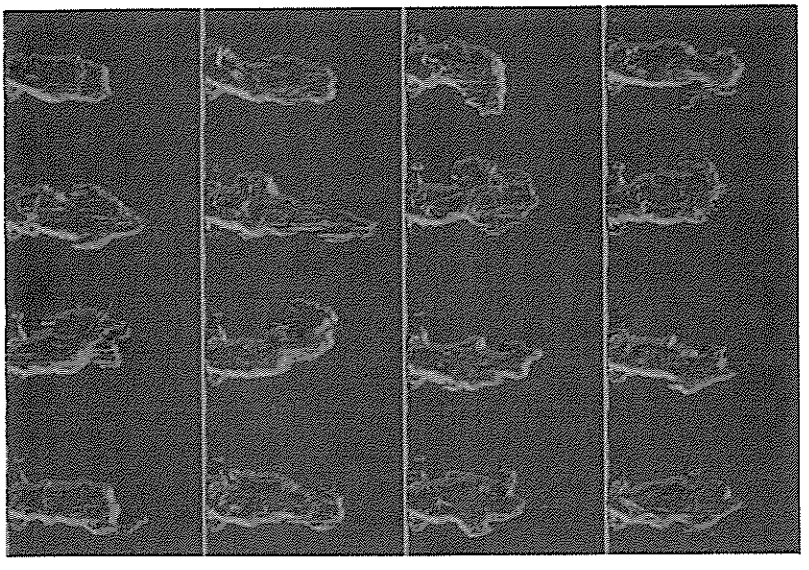


Figure 4

The approach followed is the same of the P method, but it works like the pyrometry two-colour technique on the ratio of the luminous intensity in two wavelengths, rather than on the absolute values of these intensities. This algorithm requires the calculation of some scalars.
These concern the geometric features of the flame (length, area) and the locating of the axis of symmetry. This allows to apply the algorithms of temperature reconstruction in that the flame is supposed to be with a cilindrical symmetry.

4. RESULTS

The MP algorithm is implemented in C language on a IBM PS/2.
The software allows to perform these operations:

- choice of the flame image in which to operate (red and green component);

- calculation of the scalars;

- temperature reconstruction.

The thermic map is represented as a flame image in which each colour is representative of a range of temperature (Kelvin degrees).

The results generated from this algorithm are equal to the values expected. This has been applied in the central areas of the flame and here the values of temperature are between 1400 and 1700 Kelvin degrees. The very important feature of PM algorithm is the lack of sensibility to the errors of measure, just the opposite of the OP and P algorithms. Furthermore the MP algorithm converges more rapidly than the OP and P algorithms.

5. CONCLUSION

It has presented the MP algorithm and has been shown that its storage and computational costs are lower than the other existing methods.
However in order to obtain a characterization of flame images as much as possible meaningful the calibration of the system is essential. In fact the MP algorithm appears very sensitive to little variations of some input values, for istance the wavelength.

REFERENCES

[1] Studio di metodologie per la ricostruzione di mappe termiche, relazione tecnica DIBE (Universita' di Genova, 1987).

[2] Gilardi, E., Rapporto tecnico conclusivo 01 (Ansaldo Ricerche, 1987).

[3] Rizzo, E., Rapporto tecnico conclusivo 02 (Ansaldo Ricerche, 1987).

[4] Verrecchia, P., Rapporto tecnico conclusivo 03 (Ansaldo Ricerche, 1988).

[5] Verrecchia, P., Rapporto tecnico conclusivo 04 (Ansaldo Ricerche, 1988).

[6] Verrecchia, P., Rapporto tecnico conclusivo 05 (Ansaldo Ricerche, 1989).

[7] Verrecchia, P., Rapporto tecnico conclusivo 06 (Ansaldo Ricerche, 1989).

[8] Young, T.Y. and Fun, K., Handbook of pattern recognition and image processing, (Academic Press, 1986).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

UNIVERSAL PATTERN-MATCHING INTERFRAME CODING OF VIDEO SIGNALS

Takahiro SAITO +, Ryuji ABE +, Takashi KOMATSU + and Hiroshi HARASHIMA ++

+ Department of Electrical Engineering, Kanagawa University
3-27-1 Rokkakubashi, Yokohama, 221, Japan
++ Engineering Research Institute, The University of Tokyo
2-11-16 Yayoi, Tokyo, 113, Japan

To escape the pattern-set mismatch problem, we formerly enhanced the coding approach based on pattern-matching with both the concept of self-organization of a pattern-set and the concept of coding via copying, thus developing a novel universal pattern-matching coding system for compression of still images. Extending the concepts to low-rate video compression, the work herein develops a new universal pattern-matching interframe coding system and demonstrates its superiority over the conventional scheme by means of computer simulations.

1. INTRODUCTION

The spatial vector quantizer (SVQ)[1] can be defined as a kind of pattern-matching coder where pel subblocks are replaced by some matching-pattern. In SVQ, the pattern-set (codebook), a collection of possible matching-patterns, is generated from training images. If the statistical property of a compressed image is different from that of the training images, excess quantization error results, which is referred to as the pattern-set mismatch problem.

To escape this problem basically, we formerly enhanced the SVQ approach with both the concept of self-organization of a pattern-set and the concept of coding via copying, and developed a universal pattern-matching coding system for compression of a still image[2]. We refer to this coding system as PMC for short. The concept of self-organization of a pattern-set was first applied to facsimile data compression by Pratt et. al.[3][4]. Furthermore the concept of coding via copying was first introduced for 1-dimensional reversible (noiseless) coding by Ziv and Lempel [5][6], and we subsequently extended the concept to 2-dimensional irreversible coding[2]. We demonstrated the superiority of the PMC system over the adaptive DCT coding system at a low bit rate by means of computer simulations[2].

The work herein extends the concept of PMC to low-rate video compression, and develops a new universal pattern-matching interframe coding system, called PMIC for short. The concept of PMIC comprehends the conventional interframe coding with the block-matching displacement estimation[7] as a specific variant, and computer simulations on monochrome image sequences demonstrate that the PMIC system outperforms the conventional scheme.

2. UNIVERSAL PATTERN-MATCHING CODING (PMC)[2]

2.1 Concept

PMC employs the concept of coding via copying. The concept is to encode future subblocks with the waveform distortion less than the allowed distortion value D* via a maximum-length copying from a frame memory containing the recently decoded subblocks. If any copying operation does not yield the waveform distortion less than the allowed distortion value D*, then the incoming subblock is encoded by some conventional image coding system, which is referred to as residue coding. Figure 1 illustrates the copying operation, which is identified by both the address of the starting point and the number of the copied subblocks. Compression is achieved by transmitting them.

This concept, however, is not feasible, because it takes a large computational effort to search the frame memory for the optimum starting point. To alleviate the computational requirement, we restrict a search area used for determining the starting point to some small region. We define all the subblocks contained in the search area as a subblock-template. The coder compares an input subblock with subblock-templates, and starts the copying operation at the location of the subblock-template yielding the best match. This search procedure is identical to that employed in the block-matching displacement estimation[7].

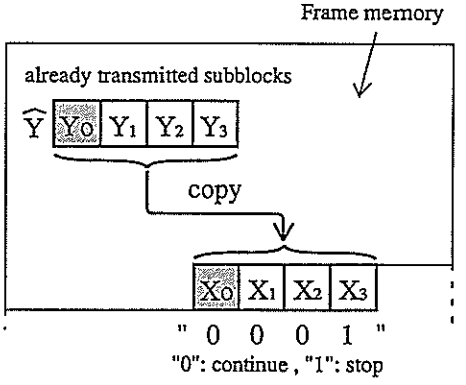


Figure 1 - Copying operation

To restrict a search area reasonably, PMC employs the concept of self-organization of a search area. The search area is restricted to some small regions A_{00}, A_{01}, ..., A_{0m} as shown in Fig.2. Both the near-search-area A_{00} and the far-search-area A_{0j} (j>0) contain 16 subblock-templates. The set of the far-search-areas is empty at the beginning of the coding process, and is gradually built up by adding the square region containing the residue-encoded subblock to the set and deleting the least useful far-search-area from the set. We adopt the sequential list employing the move-to-front heuristic[8] as an auxiliary data structure in order to maintain the set reasonably.

2.2 Coding Algorithm for Compression of Still Images

The coding algorithm for compression of still images is described below :

(0) Input subblock :
Partition a given input image into nonoverlapping square subblocks of 4 x 4 pels, scan subblocks from left to right and from top to bottom, and feed them to the coder as an input subblock.

(1) Search the near-search-area :
Compare a given input subblock with all the subblock-templates contained in the near-search-area, and determine the best subblock-template yielding the minimum waveform distortion. If the minimum distortion is less than the allowed distortion value D*, then start the copying operation at the location of the best subblock-template ; if not, proceed to the step (2).

(2) Search the set of the far-search-areas:
Choose the best subblock-template yielding the minimum waveform distortion from among all the subblock-templates contained in the set. If the minimum distortion is less than the allowed distortion value D*, then start the copying operation at the location of the best subblock template ; If not, proceed to the next step (3).

(3) Residue Coding :
Encode a given input subblock by using the conventional image coding system, and update the set of the far-search-areas.

In the copying operation, the number of the copied subblocks is determined by the following procedure. In Fig.1, the coder computes the waveform distortion between the input subblock X_1, and the past subblock Y_1. If the computed distortion is less than the allowed distortion value D*, then the coder continues the copying operation and applies the above procedure to the next input subblock X_2 ; if not, the coder stops the copying operation.

To better not only a rate versus distortion performance but also reconstruction quality, the allowed distortion value per subblock D* is adaptively controlled according to a standard deviation σ of an input subblock :

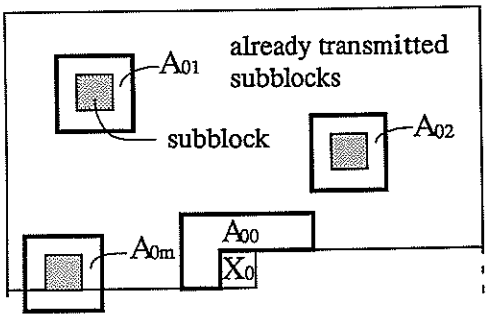
$$D^*(\sigma) = \begin{cases} d\sigma^T & , \sigma \geq 1 \\ d & , \sigma < 1 \end{cases} \qquad (1)$$

, where the coding rate depends on the parameter d.

In addition the coder transmits the overhead information identifying the coding mode.

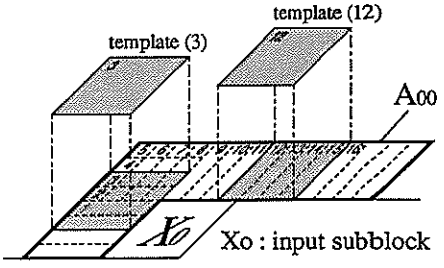
3. EXTENSION TO VIDEO COMPRESSION

We extend the concept of PMC to low-rate video compression, and develop a universal pattern-matching interframe coding system called PMIC for short. The concept of PMIC comprehends the conventional interframe coding as a specific variant.

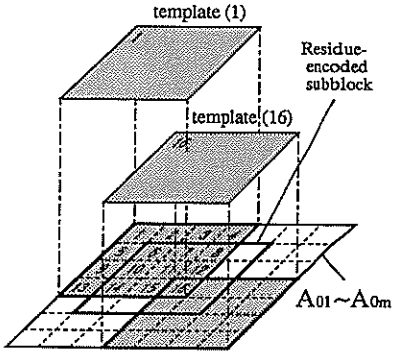


Xo : input subblock
Ao : near-search-area
A_1~A_m: far-search-areas

(a) Search area



(b) Near-search-area



(c) Far-search-area

Figure 2 - Definition of search area

The initial frame of an image sequence is encoded as a still image by the same coding algorithm as described in Chap.2. In frames succeeding to the initial frame the coder searches not only the present frame to be decoded but also finite past decoded frames for determining the proper starting point of the copying operation. We herein define the search areas used for determining the starting point as shown in Fig.3.

In Fig.3, the regions A_{00}, A_{10}, ..., A_{n0} are referred to as the near-search-area, whereas the regions A_{0j} ($j \neq 0$) are referred to as the far-search-area. The near-search-area A_{00} of the present frame K is defined as shown in Fig.2(b), and referred to as the present-near-search-area. The far-search-areas are defined as shown in Fig.2(c), and their set is maintained in the same manner as stated in Chap.2. The near-search-area A_{10} of the previous frame K-1 is referred to as the previous-near-search-area, whereas the near-search-areas A_{20}, ..., A_{n0} of the past frames K-2, ..., K-n are referred to as the past-near-search-area. The previous-(past-) near-search-areas is defined as 14 x 14 (10 x 10) square region which centers the decoded subblock in the respective frame at the same spatial position as an incoming subblock, and contains 121 (49) subblock-templates, which are classified into two categories, one stationary subblock-template and 120 (48) displaced subblock-templates. The two categories of the subblock-templates are treated separately. The stationary subblock-template is defined as the subblock-template at the same spatial position as an incoming subblock, whereas the displaced subblock-template is defined as a subblock-template outside the stationary subblock-template.

We explain the conceptual intentions of several search areas accoding to a simple image model [9] as shown in Fig.4, which is restricted to rigid objects translatorily displaced in the image plane. In Fig.4, the frame is divided into four different regions : moving objects, stationary backgrounds, uncovered backgrounds, and backgrounds to be covered. Figure 4 conceptually shows the copying operations in these different image regions. In the regions of stationary backgrounds and moving objects with small displacement, the coder will probably copy from the previous-near-search-area. In the region of stationary backgrounds, however, if large temporal changes between the present and previous frames are caused by noise or illumination changes, the coder will probably copy from one of the past-near-search-areas. Additionally, in the uncovered background region, the coder will probably copy from one of the past-near-search-areas. In the region of moving objects, if the displacement is very large, the coder will probably copy from the present-near-search-area, because there must be a high spatial correlation due to the effect of camera integration. Furthermore, in the case of a scene change, the coder will probably copy from the present near-search-area or one of the far-search-areas.

The coder uses various types of search area to search for the starting point of the copying operation. The search and coding procedures are organized in the order of (1), (2), ..., (7) as described below.

(1) Can the coder copy from the stationary subblock-template in the previous-near-search-area ? If no, proceed to the step (2).
(2) Can the coder copy from any stationary subblock-template in the past-near-search-areas ?
If no, proceed to the step (3).
(3) Can the coder copy from any displaced subblock-template in the previous-near-search-area ? If no, proceed to the step (4).
(4) Can the coder copy from any displaced subblock-template in the past-near-search-areas ?
If no, proceed to the step (5).
(5) Can the coder copy from the present-near-search-area ?
If no, proceed to the step (6).
(6) Can the coder copy from any far-search area ? If no, proceed to the step (7).
(7) Residue Coding

In addition the coder dynamically variable-length-encodes[10] the overhead information, by which the decoder is informed which coding mode the coder has chosen from among the above-mentioned 7 coding modes.

The PMIC system is simulated on monochrome image sequences under the following conditions :
(1) The test sequence is composed of 60 frames, and each frame consists of 240 lines and 256 pels per line. The sequence is composed of four different scenes each of which contains objects that move violently, and scene changes take place at the 16-th, 31-st, and 46-th frames.
(2) In Fig.3, the value of n is fixed at 9. The coder, then, can search at most 9 past decoded frames.
(3) The size of an input subblock is 4 x 4.

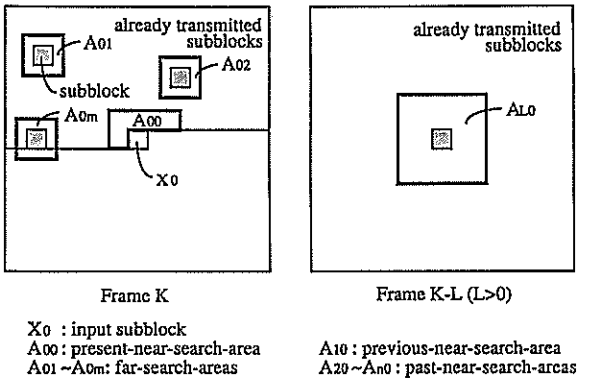


Figure 3 - Definition of search area

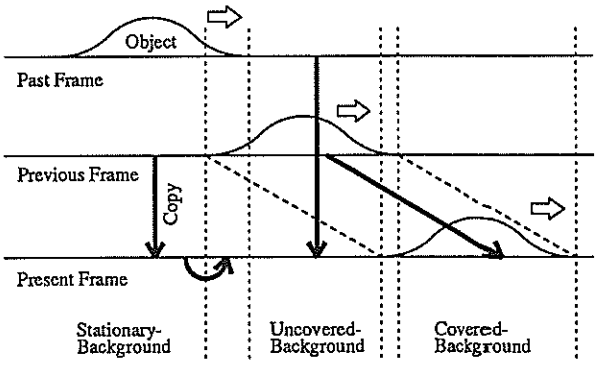
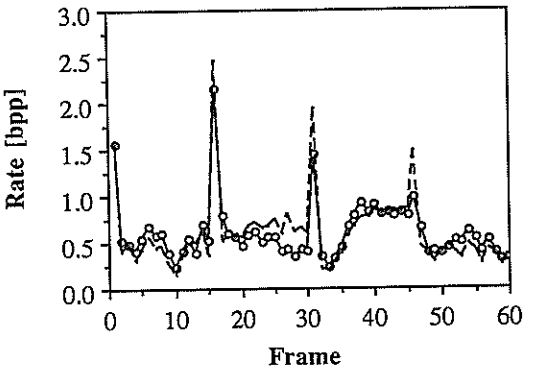
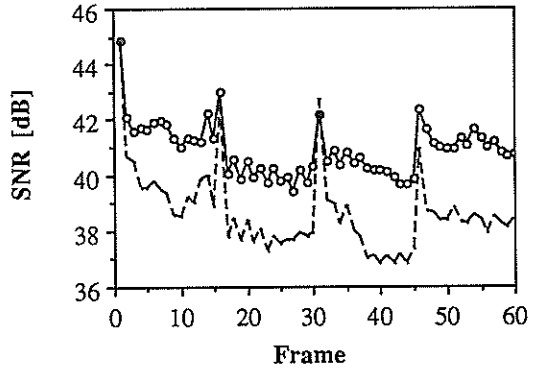


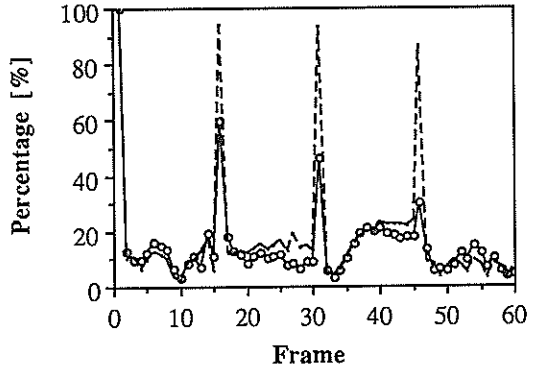
Figure 4 - Image model and copying operation



(a) Bit rate



(b) SNR



(c) Percentage of residue-encoded subblocks

Figure 5 - Coding performance

—o— PMIC
---- Conventional scheme

(4) The maximum number of available far-search-areas is fixed at 64.
(5) We employ the conventional orthogonal transform intraframe coder [11] for residue coding, which coder uses extrapolative prediction and the discrete sine transform.
(6) We employ the mean absolute error as the measure of the waveform distortion to alleviate computational complexity, and accordingly fix the value of the parameter T in Eq.(1) at 0.3.

Figure 5 shows a coding performance of the PMIC system compared with that of the conventional scheme employing the block-matching displacement estimation[7] where a displacement vector is determined in the search area of 14 x 14 by the full search. Figure 5 demonstrates that the PMIC system outperforms the conventional interframe coder and that it is effective for compression of image sequences which contain scene changes and moving objects with large displacement.

4. CONCLUSIONS

We formerly developed the concept of the universal pattern-matching coding(PMC) for compression of a still image[2]. By extending the concept of PMC to low-rate video compression, we develop a new interframe pattern-matching coder, called PMIC. Computer simulations demonstrate that PMIC is useful and potential as a technique for encoding an image sequence at a low bit rate.

REFERENCES

[1] N.M.Nasrabadi and R.A.King, "Image coding using vector quantization", IEEE Trans. Commun., Vol. COM-36, pp.957-971, Aug. 1988.
[2] T.Saito et.al., "Self-Organizing pattern-matching coding for picture signals", Proc. IEEE Int. Conf. Acoust., Speech & Signal Process., pp. 1671-1674, May 1989.
[3] W.K.Pratt et.al., "Combined symbol matching facsimile data compression system", Proc. IEEE, vol.68, pp.786-796, July 1980.
[4] O.Johnsen and J.Segan, "A pattern matching technique for facsimile coding", Proc. IEEE Int. Conf. Commun., pp.2G.2.1-2G.2.7, June 1982.
[5] A. Lempel and J.Ziv, "On the complexity of finite sequences", IEEE Trans. Inform. Theory, vol. IT-22, pp.75-81, Jan. 1976.
[6] J.Ziv and A.Lempel, "A universal algorithm for sequential data compression", IEEE Trans. Inform. Theory, vol. IT-23, pp.337-412, May 1977.
[7] T.Koga et.al., "Motion-compensated interframe coding for video conferencing", Proc. Nat. Telecom. Conf. pp.G5.3.1-G5.3.5, Dec.1981.
[8] D.E.Knuth, "The art of computer programming, vol.1, Sorting and searching", Addision-Wesley Reading, Mass., 1973.
[9] A.Furukawa et.al., "Motion-adaptive interpolation for video conference pictures", Proc. Int. Conf. Commun., pp.707-710, May 1984.
[10] D.E.Knuth, "Dynamic Huffman Coding", J.Algorithm, vol.6, pp.163-180, 1985.
[11] N.Yamane et.al., "An image data compression method using extrapolative prediction-discrete sine transform; in the case of two-dimensional coding", Trans. IEICE Japan, vol.J71-B, pp.717-724, June 1988.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

MULTIPLE RESOLUTION PROGRESSIVE VECTOR QUANTIZATION FOR IMAGE SEQUENCES

Fabio LAVAGETTO, Sandro ZAPPATORE

DIST - Universita' di Genova, Via Opera 11a, 16145 Genova, Italy

This paper presents the basic characteristics of a coding scheme that employs the Vector Quantization technique to allow a high performace code-book vector update. The Multiple Resolution Progessive Vector Quantization (MRP-VQ) strategy consists of different coding sessions: the first one is devoted to produce a low resolution vector code-book, the others use the error sequences to generate code-books which add local detail to the reconstructed images. Therefore, the resulting code is composed by different resolution code segments and by a code-book update segment. The size of the segments depends on source activity, channel capacity and required reconstruction performances. Some preliminary experimental results are also presented illustrating the performances of the proposed algorithm.

1. INTRODUCTION

The basic target in image sequences coding is exploiting the large amount of temporal correlation along the sequence in order to reduce the bitrate, while preserving an acceptable reconstruction quality at the receiver.

Motion compensation and conditional replenishment approaches have by far proven to supply robust and efficient coding schemes. These techniques do not exploit, however, a priori knowledge about the intensity statistics characterizing each specific video source. A Vector Quantization approach, on the contrary, fundamentally relies on the hypothesis of a strong stationarity in the images statistics, so that through a fairly long observation of the video source it is possible to build up a representative codebook of vectors capable of adequately modeling the source itself.

VQ coding schemes have been traditionally employed in static image coding applications, where the learning phase is conducted on training images belonging to the same class. VQ approaches to sequence coding have recently been reported showing promising experimental results [1],[2]. In these coding schemes the critical problem to face is the frame adaptivity of the codebook which has to cope with temporal image statistics variations. The only way to solve this problem is supplying a very large reconstruction codebook or periodically updating this latter. In the former approach it is necessary to carry the codebook building phase on a quite long observation interval in order to have a meaningful experience of the source dynamics. Furthermore it is necessary to constrain the codebook not to overcome a given dimension in order to bound the resulting bitrate or, conversely, to give the receiver the capability to switch among a set of smaller codebooks. In the latter approach an updating mechanism must be devised capable to temporally track the statistics variations in the incoming images and periodically replenish the reconstruction codebook.

Care must be payed to the updating performance in terms of codebook replenishment overhead and reconstruction quality.

The coding scheme here presented employes Vector Quantization technique in a fashion such to allow a high performance codebook vector updating mechanism. The Multiple Resolution Progressive Vector Quantization (MRP-VQ) algorithm basically consists of a first low resolution coding session followed by higher resolution sessions, which add local detail to the reconstructed images. The low resolution coding session is encharged to assure a minimal reconstruction quality at the receiver with a fairly low bitrate. The high resolution coding sessions increase the global bitrate by delivering detail information onto the channel.

The possibility to control the bitrate on the basis of the reconstruction quality, enables an efficient codebook updating policy. Keeping the bitrate fixed, it is possible to insert into the code stream the necessary codebook updating bits in place of some high detail bits, without lowering too much the reconstruction performances. Code-bits allocation policy and codebook updating policy are dynamically supervised by an overall coding manager which is in charge of maintaining a fixed channel bitrate while assuring acceptable reconstruction quality.

2. CODING SCHEME

The coding scheme is organized in a four layer architecture, where each layer is encharged to vector quantize a certain part of the video information.

The first layer, the basic one, processes the input video data and produces a high priority code which is always sent onto the channel. This code is also used, at the transmitter end, to compute an error sequence obtained as the difference between the original frames and the corresponding reconstructed ones.

The second layer works on the error sequence at a higher spatial resolution and produces a code which carries additive information. This last code, jointly with the first one yielded by the basic layer, is used to obtain a new error sequence which is eventually processed by the following layers. The code produced by these last layers is given a lower priority and is sent onto the channel depending on the code-bit allocation policy.

The resulting code can therefore be seen as composed of different code segments: the basic layer segment, the high layer segments and the codebook updating segment. The size of each segment is a critical issue, whose choice fundamentally depends on the video source activity, on the channel bandwidth and on the required reconstruction performances.

The basic layer segment, which is always present and complete, must be sized in order to guarantee a certain reconstruction quality also in case of burstiness peaks in the source dynamics. On this event the bandwidth will be typically entirely occupied by the basic layer segment and the codebook updating segment. The high layer segments, which add local detail, are inserted to fill the bandwidth left and their size depends on the code-bit allocation policy performed by the coding manager.

Being the encoding scheme based on vector quantization, the size of the code segments is an explicit function of the size of the various codebooks. Another critical element is the choice of the threshold values which determine when high layers are to be invoked and to which extent they must be used.

3. CODEBOOK CONSTRUCTION MECHANISM

The scheme which has been used to build up the various codebooks is a binary tree structure characterized by a self-adaptation mechanism which overrules its growing process to fit the statistics of the signal to code [3]. The codebook for the basic layer is built in order to assure a minimal level of the reconstruction quality at a rather low bitrate. The bidimensional vector support must be sized in order to keep the bitrate low and not to loose too much spatial correlation. The codebook size must be sufficiently large for reproduction fidelity. Experimental tests on sequences of 256x256 pixels images, 8 bit/pixel, have given best results for a vector support size of 16x16 pixels. A reference performance figure of 20 dB of SNR has been employed to size the codebook, which is therefore uniquely dependent on the amount of sequence activity.

The codebooks for the optional layers have been computed using smaller and smaller vector supports and operating on the error sequence as explained in the previous section. Working at higher spatial resolution enables detail information to be progressively picked up and transmitted while partially reducing the unpleasant block effect of the reconstructed sequence. The vector sizes which have been used are respectively of 8x8, 4x4 and 2x2 pixels. Each codebook size has been chosen in order to produce an increment of 3 dB in the SNR, for each new added layer.

4. CODEBOOK UPDATING POLICY

During the training phase simple statistics on the dynamics of the representative sequence are computed such as the probability of block updating, the probability to stop at the 8x8 resolution step, to stop at the 4x4 step and at the 2x2 step. From these statistics a mean frame bitrate is derived as well as a measure of its burstiness to model the specific video source. These parameters fix the threshold values of the codebook updating algorithm.

During the coding phase each incoming image vector is routed through the various encoding trees to select the relative descriptors and while doing so they also update the descriptors themselves. When the dynamically updated descriptors differ from the original ones more than a given threshold, they are labeled as no more descriptive vectors and will be updated as soon as possible.

A topological variation of the tree structure can also be necessary, namely a pair of leaf nodes is cut off while another leaf node is conversely splitted. This is transparent to the reconstructor, which must only replenish some lookup table locations and correctly reassign the addresses. The codebook updating task is up to the coding manager, which has an overlook on the whole coding process status. Depending on the dynamics in the frame coding process and on the codebook updating urgency, the manager assembles the structure of the code stream according to the strategy presented in the following section.

Experimental tests have applied the updating mechanism only to the first two codebooks, the remaining two remain fixed as configured during the learning phase.

5. CODE-BIT ALLOCATION POLICY

Variable length code (VLC) has been employed to encode the data delivered by each layer of the coder. When a codebook is updated and the relative updating code is sent to the reconstruction end, the entropic code is recomputed and the VLC lookup table is accordingly replenished.

Let B be the available bandwidth and C1 the basic layer code, B-C1 bit/sec are left for further code.
Let S be the number of dB of the SNR obtained by means of the unique code C1.
Let U be the codebooks updating code:
if U < B-C1 all the updating code is sent, on the contrary only the most important updating data are sent to reach the bandwidth B.

Importance is represented by the SNR improvement introduced by the updating itself. The remaining (B-C1-U) bits are employed to send high layers code.
The basic layer code C1 performs a global refresh of the screen while the high layers code perform only a local refresh adding the necessary detail. The high layers code C2, C3 and C4, one for each encoding layer, refresh the screen through a circular scanning from the centre of it and under the control of the thresholds T1, T2 and T3. These thresholds are redifined frame by frame depending on the value (B-C1-U) which represents the amount of still available bits.

6. CONCLUSIONS

The coding scheme here presented has shown high adaptivity to variations in the image statistics while assuring a fixed minimal reconstruction quality.

Figs. 1-4 show some results achieved by the application of the MRP-VQ algorithm. The code-bit allocation policy and codebook updating policy which have been employed are still preliminary and show promising possibility to improve the overall performance of the coding scheme. Applications in Variable Bit Rate coding for Packet Switched Networks are foreseen [5],[6],[7],[8],[9].

REFERENCES

[1] M.Goldberg, H.Sun, "Image Sequence Coding Using Vector Quantization", IEEE Trans. on Com., 34, n.7, 1986, pp. 703-10.

[2] M.Goldberg, H.Sun, "Frame Adaptive Vector Quantization for Image Sequence Coding", IEEE Trans. on Com., 36, n.5, 1988, pp. 629-35.

[3] F.Lavagetto, S.Zappatore, "An Unbalanced Tree Structure for Frame Adaptive Vector Quantization of Image Sequences", SPIE/SPSE Symp. on Electr. Imaging, S.Clara Cal.,Feb.,11-16 1990.

[4] F.Lavagetto, S.Zappatore, "Variable Bit Rate Progressive Vector Quantization for Packet Switched Networks", Int.Workshop on Packet Video, Morristown N.J.,March 22-23 1990 (Submitted paper).

[5] S.Dixit, Y.Feng, "Adaptive Vector Quantization of Video for Packet Switched Networks", Proc. ICASSP '89, pp. 1870-73.

[6] H.Jozawa, H.Tominaga, "An Adaptive Coding Procedure for Packetized Video Images including Still and Moving Pictures", Int.Workshop on Packet Video, Torino, Sept. 8-9, 1988.

[7] N.Dal Degan, E.Franzini, E.Marcozzi, "Performance Analysis of a Variable Rate Video Coder for Packet Networks", Int.Workshop on Packet Video, Torino, Sept. 8-9, 1988.

[8] W.Verbiest, L.Pinnoo, B.Voeten, "The Impact of the ATM Concept on Video Coding", IEEE J. on Sel. Areas in Com., 1988, pp. 1623-311.

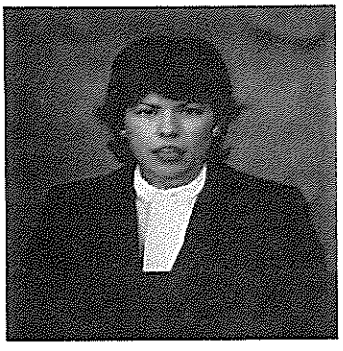


Fig. 1 Original frame from sequence Claire

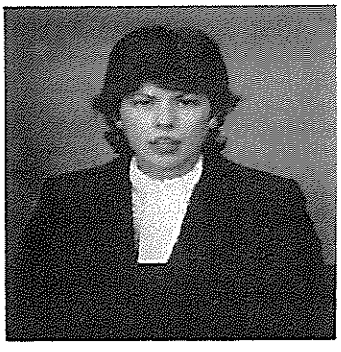


Fig. 2 Coded frame by MRP-VQ at 0.3 bpp

Fig. 3 Original frame.

Fig. 4 Coded frame by MRP-VQ at 0.5 bpp.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

VECTOR QUANTIZATION IN IMAGE SEQUENCE CODING

J. Huguet, L. Torres

Dept. of Signal Theory and Communications
E.T.S.I. Telecomunicació - UPC
Apdo 30.002 08080 BARCELONA, SPAIN

Abstract

A 3D Vector Quantization scheme and the subsequent results obtained in image sequence coding are presented as an efficient technique that exploits the correlation in space and time. A comparison between 2D and 3D vector quantizer performance in image sequence coding is given.

I. Introduction

The goal of a good coding technique is to reduce the information as much as possible, in order to be transmitted or stored efficiently. This is accomplished by removing the information redundancy contained in the original signal and sending or storing only the most significant parts of the image information. Among the great number of available techniques to extract the relevant part of the information signal, vector quantization (VQ) has proved to be very effective. Vector Quantization was used first with speech signal (1D-signal) and the good results results observed led to extend its application to image signal (2D-signal). In a further stage it was used with image sequence signal (3D-signal). A brief revision of these advances in vector quantization and some comparative results are treated here.

II. Vector Quantization

In contrast to scalar quantization, vector quantization performs the quantization not treating each sample individually but previously grouping a number of consecutive samples forming a vector and then quantizing this vector. Let X be an input vector from the original signal and let Y represent a codeword or reproduction vector. The codebook contains all the available codewords and the quantizer replaces the input vector by the codeword that is more similar to X. Using the euclidean distance as a distortion measure, the quantizer searches for the codeword located at the minimum distance from the input vector. The quantization process is usually termed Y=Q(X). In a mathematical sense, the quantizers establishes a correspondance between a N-dimensional space V to a finite codebook W= [Y_1, . . . , Y_M]

This Paper has been supported by PRONTIC 105/88.

$$Q: V \rightarrow W$$

A vector quantizer is completely described by the codebook and the Voronoi regions R=[R_1, . . . , R_M].
The partition of the input space V into the Voronoi regions depend on both the codebook and the distortion measure used.

III. Vector Quantization in speech and image coding.

The LBG algorithm [1] has been widely used in speech and image coding to generate the codebook. In speech coding the unidimensional signal is divided into vectors which remain as unbreakable units during the coding process, while in image quantization the two dimensional signal is divided into rectangular blocks that will be the units in the coding process (usually known as 2D-VQ) as shown in Fig. 1.

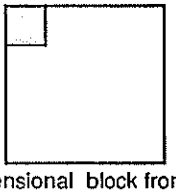


Figure 1. Two dimensional block from a single image in a 2D vector quantizer.

The size of the block should be large enough to group spatially correlated pixels but sufficiently small to avoid the so called "blocky effect". To achieve low coding rates, large blocks are needed. As it will be shown, a 3D vector quantizer allows to group a larger number of pixels than a 2D scheme thus avoiding the "blocky effect" that a 2D vector quantizer shows when grouping the same number of pixels from the same image.

IV. Vector Quantization in Image Sequence coding

The first intuitive way to use vector quantization in image sequence coding is to use a 2D vector quantizer, where each image of the sequence is quantized individually and independently of previous or future images of the sequences.

A closer look at the problem reveals that only the spatial correlation is exploited in a 2D vector quantizer. In case we want to exploit the spatial and temporal correlation contained in an image sequence, a 3D vector quantizer should be used. The image sequence, considered as a 3D signal, is divided into 3D cubes that will be the quantization units. Each cube contains the pixels within a small region in the image (spatially correlated) and the pixels of the next images of the sequence (time correlated). Thus it verifies that all pixels grouped within a cube are correlated.

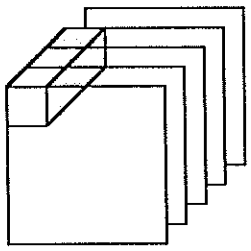


Figure 2. Three dimensional cube from an image sequence in a 3D vector quantizer.

In order to use the LBG algorithm, we need to introduce a new measure of distortion between cubes. A simple square error has been chosen, calculated as follows for a MxNxK cube size

$$d(X,Y)= \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{k=0}^{K-1} (x_{mnk} - y_{mnk})^2$$

where X is a cube of the image sequence and x_{mnk} a pixel within the cube. Similarly, Y represents a codeword cube of the codebook and y_{mnk} a pixel of that codeword. The distortion per pixel that will be used as a comparison measure between quantizers is computed as

$$\text{distortion per pixel} = d(X,Y) / (MxNxK)$$

In the coding process, a full search is carried out for each cube implying the calculation of the distance to all the codewords. Full search is highly processor consuming when the codebook is large, but gives the best (optimum) results in the square error sense. In order to reduce the computational cost of the full search algorithm the fast codebook search algorithm described in [2] has been used, reducing the encoding complexity to less than the half when using a square distortion measure. This algorithm discards a bad codeword with very few operations without calculating the euclidean distance which requires N multiplications and 2N-1 additions / substractions. A property of this algorithm is that it never discards a codeword whose distance is less than that of the best codeword tested so far. The final result is then the same as if all euclidean distances had been calculated and compared.

On the other hand, a tree search algorithm or a multistage structure [3] can be used to accelerate the encoding process at the cost of some quality degradation and additional complexity. In fact, many of the more sophisticated schemes used in 1D and 2D vector quantizer can also be applied to a 3D-VQ scheme because the underlying idea of the algorithms remains the same.

V Results

We have used both a 3D vector quantizer to code an image sequence named "ma" (that contains 24 images) for three different cube sizes and a fixed number of codewords chosen to be 1024, using the full search algorithm. The images were 256x256 pixels and each pixel ranged from 0 to 255 (gray level). The results obtained with a 3D and 2D vector quantizer are presented in tables I and II, while Fig. 3 contains a quantized image of each example (we just show one image of the sequence as it is impossible to include the whole sequence, and even that would not help to completely figure out the quality of image sequence seen in real time). Thus for a 4x4x2 pixel cube size and 4096 codewords gives a 0.375 bits/pixel with a mean distortion per pixel of 9. The same bit rate in a 2D vector quantizer for a 4x4 pixels block size (the same spatial size) limits the codebook size to 64 codewords and yields a distortion of 30. Notice that the first case of table I and II are in fact the same and so are their results. The codebook was generated in all cases from the same long sequence.

The (S/N) of tables I and II is taken as the peak signal to noise ratio, computed as

$$(S/N) = 10 \log(255^2/\text{dist. per pixel})$$

The images of Fig. 3 clearly reveal that the quality of the image sequences quantized with the 3D vector quantizer are significanly better than those quantized with the 2D vector quantizer for the same bit rate. In the 2D case, the effect of reducing the bit rate drastically reduces the codebook size, with a great influence on the final quality of the images.

VI Conclusions

These results confirm that a 3D-VQ scheme gives better quality images than a 2D-VQ, specially when low bit rates are desired. In other words, additional redundancy is removed when grouping pixels in three dimensional blocks. It is obvious that a 2D vector quantizer is a particular case of a 3D vector quantizer (the first case of table I shows that). On the other hand, as the codebook is larger the computational cost required in the full search encoding process is also greater. That is why a great number of fast encoding algorithms for vector quantization have grown up in the last years.

References

[1] Linde, Y., Buzo, A., and Gray, R.M.
"An algorithm for vector quantizer Design"
IEEE Transactions on Communication.
COM-28, pp. 84-95. January 1985.

[2] J. Huguet and L. Torres.
"Fast codebook search algorithm in vector quantization". Submitted for publication to Vision Communications and Image Processing'90.
Lausanne. 2-4 October 1990

[3] R.M. Gray
"Vector Quantization"
IEEE ASSP Magazine, Vol 1, April 1984.

[4] M. Goldberg and H. Sun
"Image sequence using vector quantization".
IEEE Trans. Comm. July 1986

Cube Size	Number of codewords	Bit Rate (bits/pixel)	Distortion per pixel	(S/N) dB
4x4x1	4096	0.75	8	39.10
4x4x2	4096	0.375	9	38.58
4x4x4	4096	0.18	10	38.13

Table 1. Results obtained with a 3D Vector Quantizer with the image sequence "ma".

Cube Size	Number of codewords	Bit Rate (bits/pixel)	Distortion per pixel	(S/N) dB
4x4	4096	0.75	8	39.10
4x4	64	0.375	30	33.35
4x4	8	0.18	67	29.87

Table 1. Results obtained with a 2D Vector Quantizer with the image sequence "ma" for the same bit-rate of table I (the number of codewords is modified accordingly).

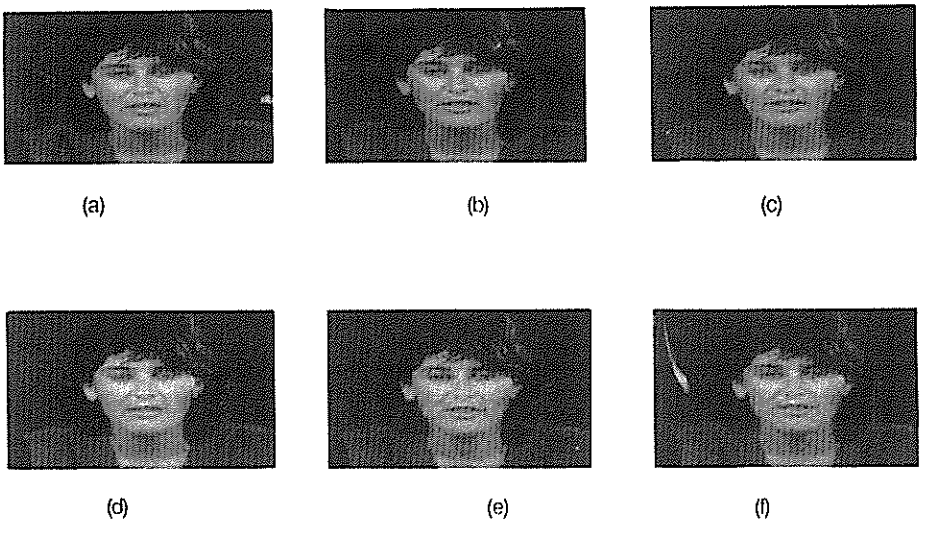


Figure 3. Single image of the sequence quantized with different block (or cube) size when using a 2D (or 3D) vector quantizer scheme. (a) Original image of the sequence. (b) Cube of 4x4x1 pixels. (c) Cube of 4x4x2 pixels. (d) Cube of 4x4x4 pixels. (e) Block of 4x4 pixels and 64 codewords. (f) Block of 4x4 and 8 codewords.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

AN ADAPTIVE APPROACH TO COLOR-PICTURE CODING

Fabio Arduini, Daniele D. Giusto, and Gianni Vernazza

Department of Biophysical and Electronic Engineering
University of Genoa, Via Opera Pia 11A, I-16145 Genoa, Italy

This paper is aimed at presenting and discussing some results obtained by applying an adaptive coding system to color pictures. The system integrates into a hybrid approach two well-known (and significantly improved) techniques (i.e., vector quantization and polynomial approximation) by analyzing edge and texture information.

1. INTRODUCTION

The aim of this work is to present an innovative coding system to be applied to color pictures. Such a system is able to compress data in a flexible way, considering that the analysis of an image usually points out two important aspects. More precisely, some image areas (i.e., those with homogeneous grey levels) appear uniform, hence such areas are of minor importance for the human user. Instead, the areas that contain the main characteristics of a scene (i.e., edges and textures) are of major importance because the human eye focuses naturally on them. These considerations have led to the development of a system able to operate in an adaptive way on the scene examined. In other words, the system contains a pre-coding module devoted to the generation of an edge-and-texture map, used by the coder during the integration of different techniques, as described in the following. It should be noted that this system is the modified version of a previously developed one [1], in which the coding step is driven by an understanding system that locates the various areas present in a scene, and associates some degree of importance with each of them. Then, on the basis of these different degrees, the areas are coded in different ways by applying vector quantization (to the most important regions) or polynomial approximation (to the least significant ones). However, this previous system is strongly dependent on the kind of scene to be coded, that is, one needs different databases and specific decision rules for the different cases considered. So, to obtain a system able to operate on every kind of picture, we have implemented this new version of the original system: it merges vector quantization and polynomial approximation into an adaptive coding scheme, without need for any understanding pre-task. In the following, the architecture of the system is described, as well as the edge-and-texture map and the two coding submodules. Finally, the results obtained on standard images are reported and discussed.

2. THE CODER

As previously mentioned, this module utilizes two different submodules in an adaptive way: one applies the vector quantization technique, and the other uses polynomial approximation.

2.1. The Vector Quantizer

Vector quantization is a transformation technique based on the conversion of a square block (of K pels) into a vector [2]. This vector is then quantized via a comparison with a set of standard vectors constituting the codebook. The codebook vector most similar to the examined vector is chosen as the coding vector. Because the codebook is well known, it suffices to transmit only one symbol (denoting the vector); this symbol is then decoded, and, in the reconstructed image, the block is represented by the coding vector. The high quality of the codebook depends on the training set used for its construction: a statistically valid training set will generate an efficient codebook.
As regards the present implementation of this coding technique, a quantizer is proposed that is characterized by three innovative aspects:
- use of a neural net (i.e., a Kohonen self-organizing feature map [3]) during the phase of the codebook generation;
- use of a merge-and-split operator able to furnish a more effective codebook;
- use of a predictor to exploit the correlations between neighbouring blocks in order to reduce, in a most drastic way, the compression ratio without loss in quality.

The starting point of the present work is the classical algorithm proposed by Linde, Buzo and Grey [4]. It is well-known that, given an

initial codebook, such a method produces only a local minimum for distortion. Many algorithms to overcome this drawback have been proposed in the literature (e.g., random choice from the training sequence, splitting, coarse scalar quantizer, etc.). The Linde-Buzo-Grey algorithm exhibits another drawback, namely, a considerable processing time for the codebook computation.
Application of a neural net can sharply reduce the overall processing time, as it provides a starting point that is already close to a suboptimal solution; moreover, a neural net ensures an improvement in the final solution, since the starting point can be obtained by a non-supervised classifier.
Once the codebook generation has been achieved by the Linde-Buzo-Grey technique (i.e., at the end of the iterations), if one examines the partitions associated with the vectors obtained, one can notice that a considerable number of vectors are not fully utilized, that is, they are associated with partitions of negligible size. A more uniform distribution of the training sequence in the codebook space would result in a notable improvement in the quality of the coded images, in that the set of available patterns would be able to better suit the vectors used most frequently in natural images, rather than suit the peculiarities of the images of the training sequence.
In order to obtain this uniform distribution, a merge-and-split algorithm for the codebook vectors has been developed, which can be called at each step of the Linde-Buzo-Grey algorithm for the twofold purpose of eliminating the vectors associated with partitions of negligible size (merging), and of replacing them with the vectors associated with the new smaller partitions obtained by subdividing (splitting) the ones of larger size, which approximate a larger number of vectors of the training sequence.
An additional improvement in the coder has been obtained by using an RGB predictor. The basic idea of the development of a predictor is provided by the address-vector-quantization technique [5]. Such a technique produces a further reduction in the redundancies of the quantized images by exploiting the residual correlations between neighbouring vectors.
During the learning phase (codebook generation), an address codebook is produced (i.e., a codebook where each vector is made up of the addresses of 4-connected vectors, in the Linde-Buzo-Grey codebook). In addition, four transition matrices are used, whose inputs contain the number of times a vector is immediately followed by another one in a certain direction, in the images of the training sequence.
After vector quantization, for each vector of the address codebook a score function is computed, which is based on the values contained in the aforesaid matrices and on the use of the values related to the neighboring blocks that appear first in a raster scanning. The address codebook is rearranged according to these values, and some vectors are included in the so-called active region. If the vector to be transmitted is contained in the active region, the whole set of four vectors will be coded by using only one address inside the active region; otherwise, it will be necessary to transmit the information about the single vectors. Therefore, the increase in the compression ratio depends on the probability of a correct prediction.
Performances can also be improved by considering the correlations among the different spectral bands; this fact leads to the use of particular transition matrices that refer to the correlations among blocks in different spectral bands.

2.2. Polynomial Approximation and Hybrid Coding

In this section, the coding submodule that implements the method of polynomial approximation [6] is described. This method belongs to the second-generation ones; it is based on image description by means of regions (i.e., it needs a segmentation step before the coding one). Two kinds of information are generated: one refers to regions, the other to the grey-level values inside each region. Because the segmenting module utilizes the split-and-merge algorithm, regions are defined as sets of squares of variable size: grey-level approximation inside them is performed using 2-D polynomial functions.
It is worth noting that the coding procedure is not restricted to signal approximation only, as it also includes the segmentation process. During the segmentation phase, one has to take into account the possible presence of edges or textural areas inside a region; in this case, it is mandatory to perform the splitting process because polynomials are not able to approximate wide regions characterized by the presence of these features.
It should be pointed out that, while an edge can be coded by polynomials by accepting a high fragmentation of the squares, textural areas remain an unsolved problem of this approach. Polynomials cannot follow the behaviour of a textured surface with acceptable accuracy. To overcome this drawback, the proposed solution is to integrate vector quantization into polynomial approximation; in other words, one uses a hybrid coding technique that automatically switches from one kind of approximation to the other when some particular conditions are fulfilled.
Coming back to the global architecture, it

should be noted that edge and texture information is provided by a separate module that furnishes a map of the regions in which the splitting process must be performed. In particular, edges are located by applying the Perona-Malik operator [7], able to furnish edges not affected by distortion. Instead, textural areas are found by applying a simple operator that analyzes the histogram of the image obtained by convolving the original one with a Laplacian kernel, and locates the areas that do not look uniform.

As said above, the split-and-merge process and the approximation are performed together, for the former is driven by the results of the latter. Obviously, split-and-merge segmentation is divided into two further steps. The first involves the partition of an image into squares (pyramidal approach), and information about the squares is organized into a quadtree. The squares are examined by the approximation module in order to decide if they have to be split again. To accomplish this task, for each square a 2-D polynomial function is computed that best fits the original data. After error minimization, the resulting error is compared with a threshold to assess the quality of the approximation. If this test yields unsatisfactory results (or if an edge or a textural area is still present) the splitting process is performed, followed by another approximation process for each square generated.

A very important problem, not completely solved, is the choice of the points to be used for error computation (obviously, for computation-time reasons, it is not possible to consider every point of a region). This choice is actually a pseudo-random one because such points are located among those near the characteristic ones (e.g., local maximum or minimum points).

When the splitting process has been performed, one has to perform the second step (i.e., the merging process) to obtain a simpler scene description and reduce the tessellation effects due to the splitting process. The merging module is the most problematic of the whole system because it has to manage a very complex data structure. In short, the quadtree is integrated with a list of links between contiguous regions, to be considered in the merging phase. Each region is pointed by several links, thus allowing an easy updating of the structure whenever a merging operation is performed. Clearly, in this way, the original quadtree is strongly modified by these links, and results in a very complex graph. The merging phase computes the fusion cost (i.e., the increase in distortion) for each link; then, after choosing the link with the lowest cost, operates the fusion between the two related regions.

At the end of the merging process, the whole scene is described by a graph and by the coefficients of the polynomials associated with each region. However, considering that regions are bounded by very particular edges (i.e., they are rectilinear and oriented in two directions only, due to the segmentation approach followed), it is more advisable to give up this kind of description and to code regions according to their edges. In this way, by using a very simple contour-coding technique, one can reduce significantly the amount of information required. In the decoding phase, starting from the contour maps, regions are recovered by applying a filling algorithm; then, by using the coefficients of the 2-D polynomials, or the quantized vectors for edges or textural regions, the scene surface can be reconstructed to a good approximation.

3. RESULTS

In this section, we present the results of the application of this approach to the coding of two classical images, that is, the picture of Lena and peppers. The original images are shown in Fig.1: they are 512*512 pels in resolution, and 24 (8 bits per plane) bits in intensity. In Fig.2, the edge-and-texture map is displayed, together with the split one, for the peppers image. Finally, in Fig.3 the results on both images are presented. For Lena's image, the obtained compression ratio is about 60:1, associated with an SNR of about 29 dB. For the peppers image, we reached a compression factor of about 90:1, together with an SNR of about 26 dB. The codebook used was 256 vectors in size, and each vector was 4*4 pels in dimensions; the probability of a correct prediction was over 90 percent. Concerning to the polynomials, we used 2nd-order functions, so each region was represented by means of 6 coefficients.

4. CONCLUSIONS

A new hybrid technique that integrates vector quantization into a polynomial approximation scheme has been presented. The integration can be performed in an adaptive way by analyzing edges and textural areas. In such areas, polynomial approximation does not reach sufficient accuracy, so they are coded by applying vector quantization. The vector quantization module is characterized by some improvements that have been obtained by using a neural net and a merge-and-split operator in the codebook-generation phase, and by developing a multispectral spatial predictor. The proposed approach allows good performances, as shown by some results obtained on standard images.

5. REFERENCES

[1] Giusto et al., "A new adaptive approach to picture coding", Annales des Telecommunications (to be published).

[2] Gray, "Vector Quantization", IEEE-ASSP Magazine, Vol. 1, No. 2, pp. 4-29, 1984.

[3] Kohonen, Self-Organization and Associative Memory, Springer Verlag, Berlin, 1984.

[4] Linde et al., "An Algorithm for Vector Quantizer Design", IEEE Trans. on Communications, Vol. 28, No. 1, pp. 84-95, 1980.

[5] Feng and Nasrabadi, "Address-Vector Quantization: An Adaptive Vector Quantization Scheme Using Interblock Correlation", Proc. SPIE, Vol. 1001, pp. 214-222, 1988.

[6] Kunt et al., "Recent Results in High-Compression Image Coding", IEEE Trans. on Circuits and Systems, Vol. 34, No. 11, pp. 1306-1336, 1987.

[7] Perona and Malik, "Scale-Space and Edge Detection Using Anisotropic Diffusion", Univ. of Berkeley, Tech. Rep. UCB/CSD 88/483, Dec. 1988.

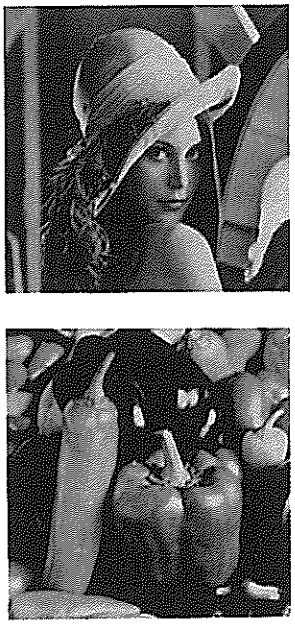


Fig.1. The two original images (Lena and peppers).

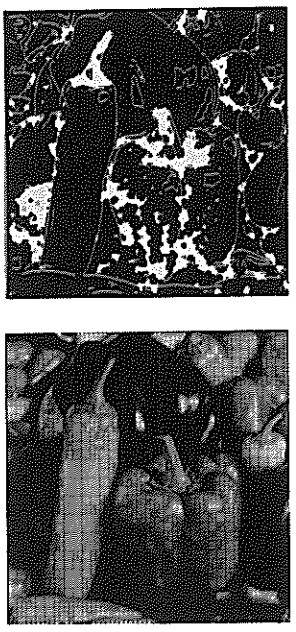


Fig.2. The edge-and-texture and split maps for the peppers image.

Fig.3. The final decoded images: Lena (compression factor: 60:1, SNR: 29 dB) and peppers (compression factor: 90:1, SNR: 26 dB).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

PARALLEL ADAPTIVE MULTISTAGE VECTOR QUANTIZATION FOR DIGITAL VIDEO COMPRESSION

Javier R.Fonollosa, José A. Rodríguez-Fonollosa

Departament de Teoria del Senyal i Comunicacions
Universitat Politècnica de Catalunya
Apdo. 30.002, 08080 Barcelona, Spain.

A new video data compression algorithm is proposed in this communication. It is based on an adaptive vector quatization scheme for DCT blocks of color images. The adaptation algorithm is very simple and does not require additional information to be sent to the decoder. The multistage structure allows reasonable memory requirements both in the coder and decoder.

1. INTRODUCTION

When transmitting or storing image sequences it would be desirable to send or store the minimum amount of information necessary to reconstruct the sequence to a given level of picture quality. In this paper different methods, all comprising Vector Quantization and Discrete Cosine Transform, are proposed, which seem to be useful in this endeavor.

2. VECTOR QUANTIZATION (VQ)

2.1. General description

Quantization, the process of approximating continuous-amplitude signals by digital (discrete-amplitude) signals, is an important aspect of data compression or coding. The independent quantization of each signal value or parameter is known as scalar quantization, while the join quantization of a block of parameters is termed vector quantization. This is a process of redundancy removal that makes effective use of four interrelated properties of vector parameters: linear dependency (correlation), non linear dependency, shape of the probability density function (pdf), and vector dimensionality itself. In contrast, scalar quantization can utilize effectively only linear dependency and pdf shape [1].

2.2. Multi-Stage Vector Quantization (MSVQ)

Vector Quantization can offer substantial performance advantages over scalar quantization. It is known to be asymptotically optimal according to the Rate Distortion Theory [2], but unfortunately at considerable computational and storage costs. These requirements are exponential in the number of bits per vector.

For instance, when 8•8 blocks are to be quantized, 64 dimension vectors form the codebook. Coding to 0.5 bits per pixel, which can not be regarded as excessive to most of the applications, means 32 bits per block and consequently 2^{32} reconstruction vectors (more than a thousand Gigabytes of memory) which is evidently impractical.

In that case a Multi-Stage scheme is suitable [3]. Multi-stage has always been seen as a suboptimal VQ scheme with reduced complexity and storage. It consist in successively approximating the input vector in several cascade VQ stages, where the input vector for each stage is the quantization error from the preceding stage.

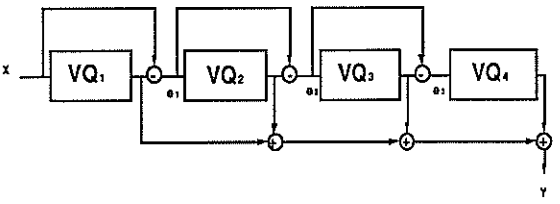


Figure 1. Multi-Stage VQ scheme with 4 stages

Using 4 stages of 8 bits each (fig. 1) will reduce complexity and storage, requiring 4•256=1024 codewords (half Megabyte).

2.3. Adaptive Multi-Stage Vector Quantization (AMSVQ)

The multi stage structure can be used to develop a continuously adaptive VQ system [4]. The main idea of the adaptation algorithm is to update the preceding codebooks taking into account the information given by the rest of the quantizers. In the most simple scheme, with two stages, a vector **x** is quantized to give a quantized vector **y**

$$y = c + e_q$$

where c is the output of the first quantizer and e_q the contribution of the second codebook (fig.2).

$$c = q_1(x)$$

$$e_q = q_2(x - c)$$

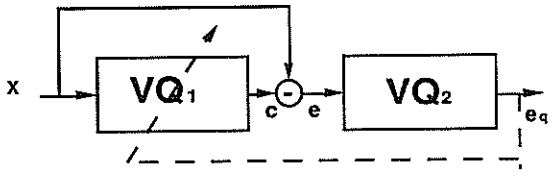


Figure 2. Adaptive Multi-Stage VQ scheme with 2 stages

Then e_q, that is an estimation of the quantization error of the first codebook, can be used to adapt the quantizer in the following way:

$$c = c + \mu\, e_q$$

where μ is the adaptation factor. This adaptation is made without the need to increase the bit rate, and is computationally very simple. Several values for μ are proposed in [5] in the range from 0.1 to 1.

3. DISCRETE COSINE TRANSFORM (DCT)

3.1. Transform coding techniques

In the codification of speech and specially video signals one of the most used technique is the Transform Coding. This kind of technique benefits from linear dependencies for efficient coding. The quantization process implied in all coding schemes is carried out, not directly in the samples, but in a biunivocal linear transformation of them.

It can be proved [2] that quantization of transformed coefficients can reduce the quantization error, even in non optimal schemes as DCT, depending on the correlation of the samples of the source.

3.2. Two-Dimensional DCT of images

Block DCT of images has been widely used in video coding schemes. It consist in dividing every frame into square sample blocks, which are DCT transformed into another block of coefficients. These coefficients define the two dimensional vector to be quantized.

3.3. Three-Dimensional extension for color images

The Discrete Cosine transform can be easily extended to color images. If the image is then defined by the luminance and two color difference frames, S(i,j,k) the Three-dimensional DCT, T(u,v,w) can be:

$$T(u,v,w) = \frac{2}{N}\beta(u)\beta(v)\beta(w)\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} S(i,j,k)\, cf(u,i,I)cf(v,j,J)cf(w,k,K)$$

where

$$\beta(0) = \frac{1}{\sqrt{2}}\ ;\ \beta(z) = 1,\ z \neq 0$$

and

$$cf(u,i,I) = \cos\frac{\pi u(2i+1)}{2I}$$

$$cf(v,j,J) = \cos\frac{\pi v(2j+1)}{2J}$$

$$cf(w,k,K) = \cos\frac{\pi w(2k+1)}{2K}$$

The 3D-DCT can benefit from of the high correlation between the luminance and the color signals.

4. CODING SCHEMES

4.1. Parallel AMSVQ of DCT blocks (PAMSVQ/DCT)

The AMSVQ method had proved to be useful coding LPC parameters of speech at very low bit rates, 1350 bps [4] and therefore good results were expected in video coding. Images are divided into blocks which are coded sequentially by the algorithm. The selected blocks are 8•8 pixels resulting in a 64 dimension vectors. Several different configurations have been tested in the AMSVQ for this scheme. The introduction of DCT before the quantization of the blocks (fig. 3) can improve performances of the algorithm without adding a great overload.

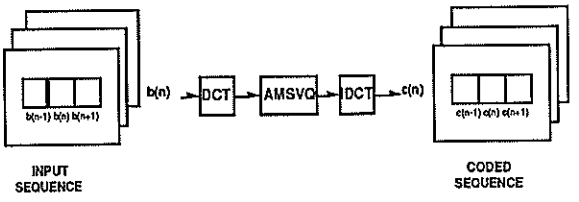


Figure 3.AMSVQ/DCT scheme applied to a video sequence

The memory requirements for the first stage codebook when it only uses 3 bits is of $2^3 = 8$ blocks of 8•8 coefficients. Therefore if each block of the image had its own adaptive first stage codebook that would only require a storage equivalent to 8 images of coefficients. That is the idea of the PAMSVQ/DCT (fig. 4). Giving one first stage adaptive codebook to each block of the image, the adaptation scheme works independently in each block and much better results should be achieved.

This scheme has been tested with 3 and 5 stages. The first stage is always adaptive and it uses 3 bits while the other stages use 8 bits and they are non-adaptive.

4.2. Parallel AMSVQ of DCT blocks with Backward or Forward Power Estimation

The training sequence used in the codebook generation in any VQ scheme has always been considered critical to the final coding performances. All the schemes presented here are adaptive in its first stage, and therefore less sensible to the training sequence than non adaptive ones. Nevertheless, the non adaptive stages can improve its performances if its input is normalized.

This idea is not new, in [6] a different approach is presented with one vector quantizer used with the normalized DCT coefficients and one scalar quantizer with the gain.

Therefore, one power estimation block should be used. The decision is then to choose one forward power estimation with associated increase in the bit rate or backward estimation (fig. 5), with no increase in the bit rate but poor accuracy.

Both schemes have been tested with 3 and 5 stages. As before, the first stage uses 3 bits and it is adaptive while the others use 8 bits and they are non-adaptive.

5. RESULTS

The above mentioned schemes have been tested with several still images and video sequences. The parallel extension of the algorithm has proven to improve the coding performances specially when applied to slow moving images.

As an example of this improvement one simulation result is given for the Backward Prediction Error PAMSVQ coding one fixed image. Three stages have been used, the first adaptive with 3 bits (8 codewords) and the second and third non adaptive with 8 bits (256 codewords). The first iteration of the algorithm is shown in figure 6. The reconstruction errors are clearly visible owing to the small number of codewords of the first stage. Once the adaptive algorithm has converged, after 10 iterations in this case, the coding errors are dramatically reduced as shown in figure 7. Similar results were obtained for other images.

The scheme converges provided the error prediction is accurate enough to improve the codewords of the first stage via the adaptation mechanism. Two stages of 8 bits each have shown to fullfil these requirements when the backward prediction power normalization structure is used.

6. CONCLUSIONS

AMSVQ/DCT is a new hybrid coding method for image and video compression. It combines Transform coding techniques with vector quantization. In addition it is automatically and continually adapted to the input signal with no increase in the bit rate.

The parallel extension of this method can achieve very good performances at compression rates in the order of 0.3 bits per sample with reasonable memory requirements (equivalent to 8 images of DCT coefficients). The bit rate is kept constant but the quality is continuously improved in the blocks with slow movement.

REFERENCES

[1] John Makhoul, Salim Roucos, Herbert Gish, "Vector Quantization in Speech Coding". Proceedings of the IEEE, vol 73, NO 11, November.

[2] N.S. Jayant and Peter Noll, "Digital Coding of Waveforms. Principles and Applications to Speech and Video", Prentice-Hall Signal Processing Series. Prentice-Hall, INC. Englewood Cliffs, New Jersey 07632.

[3] B-H Juang, A.H.Gray, Jr. "Multiple Stage Vector Quantization for speech coding". IEEE Proc. ICASSP, Paris, May 1982.

[4] José A. Rodríguez Fonollosa and Enrique Masgrau., "Adaptive Multi - Stage Vector Quantization", IEEE Proc. MELECON. Lisboa April.

[5] José A. Rodríguez Fonollosa, "Adaptive Vector Quantization applied to Speech Coding", Doctoral Thesis. Polytechnic University of Catalonia. Spain.

[6] Hai-Shan Wu, Hong-Bin Chen, "A DCT Gain-Shape Vector Quantizer for Image Coding", ICASSP 88, New York.

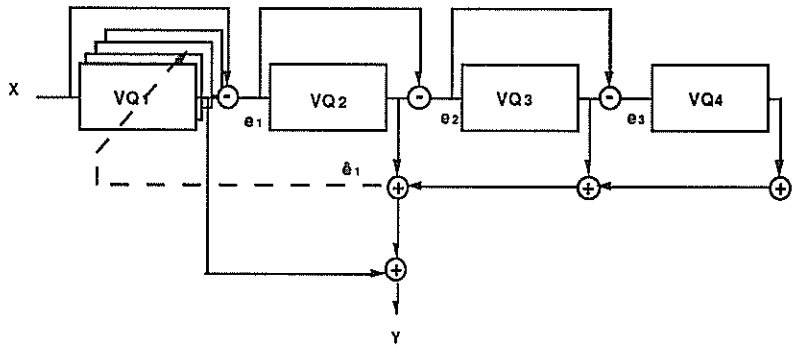


Figure 4. Parallel AMSVQ scheme

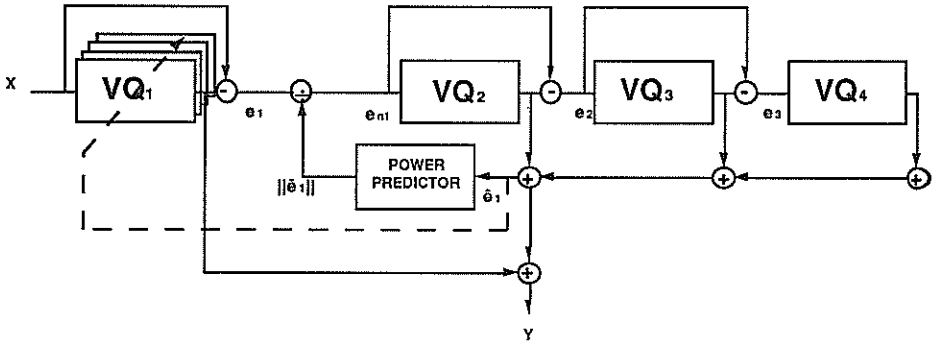


Figure 5. Parallel AMSVQ scheme with backward power estimation.

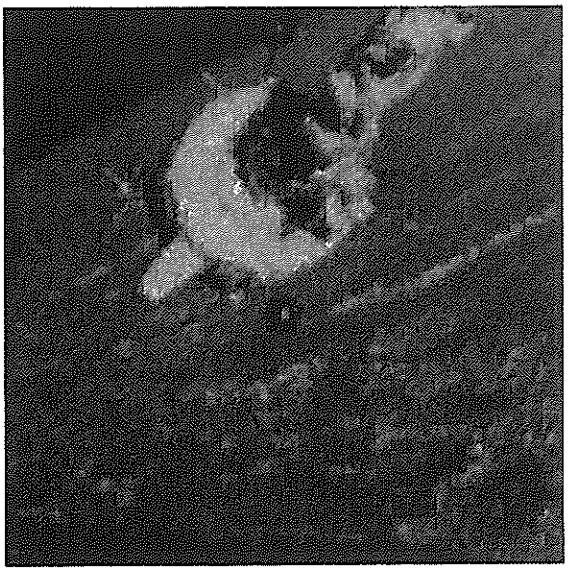


Figure 6. Image after the iteration 1. PAMSVQ scheme

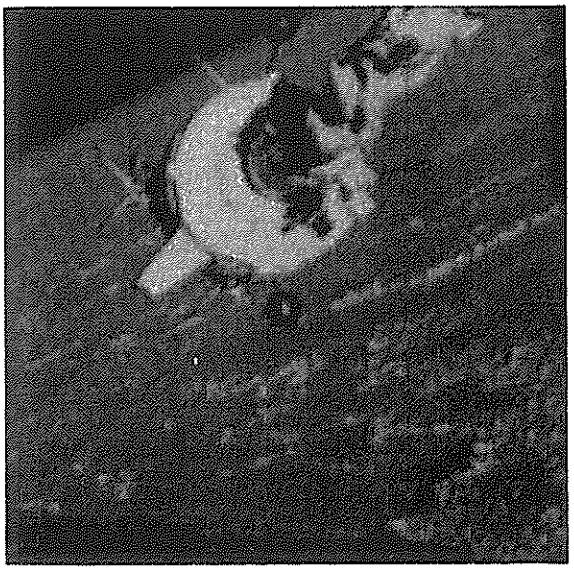


Figure 7. Image after the iteration 10. PAMSVQ scheme

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

PREDICTIVE INTERSCALE IMAGE CODING USING VECTOR QUANTIZATION

M. Antonini, M. Barlaud, P. Mathieu

LASSY Equipe I3S CNRS Université de NICE-SOPHIA ANTIPOLIS
Bat 4 S.P.I. - Rue Albert Einstein 06560 Valbonne (FRANCE)

ABSTRACT - In this paper we propose a new method for image compression associating the biorthogonal wavelet transform and an interscale prediction scheme. We use a biorthogonal wavelet transform in order to obtain a set of images at different scales and for different orientations. The method consists in predicting the position and the amplitude of the edges at a given scale using the edges of the lower scales. We also propose an interscale vector quantization scheme which permits to take into account the correlation between the wavelet coefficients inside the classification algorithm (LBG or KOHONEN neural network algorithms).

I. WAVELETS

I.1. Principle of wavelet analysis

Wavelets are functions generated from one single function ψ by dilations and translations.

$$\psi_{m,n}(x) = 2^{-m/2}\,\psi(2^{-m}x - n) \tag{1}$$

The basic idea of wavelet transform is to represent any arbitrary function f as a superposition of wavelets. Any such superposition decomposes f into different scale levels, where each level is then further decomposed with a resolution adapted to the level.

The wavelet coefficients are then computed by:

$$c_{m,n} = \langle \psi_{m,n}, f \rangle = \int \psi_{m,n}(x)\, f(x)\, dx \tag{2}$$

$$\text{and } f = \sum c_{m,n}\, \psi_{m,n} \tag{3}$$

To introduce the multiresolution notion we also define a scaling function ϕ, like the wavelet function ψ, such that:

$$\phi_{m,n}(x) = 2^{-m/2}\,\phi(2^{-m}x - n) \tag{4}$$

The projection on this family of functions $\phi_{m,n}$ gives an approximation of the signal f with resolution 2^m. The loss of information when going from an approximation of f with resolution 2^{m-1} to the coarser approximation with resolution 2^m is described by the coefficients $c_{m,n} = \langle \psi_{m,n}, f \rangle$.

I.2. Extension to sampled signals

The image signals we use are given in sampled form. These samples can be taken for the highest order resolution approximation coefficients $s_{0,n}$.

For the computation of the wavelet coefficients $c_{m,n}$ we use the following algorithm (for more details see [11]):

$$c_{m,n} = \sum_k g_{2n-k}\, s_{m-1,k} \tag{5}$$

$$s_{m,n} = \sum_k h_{2n-k}\, s_{m-1,k}$$

where $h_n = 2^{1/2} \int \phi(x-n)\,\phi(2x)\,dx$ is a low pass filter and $g_l = (-1)^l h_{-l+1}$ is a high pass filter.

These filters give exact reconstruction [7]. Equations (5), thus describe a subband coding algorithm on the sampled values $s_{m,n}$.

Furthermore, the extra ingredient in the orthonormal wavelet decomposition, comparatively to the exact reconstruction filters used in the ASSP literature, is that it writes the signal to be decomposed as a superposition of reasonably smooth elementary building blocks.

I.3. Biorthogonal wavelet basis

In image analysis, it would be convenient to use a pair of exact reconstruction filters, corresponding to an orthonormal basis with a reasonably smooth mother wavelet, and in order to have fast computation the filters should be short and symmetric as well. Unfortunately these requirements contradict each other [6].

Relaxing the orthonormality requirement and using biorthogonal bases, it is possible to preserve the symmetry of the filters which is very important in image analysis since filters with a nonlinear phase give poor results in image coding.

Biorthogonal wavelet bases were introduced by Cohen, Daubechies and Feauveau [5] and extensively used by the authors [1]: basically, there are two pairs of filters, one pair h and g for decomposition, and one pair $\tilde{h}$ and $\tilde{g}$ for reconstruction. The relations between the different filters are given by

$$\tilde{g}_n = (-1)^n h_{-n+1} \qquad g_n = (-1)^n \tilde{h}_{-n+1} \qquad \sum_n h_n \tilde{h}_{n+2k} = \delta_{k,0} \tag{6}$$

these conditions ensure exact reconstruction.

Consequently, the decomposition is still given by (5) but reconstruction becomes:

$$s_{m-1,l} = \sum_n \left[\tilde{h}_{2n-l}\, s_{m,n} + \tilde{g}_{2n-l}\, c_{m,n} \right] \tag{7}$$

There exist various extensions of the 1D wavelet transform to higher dimensions. Following the example of [11] we use a 2D wavelet transform in which horizontal and vertical orientations are considered preferential.

The wavelet decomposition of an image results in *subimages* at different resolution levels and for different orientation edges: Horizontal, Vertical and Diagonal [2].

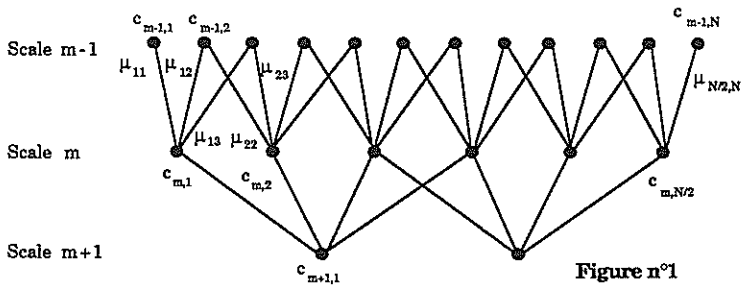
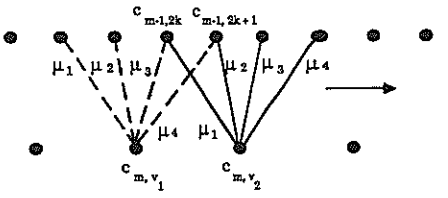


Figure n°1



1D connections between wavelet coefficients

Figure n°2

II INTERSCALE PREDICTION SCHEME

II.1. Principle

In the multiresolution decomposition of an image, there is a 'visual correlation' between the edges, or wavelet coefficients, at different scales.

We exploit this correlation and propose a method which consists in predicting the position and the amplitude of the edges at a given scale using the edges of the lower scales. This prediction is carried out independently for each subimage orientation and permits high compression rates.

II.2. One-dimensional results

II.2.1. 1D Connection of the wavelet coefficients

The goal of the method is to predict wavelet coefficients $c_{m-1,n}$ corresponding to the resolution 2^{m-1} using the wavelet coefficients $c_{m,k}$ corresponding to the resolution 2^m.

In order to *preserve spatial localisation* of singularities between the scales we must connect each of the wavelet coefficients $c_{m-1,n}$ to its close neighbours $c_{m,v}$ ($v \in V$, where V is a set of wavelet coefficients which makes up a neighbourhood of one coefficient $c_{m-1,n}$).

To determine these connections, we use the architecture of a *neural network* [8]. The architecture of the neural network is fixed by the type of multiresolution analysis and the dimension of the processed signal. Figure n°1 shows this architecture for one-dimensional signals of length 2N.

In order to make a prediction which ensures compression we use only one set of parameters μ_{ij} to predict the coefficients $c_{m-1,n}$ from the coefficients $c_{m,v}$.

The scheme of this method is depicted figure n°2. In fact, there is one set of parameters $(\mu_1,\mu_3,\ldots,\mu_{2L-1})$ for the even coefficients $c_{m-1,2k}$ and one set $(\mu_2,\mu_4,\ldots,\mu_{2L})$ for the odd coefficients $c_{m-1,2k+1}$ where 2L is the length of the filter.

II.2.2. Computation of the parameters μ_l

The relation between the coefficients $c_{m-1,i}$ and $c_{m,v}$ is given by:

$$c_{m-1,i} = \sum_{v \in V, l} \mu_l c_{m,v} + \varepsilon_{m-1,i} \qquad (8)$$

which can be write:

$$c_{m-1,i} = A_m^T(i)\,\theta + \varepsilon_{m-1,i} \qquad (9)$$

where $A_m^T(i) = [z_m(i)\ z_m(i+1)\ \ldots\ z_m(i+2L-1)]$

with

$$z_m^T = [0 \ldots 0\ c_{m,1}\ 0\ c_{m,2}\ 0\ c_{m,3} \ldots 0\ c_{m,N/2}\ 0 \ldots 0]$$

(L/2 zeros on each side)

and $\theta^T = [\mu_1\ \mu_2 \ldots \mu_{2L}]$

To compute the parameters μ_l, we must extract the predictable information; this is done using a threshold $\delta>0$, and forcing to zero the signal values included in the range $[-\delta,\delta]$.

Furthermore, if we want to privilege the high energy 'predictable' coefficients $c_{m-1,i}$, we must minimize the following criteria using a weighting factor:

$$J_\theta = \sum_i \frac{c_{m-1,i}^2}{\|A_m(i)\|^2}\,\varepsilon_{m-1,i}^2 \qquad (10)$$

for all i so that $|c_{m-1,i}| \geq \delta$

Using the gradient method with stepsize λ

$$\theta(j+1) = \theta(j) - \lambda \operatorname*{grad}_\theta J_\theta \qquad (11)$$

We find

$$\theta(j+1) = \theta(j) + 2\lambda \sum_i \left[\frac{c_{m-1,i}^2}{\|A_m(i)\|^2}\,\varepsilon_{m-1,i} A_m(i) \right] \qquad (12)$$

for all i so that $|c_{m-1,i}| \geq \delta$

II.2.3. Experimental results

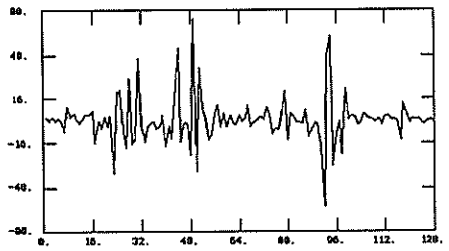
Figure n°3 shows the wavelet coefficients at the scale m=1 of a signal (256 samples) extracted from the image Lena. The wavelet coefficients were calculated by filters 5-7 associated with the biorthogonal wavelet basis given in [1] and [2] (case n°1).

Figure n°5 shows the 'predictable' part of the signal presented figure n°3; figure n°4 gives the wavelet coefficients at scale m=2 which are used to predict the signal at scale m=1 (figure n°5). The 'predictable' part is extracted using a threshold δ (Cf. § II.2.2.)

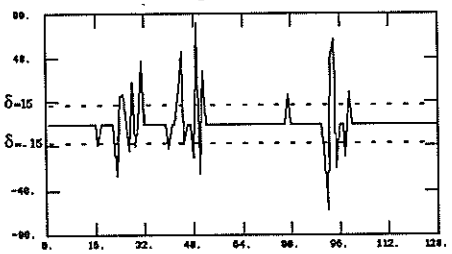
Finally, figure n°6 shows the predicted signal. The prediction filter we use in this case contains the 4 following parameters:

$\mu_1 = -0.5540$ $\qquad$ $\mu_3 = +0.2895$

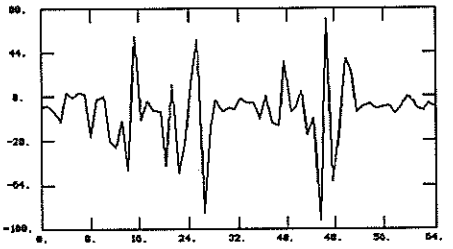
$\mu_2 = +0.3701$ $\qquad$ $\mu_4 = -0.3235$



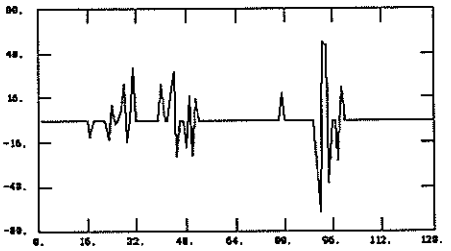
Wavelet coefficients at scale m=1
Figure n°3



'Predictable' part of signal fig. n°3
Figure n°5



Wavelet coefficients at scale m=2
Figure n°4



Predicted signal at scale m=1
Figure n°6

II.3. Extension to images

II.3.1. 2D connections of the wavelet coefficients

In the 2D case, we use a non separable square filter as an extension of the 1D filter.

The connections between the wavelet coefficients at different scales are depicted figure n°7 for a filter of length 4x4 (16 parameters) presented below:

$$\theta_{2D} = \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} \\ \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} \\ \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \\ \mu_{41} & \mu_{42} & \mu_{43} & \mu_{44} \end{pmatrix}$$

The parameters of the filter are computed, as in the 1D case, using a gradient algorithm and the same criteria to minimize.

II.3.2. Experimental results

For the 2D experimental results we use the filters 9-7 defined in [2]. The filter with length 9 has a regularity of 2 and the filters are sufficiently shorts to avoid oscillation. They seem to be well appropriate for our type of processing.

The predictable information is extracted, as in the 1D case, using a threshold δ (Cf. § II.2.2.). Here, we use δ=30 in the horizontal and vertical directions for the scale m=1, and the diagonal direction is eliminated. No visual deterioration can be observed.

The prediction is carried out between the scales m=2 and m=1 using two different prediction filters independently for the horizontal and vertical directions. The prediction results are shown figure n°9. We can compare the image figure n°9 with the scale m=1 predicted and the image figure n°10 synthesized without the resolution corresponding to m=1.

Although the PSNR of the image figure n°9 is slightly better than that of the image figure n°10 we can see that the quality of the predicted image is better, and this at low cost. In fact, if we code each filter parameter on 14 bits we get a compression of 877 at scale m=1 which corresponds about to 0.9×10^{-2} bits per pixel (bpp) to be transmitted for this scale.

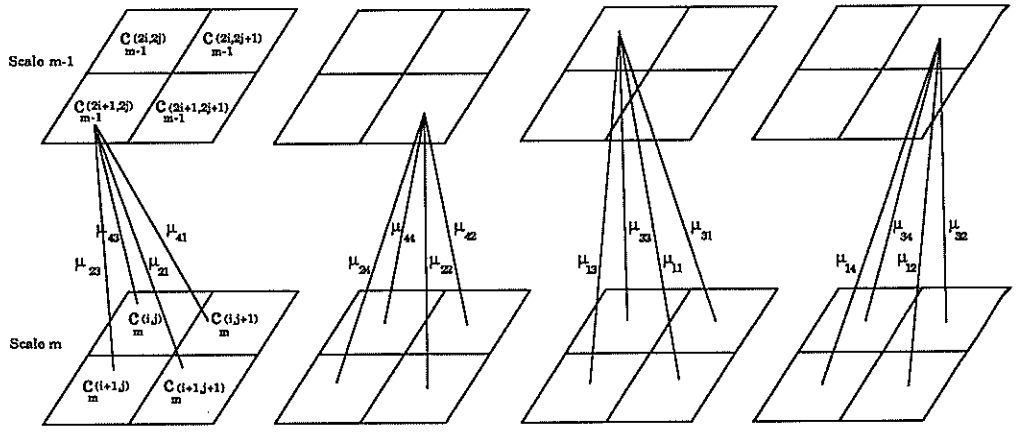
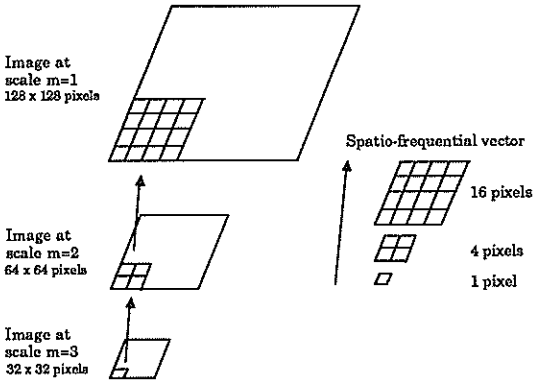


Figure n°7 *2D connections between wavelet coefficients*

Furthermore, the prediction error can be coded using a vector quantizer with a low bit rate comparatively to the original signal.

III. INTERSCALE VECTOR QUANTIZATION

Another way to consider and exploit the 'visual correlation' between the wavelet coefficients is to carry out interscale vector quantization. Each coding vector is defined both across the scale (frequency domain) and inside each direction (spatial domain): spatio-frequential vector . Thus, the correlation between the wavelet coefficients can be taken into account inside the classification algorithm. This method can be combined with the previous method presented § II.



Interscale Vector Quantization scheme

Figure n°8

The codebook is then created using a LBG or a KOHONEN neural network algorithm [13] using weighting factors to compute the distortion. These weighting factors depend on the resolution. This method permits high compression rates.

Initial result of a low bit rate coded image is presented figure n°11.

IV. REFERENCES

[1] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies "Image Coding Using Vector Quantization in the Wavelet Transform Domain" IEEE ICASSP 90.

[2] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies "Image Coding Using Wavelet Transform" submitted for publication.

[3] M. Barlaud, P. Mathieu, M. Antonini, "Wavelet Transform Image Coding Using Vector Quantization", 6 th Workshop on MDSP, Monterey California, Sep. 1989.

[4] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code", IEEE Trans. Comm. 31 (1983) 482-540.

[5] A. Cohen, I. Daubechies and J.C. Fauveau, "Biorthogonal bases of compactly supported wavelets", AT&T Bell Laboratories preprint.

[6] I. Daubechies, "Orthonormal bases of compactly supported wavelets", Comm. Pure Appl. Math. 41 (1988) 909-996.

[7] I. Daubechies, "Orthonormal bases of compactly supported wavelets. II. Variations on a theme, AT&T Bell Laboratories preprint.

[8] J.C. Feauveau "Multi-Resolution Image Segmentation by Neural Networks"

[9] A. Gersho "On the Structure of Vector Quantizers", IEEE Trans. on Inform. Theory, vol IT-28, No 2, March 1982.

[10] Y. Linde, A. Buzo, R. M. Gray, "An Algorithm for Vector Quantizer Design", IEEE Trans. on Comm., Vol. COM-28, No.1, pp. 84-95, Jan. 1980.

[11] S. Mallat "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", IEEE Trans on Pattern Anal. and Mach. intel. Vol. 11 No.7, Jul 89.

[12] P. Mathieu, M. Barlaud, M. Antonini, "Compression d'image par Transformée en Ondelette et Quantification Vectorielle", n°2 Traitement du Signal, Juin 1990.

[13] N.M. Nasrabadi, Y. Feng "Vector Quantization of Images Based Upon the Kohonen Self-Organizing Feature Maps"

[14] M.J. Smith and D.P. Barnwell, "Exact reconstruction for tree-structured subband coders", IEEE Trans. ASSP 34 (1986) 434-441.

[15] J.W. Woods, S.D. O'Neil, "Subband coding of images", IEEE Trans. on ASSP, Vol.34 No 5 oct.1986.

Figure n°9 ***Prediction at scale m=1 PSNR = 30.84 dB***

Figure n°10 ***Without the scale m=1 PSNR = 30.67 dB***

Figure n°11 ***Interscale VQ 0.54 bpp PSNR = 29 dB Filters 9-7***

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Full-Search versus Tree-Search Vector Quantization of Discrete Cosine Transform Coefficients

M. Breeuwer

Philips Research Laboratories, P.O. Box 80.000, 5600 JA Eindhoven, The Netherlands

Two picture coding systems, based on vector quantization of Discrete Cosine Transform coefficients, are compared. In one system, the complex method of full-search vector quantization is used, while in the other the less complex method of binary tree-search vector quantization is applied. It appears that to obtain equal quality, the bit rate with binary tree-search VQ should be about 30% higher than with full-search VQ.

1 Introduction

Transform coding is an efficient technique for reduction of the bit rate of digital pictures [1]. Briefly summarized, it works as follows. The picture is first divided into small blocks of e.g. 8×8 picture elements (pixels). Each block is transformed, resulting in a set of transform coefficients. A popular transform is the Discrete Cosine Transform (DCT). Bit-rate reduction is achieved by quantizing the coefficients and by efficiently mapping them on binary codewords.

Mostly, scalar quantization is used, i.e. each coefficient is quantized individually. Then, no advantage is taken of the redundancy existing among coefficients. A more efficient technique is vector quantization (VQ) [2]. However, its computational complexity is much higher than that of scalar quantization.

In [3], it was described how complexity can be reduced by using multi-stage and gain-shape VQ. With the coding system reported in that paper, digital CCIR 601 video signals can be compressed from 166 Mbit/s to about 15 Mbit/s, while maintaining high picture quality. However, the complexity of that system is still much larger than that of systems using scalar quantization.

This paper describes how complexity can be further reduced with binary tree-search VQ. First, in Section 2 the principle of VQ is briefly explained. Then, Section 3 describes a system that uses binary tree-search VQ for quantization of DCT coefficients. Results of computer simulations of the system are presented in Section 4.

2 Vector quantization

The principle of vector quantization is as follows [2]. A number of samples of the signal to be quantized is grouped into a vector. This vector is compared with a set of representative vectors, called the codebook. According to a distance criterion the closest codebook vector is selected and its index is stored or transmitted. In the decoder, the same codebook is present and the selected codebook vector is retrieved from the received index by a table-lookup procedure. Since the input vector is compared with all codebook vectors this procedure is often called full-search VQ.

Full-search VQ has a high computational complexity. A less complex method of VQ is binary tree-search VQ [4]. As is depicted in Fig. 1, a binary tree-structured codebook consists of a number of layers. During quantization, the input vector is first compared with the two codebook vectors at the first layer of the tree. The closest is chosen, and this choice determines which of the two branches leading to the second layer is followed, and so on. The index of the vector chosen at the highest layer is transmitted.

Mostly, the Euclidian distance is used to compare vectors. Then, the computational complexity of full-search VQ increases exponentially with bit rate and with the length of the input vector. For tree-search VQ the increase is only linear. This means that for large vectors and high bit rates tree-search VQ is less complex than full-search VQ. Generally, however, it has a lower performance in terms of bit rate versus quality.

Two other VQ techniques with reduced complexity are gain-shape [5] and multi-stage [6] VQ. In the encoder of a gain-shape VQ, first the gain g (square root of energy) of the input vector is calculated, is quantized to $\hat{g}$, and is transmitted as side information. The input vector is energy-normalized by dividing it by $\hat{g}$, and is subsequently vector quantized. Complexity is lower than with full-search VQ, since now the codebook only has to contain the 'shapes' of a vector. In the decoder, the quantized normalized vector is retrieved from the codebook and multiplied by $\hat{g}$.

In a multi-stage VQ, the quantization is performed in a number of stages. In the first stage, the input vector $\mathbf{v}$ is quantized to $\hat{\mathbf{v}}$ with a codebook containing a small number of vectors. The residual $\mathbf{r} = \mathbf{v} - \hat{\mathbf{v}}$ is further quantized in a second stage with a codebook especially designed for residuals, and so on. The indices of the selected codebook

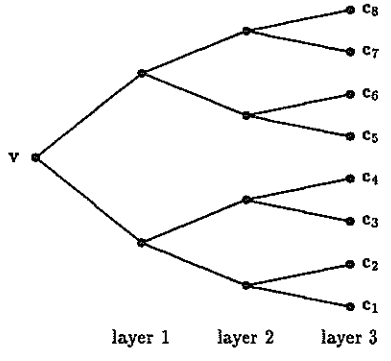


Figure 1: *Binary tree-search VQ*

vectors in each of the stages are transmitted. In the decoder, the codebook vectors selected in each of the stages are added. Computational complexity is reduced, since per stages a small number of codebook vectors is used.

3 The coding system

3.1 General description

Figure 2 shows a block diagram of the encoder. First, the luminance and the chrominance components of the picture to be coded are divided into blocks of 8×8 pixels. Each block is transformed by the DCT, giving 64 coefficients per block. One of the coefficients represents the mean pixel value in the block. Since the human visual system is highly sensitive to errors in this mean value it is separately quantized with a 9-bit uniform quantizer. The remaining 63 coefficients are vector quantized.

The human visual system is more sensitive to low-frequency noise than to high-frequency noise. This phenomenon can be turned to advantage by quantizing high-frequency coefficients more coarsely than low-frequency ones. This is realized by weighting the coefficients prior to vector quantization.

Due to the high complexity of VQ the 63 coefficients cannot be quantized as one vector. Therefore, they are divided into a number of smaller vectors called subvectors. Each subvector is quantized with gain-shape and multi-stage VQ. A comparison has been made between the results of a full-search and tree-search procedure to select the codebook vector in each stage of the multi-stage VQ.

In order to obtain a high picture quality at low bit rates, both the subvector division and the subvector quantization are performed adaptively. Each block of pixels is classified into one of five possible classes and for each class a specific division into subvectors is used (see Section 3.2). For each class specific codebooks have been generated. The sizes of the codebooks used during quantization of a subvector depend on the value of the gain of that vector (see Section 3.3). As a result, the bit rate of a coded picture depends on the amount of detail in the picture. A bit-allocation procedure is used to fix the bit rate to a pre-determined value (see Section 3.4).

For each block an index indicating the class to which the block belongs is transmitted to the decoder. Furthermore, the quantized mean value, the quantized gains of the subvectors, and the indices of the selected codebook vectors are transmitted.

3.2 Subvector division

The distribution of energy over the transform coefficients varies strongly from block to block. In principle, a fixed division of the coefficients into subvectors can be performed and codebooks can be generated for each of these subvectors. Then, for accurate quantization of the subvectors the codebooks should contain a large number of vectors, resulting in a relatively high bit rate.

A lower bit rate at the same quantization accuracy can be obtained by dividing pixel blocks into a number of classes and by using a specific division into subvectors and specific codebooks for each class. Then, the codebooks required for accurate quantization will contain a much smaller amount

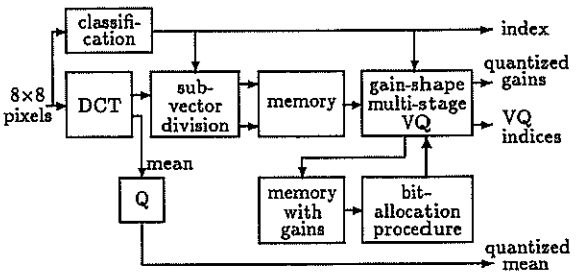


Figure 2: *Encoder*

of vectors than in the case of a fixed division, resulting in a lower bit rate. Only a small amount of side information has to be transmitted to indicate the class of a pixel block.

Experiments with different classification methods showed that the ratio of the average horizontal and vertical activity in the pixel block is a simple and efficient measure for classification. The horizontal activity A_h is defined as:

$$A_h = \sqrt{\frac{1}{8}\sum_{i=1}^{8} e_i}, \tag{1}$$

where e_i is the energy of the i-th row in the pixel block (see [3]). Equivalently, the vertical activity A_v is defined as the square root of the average of the energies of the columns in the block. The ratio $D = A_v/A_h$ indicates in which 'direction' the activity is largest. Pixel blocks are classified into one of five possible classes according to the following experimentally determined rules:

$$\begin{array}{lrcl} \text{class 1:} & & D & > 4.29 \\ \text{class 2:} & 4.29 \geq & D & > 1.62 \\ \text{class 3:} & 1.62 \geq & D & > 0.62 \\ \text{class 4:} & 0.62 \geq & D & > 0.23 \\ \text{class 5:} & 0.23 \geq & D & \end{array} \tag{2}$$

For each of the classes a specific division of the coefficients into subvectors has been defined, which is based on the following criteria. It is assumed that the DCT coefficients which contain, on average, the largest energy are the most important ones. They must be preserved as well as possible. This can be achieved by grouping them into one subvector, which enables a direct control over their quantization accuracy. Now, the codebook size required for a certain absolute quantization accuracy generally increases with the average energy of the vectors to be quantized. In order to limit the number of vectors per codebook, the high-energy coefficients are grouped into smaller subvectors than the low-energy ones. Furthermore, in order to limit computational complexity during quantization it was assumed that the length of a subvector should not exceed 16. Figure 3 gives the resulting divisions. For more details see [3].

3.3 Quantization of subvectors

In order to limit computational complexity each subvector is quantized with gain-shape and multi-stage VQ. The simplest way of performing multi-stage VQ is to use a fixed number of stages with a fixed number of codebook vectors per stage.

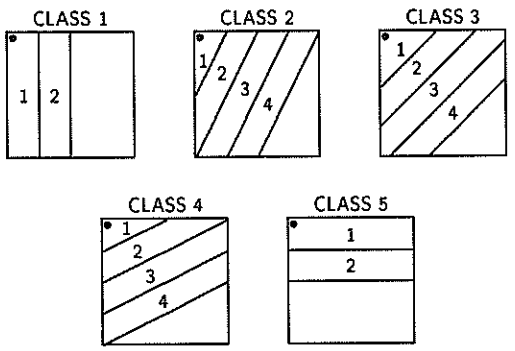


Figure 3: *Subvector division patterns for the five classes; the black dot indicates the place of the coefficient representing the mean value of the block of pixels*

This would cost a fixed amount of bits per subvector. The advantage of this method is that it is simple. It is, however, not efficient in terms of bit rate versus quality. The sizes of the codebooks required for accurate quantization will be determined by the most critical subvectors, which will generally be those with large gains. For accurate quantization of these vectors large codebooks are required. But then also subvectors with small gains are quantized with a large codebook, i.e. too many bits are used for these vectors.

A more efficient quantization can be obtained by making the number of bits used to quantize a subvector dependent on its gain. It is reasonable to assume that the larger the gain $\hat{g}$, the more bits $R(\hat{g})$ should be used. Thus a relation between $\hat{g}$ and $R(\hat{g})$ has to be found and from this relation the number of stages and the number of codebook vectors per stage have to be derived.

The relation between $\hat{g}$ and $R(\hat{g})$ was determined as follows. First, it was assumed that below a certain gain threshold g_{th} the subvector does not have to be transmitted at all, since in that case the coefficients are smaller than what can be perceived visually. The value of g_{th} was determined by visual experiments. Second, it was assumed that above a certain gain value g_{max} the bit rate does not have to be increased, since for values larger than g_{max} the noise introduced by the quantization is also below the threshold of visibility (this phenomenon is often called masking). The value of g_{max} and the corresponding amount of bits R_{max} were also determined experimentally. Third, experiments showed that for gain values between g_{th} and g_{max} a good choice is to increase the bit rate logarithmically with the gain.

Table 1 gives, as an example, the resulting relation between $\hat{g}$, $R(\hat{g})$, and the number of codebook vectors per stage for subvector 1 of class 3. It was assumed that per stage maximally 1024 codebook vectors are used. Tables as the one given in Table 1 are stored in both encoder and decoder. For every subvector of each class a specific table is used.

For each stage in the multi-stage VQ, both full-search and binary tree-search codebooks were generated from a large set of about 100 different TV pictures. The well-known LBG-algorithm was used [2]. Specific codebooks were generated for each subvector in each class. As was explained before, the sizes of the codebooks used during quantization depend on the value of the gain. Therefore, for full-search VQ codebooks with sizes 2^n, $n = 1, 2, ..., 10$ were generated. These codebook are stored both in the encoder and the decoder. For binary tree-search VQ, codebooks with ten layers were designed. Should less than 1024 vectors be used then only a limited number of layers of the tree-structured codebook is involved in the search. For example, should 256 vectors be used, then the search is performed up to and including the eighth layer.

$\log_2(\hat{g})$	nr. bits	stage 1	stage 2	stage 3
0.00	0	0	0	0
1.14	4	16	0	0
2.29	9	512	0	0
3.43	13	1024	8	0
4.57	17	1024	128	0
5.71	21	1024	1024	2
6.86	26	1024	1024	64
8.00	30	1024	1024	1024

Table 1: *Example of relation between the quantized gain of a subvector, and the number of bits used for quantization. For each stage in the multi-stage VQ the number of codebook vectors used is indicated.*

3.4 Fixed bit rate coding

Quantization according to Table 1 means that high-energy subvectors are quantized at a higher bit rate than low-energy vectors. Then, the number of bits needed to represent a picture will depend on the amount of detail in the picture. For most applications the number of bits per picture must be fixed. This section describes a simple procedure for fixing the bit rate.

It is assumed that all M subvectors $\mathbf{v}_i$ and the corresponding quantized gains $\hat{g}_i$ $(i = 1, ..., M)$ of the picture to be coded are temporarily stored in a memory and that the available bits are divided over these vectors. Furthermore, it is assumed that for every type of subvector in each class a bit allocation table similar to that in Table 1 is present.

First the total 'overhead' bit rate R_o is calculated, which is the amount of bits needed for the mean value and the classification index of all blocks of a picture, and for the quantized gains of the subvectors. If R bits are available to code the entire picture then the amount left for quantization of the subvectors is $R_a = R - R_o$. On the basis of the bit allocation tables it is determined how many bits $R(\hat{g}_i)$ would be required to quantize each subvector $\mathbf{v}_i$ when no bit rate control is applied and for each subvector this number is stored in a memory. Then the total number of bits R_{tot} needed to quantize all subvectors is equal to $R_{tot} = \sum_{i=1}^{M} R(\hat{g}_i)$.

The required amount R_{tot} is compared with the available amount R_a. If they are equal no action is needed. If the required amount of bits is too large, the subvectors must be quantized more coarsely than determined by the bit allocation tables. If the rate is too small, a finer quantization must be applied. In the remainder of this section the procedure followed when the required amount of bits is too large is described. A similar procedure can be used if the amount is too small.

First, the difference $R_d = R_{tot} - R_a$ is calculated. Now, if for quantization of the first subvector $\mathbf{v}_1$ more than zero bits would be used, i.e. if $R(\hat{g}_1) > 0$, then one bit less is allocated

to $\mathbf{v}_1$ and R_d is decreased by one. If after this decrease R_d is still larger than zero the same procedure is applied to the second subvector $\mathbf{v}_2$, and so on. If, after having treated all subvectors in this way, bits still have to be saved, the procedure is again started from the first subvector. The actual quantization starts as soon as R_d has become equal to zero.

The bit allocation tables used in the encoder are assumed also to be present in the decoder. Furthermore, it is assumed that the decoder knows at which bit rate the encoder is operating. Then the procedure described above can be exactly reproduced in the decoder on the basis of the received quantized gains of all subvectors.

4 Results and discussion

The system was evaluated by computer simulations. It was optimized for coding frames of digital CCIR 601 TV signals, which require 166 Mbit/s in uncoded form. They were coded at about 15–20 Mbit/s.

Full-search VQ appears to perform significantly better than tree-search VQ. At distances of about 1.5 m from a Barco CDCT 5151 TV monitor only minor differences can be seen between the original signals and the ones coded with full-search VQ at 15 Mbit/s. At 20 Mbit/s no significant differences are visible. For obtaining comparible picture quality, the bit rate with tree-search VQ should be about 30% higher than that with full-search VQ. However, tree-search VQ has a much lower complexity than full-search VQ. At bit rates around 15 Mbit/s, tree-search VQ requires several tens of multiplications per pixel, whereas full-search VQ requires several hundreds of multiplications per pixel.

Figure 4 gives an example of the signal-to-noise ratio versus bit rate for full-search and tree-search VQ.

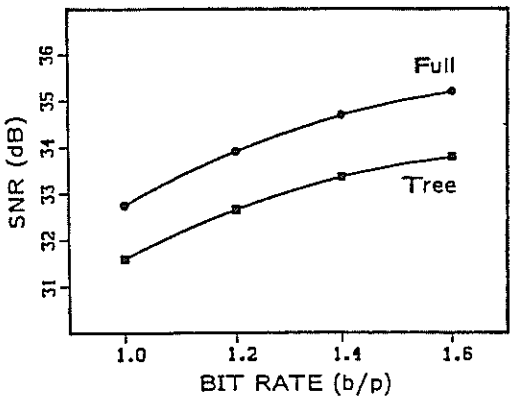


Figure 4: *Example of SNR versus bit rate*

References

[1] N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, New Jersey, 1984.

[2] R.M. Gray, 'Vector Quantization', *IEEE ASSP Magazine*, April 1984, pp. 4-29.

[3] M. Breeuwer, 'Transform Coding of Images using Directionally Adaptive Vector Quantization', *Proc. ICASSP' 88*, pp. 788-791.

[4] A. Buzo, A.H. Gray, Jr., R.M. Gray and J.D. Markel, 'Speech Coding Based Upon Vector Quantization', *IEEE Trans. Acoustics, Speech, and Sign. Proc.*, Vol. ASSP-28, No.5, October 1980, pp. 562-574.

[5] M.J. Sabin and R.M. Gray, 'Product Code Vector Quantizers for Speech Waveform Coding', *Proc. IEEE Globecom '82*, pp. 1087-1091.

[6] B. Juang and A.H. Gray, Jr., 'Multiple Stage Vector Quantization for Speech Coding', *Proc. ICASSP' 82*, pp. 597-600.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A frequency Bin Adaptive Separation Approach for Co-channel Interference Speech Suppression

Y. H. Gu and W. M. G. van Bokhoven
Dept. of Electr. Eng., Eindhoven Univ. of Technology,
P.O.Box 513, 5600 MB Eindhoven, The Netherlands

A new adaptive speech separation approach is developed based on frequency-bin time-directional LMS adaptive filtering. A short-time local Target-Interference Energy Ratio (TIR) is estimated. At each bin the algorithm is performed to adapt first to the stronger speaker's signal. In addition, a two-speakers' pitch estimation algorithm is developed by using coincidence appearances of pitch information from both the signal envelop and 'carrier' at bins. Simulations on separating synthetic speech signals with fixed pitches over a wide range of TIR between 0 dB and ±12 dB showed good results. Comparisons with HMS at -12 dB have been made. Informal listening tests and spectrograms show that the proposed algorithm is a promising approach for intelligibility enhancement of speech.

I. Introduction

The problem addressed by this study is the intelligibility enhancement of the target speech when two speech signals are additively combined in a single communication channel. In previous studies, a number of techniques have been developed [1]-[9]. Most of these methods are based on the property that voiced speech energy concentrates on pitch harmonic frequencies. On the other hand, Alexander[4] has concluded that a LMS algorithm adapts rapidly to the dominant speech corrupted by weaker interference speech. Few further progress has been reported because of this high Target-Interference Energy Ratio (TIR) restriction.

The alternative speech separation approach proposed in this paper, which, to authors' knowledge, is the first report using frequency bin LMS adaptive filtering for co-channel speech separation, is found to be consistent with the general processing methods of the auditory system.

II. System Description

As shown in the system block diagram in Fig.1, this speech separation system consists of four main parts:

(1) A wide-bandpass filterbank splitting the signal into frequency bins.
(2) An algorithm for simultaneously estimating two speakers' pitches from summed signal.
(3) Estimation of Short-time local TIR at each frequency bin.
(4) A frequency bin adaptive speech separation algorithm.

The detailed system implementation will be discussed in the following sections.

III. Estimating Two Pitches from Summed signal

This speech separation system needs accurate pitch estimation of both the target and the interference over a wide range of TIR.

In many previous studies, the dominant speaker's pitch is only concerned [2][3]. In such a case, pitch estimation can be performed by using the previously developed algorithms for a single speaker's signal. For simultaneously estimating two pitches, Parson[1] uses histogram of spectral peak submultiples, and Weintraub[9] uses a neural firing event coincidence function.

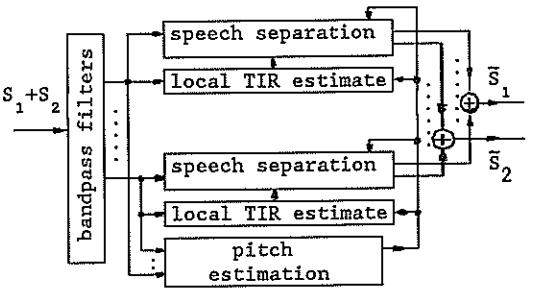


Fig.1 Frequency bin speech separation system

The two pitch estimation algorithm described in this paper uses envelop autocorrelation and 'carrier' period at each bin.

Filter Bank Performed by RSTFT

The running short-time Fourier transform (RSTFT) can be explained in terms of the simultaneous output of a bandpass filterbank with uniformly spaced center frequencies from 0 to π. Given x(t), define the transform as

$$X(n,\omega_k) = \int_{-\infty}^{\infty} x(n+m)w(m)\, e^{-j\omega_k m}\, dm \qquad (1)$$

where w(m) is symmetric window with length L. $X(n,\omega_k)$ can be explained as the complex output of a bandpass filter with center frequency $\omega_k = \frac{2\pi k}{N}$, ($N \geq L$), where N is the total

number of bandpass filters. The corresponding impulse response of the complex filter is

$$h_k(n) = w(-n)\, e^{j\omega_k n} \qquad (2)$$

Choosing the window length comparable to the pitch period, the spectrogram will keep high time resolution with periodicity and formant.

Pitch Estimation Algorithm

1)At each bin, short-time envelop autocorrelations are calculated. The average intervals between two successive signal peaks over short-time duration are also calculated as 'carrier' period, and the time indices of 'carrier' period multiples are marked.
2)At each time index, the numbers of the appearance of envelop autocorrelation peaks and 'carrier' multiples are accumulated respectively over all bins. The autocorrelation values associated with the peaks are also accumulated at each time index.
3)The first several indices with the highest values are picked up as pitch candidates, and the weighted sum of these values are calculated in order to select pitches.

IV. Adaptive Frequency Bin Speech Separation

One of the important features of this speech separating system is that it is consistent with the processing methods of the human auditory system without mimicing its behaviour. Psycho-acoustic experiments indicate that the ear performs some sort of running short -time spectral analysis on the acoustic wave-forms, by decomposing a signal into isolated frequency components with further processing done essentially along the time axis.

It is noticed that the globally stronger speaker's signal can also be the weaker one in some frequency region. In general, target and interference signal can dominate differently in the various frequency bins. By splitting the signal into frequency bins, speech separation is divided into a group of monotonic problems, each containing one stronger speaker.

Short-Time Frequency Bin TIR Estimation

Short-time frequency bin TIR is roughly estimated in order to decide the associated stronger voiced speaker at each bin.

First, the bin signal y(t,k) is half-wave rectified to $\tilde{y}(t,k)$ and the third order moment with the lag of each speaker's pitch period P_m is calculated

$$Z_m(t,k) = \tilde{y}(t,k)\tilde{y}(t-P_m,k)\tilde{y}(t-2P_m,k) \qquad (3)$$

where P_m can be provided by the above pitch estimation algorithm. The short-time frequency bin TIR is then estimated by the energy ratio

$$TIR(t_0,k) = \sum_t Z_T^2(t,k) \Big/ \sum_t Z_I^2(t,k) \qquad (4)$$

where $t_0 \in \{ t \}$.

Frequency Bin Adaptive Speech Separation via LMS filtering

Let's consider the situation of a signal composed by one target and one interference speaker. The bandpassed signal at frequency bin k is

$$y(t,k) = S_T(t,k) + S_I(t,k) \qquad (5)$$

where $y(t,k)=Re[X(t,k)]$. For voiced-speech, it is highly correlated among the successive periods. Supposing S_T and S_I are statistically uncorrelated, the one pitch period P_m delayed autocorrelation of bin signal y(t,k) is

$$E(y(t,k)y(t-P_T-j,k)] = R_{S_T}(j,k) + R_{S_I}(P_T+j,k) \qquad (6)$$

where target pitch period P_T is chosen, which can be provided by the above pitch estimation algorithm. A similar analysis holds for the interference pitch P_I. When one speaker's signal dominates at the bin, the first term in (6) becomes the main part. A LMS algorithm can then be very effectively used to adapt to the stronger signal using periodic correlation.

The separation error at each frequency bin depends on the difference of the local energy, the pitch and the 'carrier' period between the two speakers. The overal error of the separated speech is the error sum of the associated one over all frequency bins.

The separation algorithm consists of three steps:

1)Decisions are made using the estimated $TIR(t_0,k)$ on which speaker is locally stronger, and whether or not this speaker's signal is strongly dominant at the bin.
2)A LMS algorithm[10] is applied to adapt to the stronger speech signal by using the periodic correlation. The output of the adaptive filter d(t,k)

$$d(t,k) = \sum_{j=0}^{n-1} W_j(t,k)\, y(t-P_s-j,k) \qquad (7)$$

is assigned to the stronger speaker as the separated signal. The weights of each bin are updated at every time sample

$$W_j(t+1,k) = W_j(t,k) + 2\mu_k e(t,k) y(t-P_s-j,k) \qquad (8)$$

where $j=0..(n-1)$, $\mu_k = \dfrac{\mu_0}{n\, E[y^2(t,k)]}$ is the step size at bin k, which is normalized by the short-time energy of that bin. μ_0 is a positive small constant, n is filter order.
3)If the stronger speaker's signal is not strongly dominant, the residual

$$e(t,k) = y(t,k) - d(t,k) \qquad (9)$$

is assigned to the weaker speaker as the separated signal, otherwise zero is assigned.

In the situation where the signal consists of one voiced and one unvoiced speech, the voiced speech signal is adapted first.

V. Simulations and Results

Computer simulations have been tested in two steps in order to check:

1. The effectiveness of the separation algorithm.
2. The convergence speed of the algorithm for

adaptation to speech sentences with time-varying vocal tract parameters.

(1) Separation of Synthetic Voiced-Speech

Two synthetic speech signals, each containing three formants with a specific pitch, are added with properly chosen TIR between 0 dB and ±12 dB. The adaptive separation algorithm is applied using only the summed signal, where μ_0=0.1 is chosen. The separated speech signals are then compared with the clean ones. Fig.2 shows the LPC spectra.

(2) Comparisons with HMS at -12 dB TIR

Comparisons with HMS method[2] on separated speech at -12 dB indicate a better performance of our method. Fig.2.(b3) shows an example.

(3) Separation of Summed Synthetic Speech Sentences with fixed pitches

Two synthetic sentences with different fixed pitches are added with the chosen global TIR between 0 dB and ±12 dB. The bin TIR and the μ_k are calculated in each frame. Simulations showed that the algorithm can adapt quickly to speech sentences. Informal listening tests showed intelligibility enhancement of the target speech. The spectrograms also indicated similar results. Fig.3 shows an example.

(4) Pitch Estimation from Co-channel signal

The pitch estimation algorithm is tested on summed synthetic voiced speech over a wide range of TIR. Fig.4 shows an example.

VI. Conclusions and Future Expectations

An adaptive speech separation algorithm has been developed for co-channel interference speech suppression over a range of TIR between 0 dB and ±12 dB which seems promising. The system has been tested on summed synthetic speech signals of two speakers with fixed pitches, which showed intelligibility enhancement of the target speech. The two speakers' pitch estimation algorithm has been tested, and provided good pitch estimation to the separation system. Further research is continued on the improvement of the two speakers' pitch estimation and the refinement of the system in order to adapt to natural speech.

References

[1]T.W. Parsons , " Separation of Speech from Interfering Speech by Means of Harmonic Selection ", J. Acoust. Soc. Am., Vol 60, No.4, PP 911-918, 1976.

[2]B.A.Hanson & D.Y.Wang, "The Harmonic Magnitude Suppression Technique for Intelligibility Enhancement in the Presence of Interfering Speech", PP.18.A.5.1-5.4, ICASSP 84.

[3]J.A.Naylor & S.F.Boll, "Techniques for Suppression of an Interfering Talker in Co-channel Speech", PP. 205-208, ICASSP 1984.

[4]S.T.Alexander,"Adaptive Reduction of Interfering Speaker Noise Using the Least Mean Squares Algorithm", PP.728-731, ICASSP 87.

[5]D.G.Childers & C.K.Lee, "Co-channel Speech Separation ", PP. 181-184, ICASSP 1987.

[6]T.F. Quatieri and R.G. Danisewicz, " An Approach to Co-channel Talker Interference Suppression using a Sinusoidal Model for Speech ", PP. 565-568, ICASSP 1988.

[7]K.Min, et al., "Automated Two Speaker Separation System ", pp. 537-540, ICASSP 1988.

[8]C.Rogers, et al., " Neural Network Enhancement for a Two Speaker Separation System", PP 357-360, ICASSP 1989.

[9]M.Weintraub, "A Computational Model for Separating Two Simultaneous Talkers ", PP. 81-84, ICASSP 1984.

[10]B.Windrow, et al., "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter ", Vol. 64, No.3, PP. 1151-1162, Proc. IEEE, 1976.

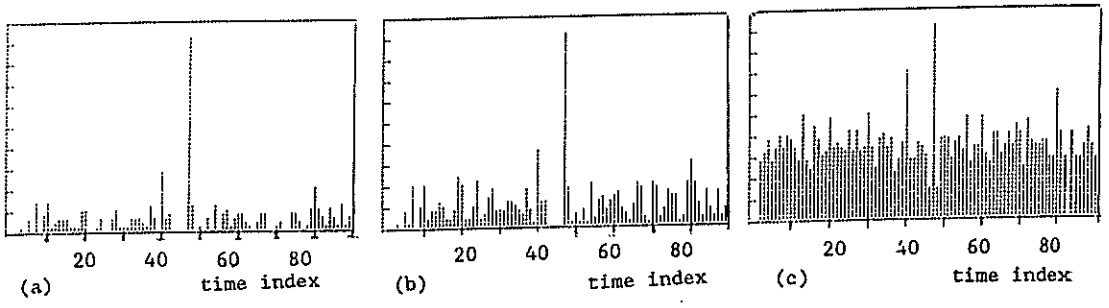


Fig.4. Estimation of Two pitches from co-channel signal at 0 dB TIR
(p1 = 40 samples, p2 = 47 samples, fs = 8 kHz)
(a) Appearance numbers of envelop autocorrelation peaks
(b) Accumulated values associated with the autocorrelation peaks
(c) Appearance numbers of 'carrier' period multiples

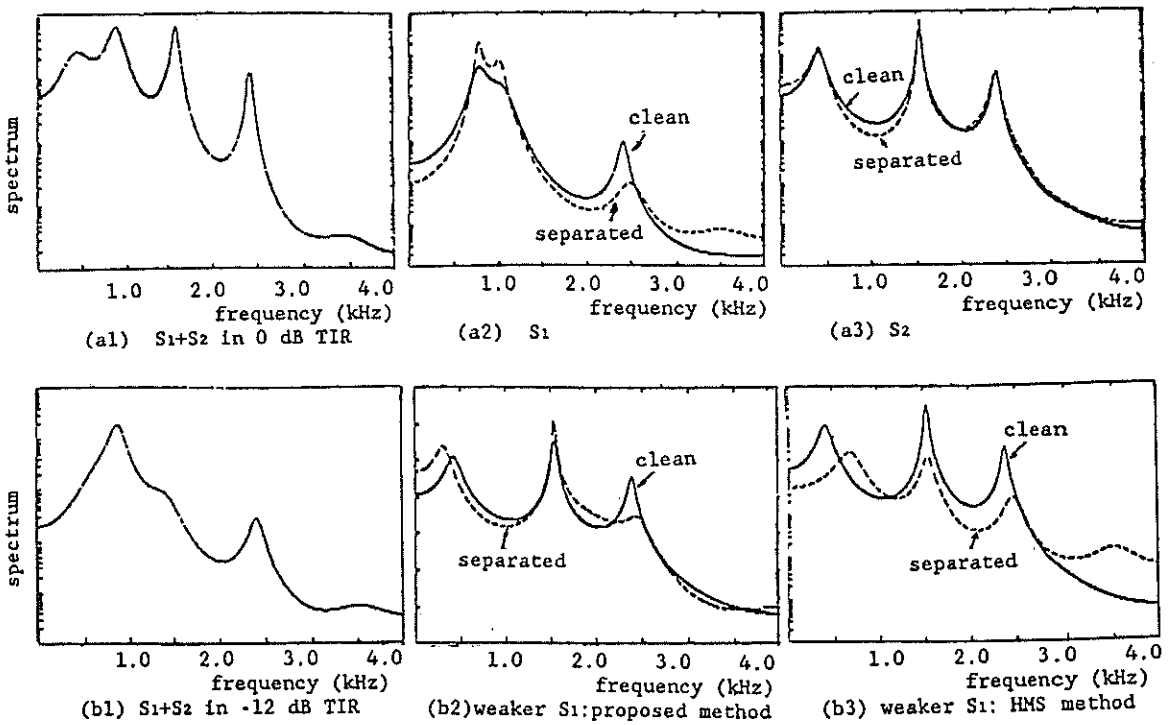


Fig.2. LPC spectra of the speech

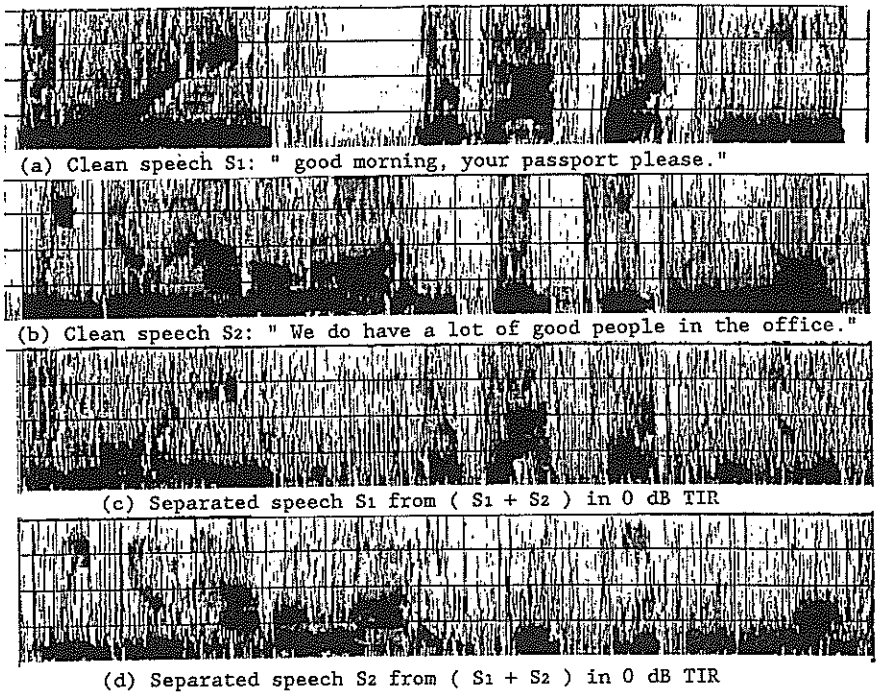


Fig.3 Speech spectrograms

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ON USING THE COHERENCE FUNCTION FOR NOISE REDUCTION

Régine LE BOUQUIN, Gérard FAUCON

Laboratoire Traitement du Signal-IRISA, Université de Rennes I, Campus de Beaulieu,
35042 Rennes Cedex, France

Our concern is the estimation of a signal disturbed by an additive noise when M observations are available ; on each channel, we've got an observation composed of signal + noise ($x_i = s_i + b_i$). The aimed application is the enhancement of noisy speech for radio-mobile applications. We present here two methods, both of them based on the coherence function. In the first one, we directly use the coherence fonction ($\rho(f)$) to weight the observation x_1 in order to estimate the signal s_1 (chosen as the reference signal) ; in the second one, we use this function to determine a speech/noise classification algorithm : then we evaluate the noise spectra when theoretically no speech is present with a view to a possible spectral subtraction to reconstruct the original speech signal. Finally some objective and subjective tests will be performed.

1. INTRODUCTION

This paper deals with the problem of the continuous estimation of a signal disturbed by an additive noise when two or M observations $s_i + b_i$ are available ; our aimed application is the enhancement of noisy speech recorded in a car. Since it's difficult, for technical reasons, to make recordings on many sensors in a moving car to create our data base, the procedures we propose are described with M = 2 ; of course it's quite easy to generalize our problem to a situation using M (M > 2) observations.

In the following we'll call x_1 and x_2 the observations on the first and the second channels:

$$x_1 = s_1 + b_1$$
$$x_2 = s_2 + b_2$$

Figure 1 represents the module of the coherence function between b_1 and b_2 in the following case : the first microphone is placed on the sun-visor in front of the speaker and the second one is placed on the stile of the left door beside him, so that the distance between the two sensors is about 40 cm. We note that $|\rho_{b_1 b_2}(f)|$ is about 0.7 in the low frequencies but it rapidly decreases in the higher frequencies ($\simeq 0.3$). So, by assuming $|\rho_{b_1 b_2}(f)|$ weak, it seems that the value of the coherence function between observations will be an effective criterion to determine if speech signals exist or not as we'll see it next. Let us note that no knowledge or learning of the characteristics of noises is necessary and the proposed processes directly work on present observations.

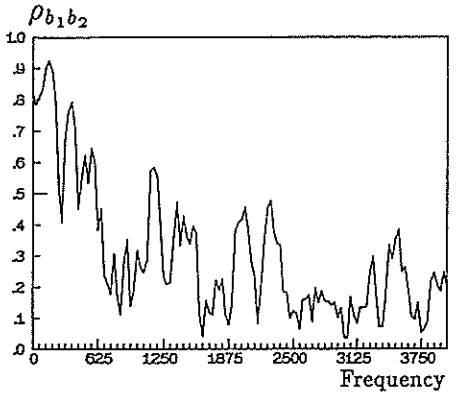


Figure 1 : Coherence between noises $\rho_{b_1 b_2}(f)$

2. The basic principle : the coherence function

2.1. Definitions and Hypotheses

The coherence function between two signals x and x' is defined like this :

$$\rho_{xx'}(f) = \frac{\gamma_{xx'}(f)}{\sqrt{\gamma_{xx}(f)\,\gamma_{x'x'}(f)}}$$

where :
$\gamma_{xx}(f)$ and $\gamma_{x'x'}(f)$ are respectively the spectral densities of signals x and x'.
$\gamma_{xx'}(f)$ is the interspectral density between x and x'. We define another value, called the Magnitude Squared Coherence (MSC), as :

$$MSC(f) = |\rho_{xx'}(f)|^2 .$$

The only hypotheses we need are :
a) speech signals and noises are independent
b) signals s_1 and s_2 are derived from the signal issued from the speaker's mouth by

filtering inside the car ; the source signal is unique and the module of the coherence between the signals picked up by each microphone is close to 1 on all the frequency bandwidth [0..4kHz].

c) as for noises b_1 and b_2, the correlation principally depends on the distance d between microphones, but also on their location and nature. Owing to the distance d we measure, we may assume the noises are spatially uncorrelated.

2.2. Computation of the spectral densities

Each observation is sampled at a sufficiently high frequency rate to prevent frequency aliasing. Let D be the sample period of the sequence ($D = 1.25\ 10^{-4}$ sec.). We define a window function w(n) , which is a low-pass filter such as a Hamming window. The short-term spectrum [1] of a signal $x_1(n)$ is computed every T seconds ($T > D$) on the modified sequence $x^w_{1,k}(n) = x_1(n)\ w(k-n)$. It may be written :

$$X_1(mF,kT) = \mathcal{F}[w(kT-nD)\ x_1(nD)]$$

where m is the frequency index and k defines the position of the window. $\mathcal{F}$ is the discrete Fourier transform and $X_1(mF,kT)$ is defined by :

$$X_1(mF,kT) = \sum_{n=\frac{kT}{D}-\frac{1}{2}N}^{\frac{kT}{D}+\frac{1}{2}N-1} w(kT-nD)\ x_1(nD)\ e^{-imnFD}$$

The bandwidth B of $X_1(mF,kT)$ is the same as that of the window. We define it as the lowest frequency for which the log magnitude spectrum remains 42 dB below the maximum value. For a window of L points, $B = 2/LD$. According to the Nyquist theorem, the short-term spectrum must be sampled at least at a rate of 2B samples per second. Thus, $T = 1/2B = LD/4$. After computing $X_1(mF,kT)$ and $X_2(mF,kT)$, the spectral densities are estimated using a time averaging by a simple recursive calculation :

$$\gamma_{x_i x_j}(mF,kT) = a\ \gamma_{x_i x_j}(mF,(k-1)T) + X_i(mF,kT)X_j^*(mF,kT) \quad (i,j=1,2)$$

3. THE FIRST METHOD

According to what we already said, the objective of the coherence function ($\rho(f) = \rho_{x_1 x_2}(f)$) consists, for each frequency, in turning off uncorrelated signals and passing correlated ones. To this end, we choose the following decision criterion :

- if $|\rho(f)| \geqslant S_{max}$, the speech signal is predominant and must be passed without distortion
- if $|\rho(f)| \leqslant S_{min}$, we assume there are only disturbing noises
- if $S_{min} < |\rho(f)| < S_{max}$, we assume there are signal + noise and the observation x_1 will be weighted by a function of $\rho(f)$.

In a more general manner, we'll write :

$$\hat{S}_1(f) = (MSC(f))^{\alpha}\ X_1(f), \quad S_{min} < |\rho(f)| < S_{max}$$

$$\hat{S}_1(f) = X_1(f), \quad |\rho(f)| \geqslant S_{max}$$

$$\hat{S}_1(f) = (S_{min})^{2\alpha}\ X_1(f), \quad |\rho(f)| \leqslant S_{min}$$

The filtering is all the more rigid as α is great ; the first equation may be modified to take into account the thresholds and avoid discontinuities in the weighting function.

We justify this algorithm with regard to the Wiener filtering in a particular case. We assume that a preprocessing may be achieved to get the speech signals on each observation almost identical ; we learn a transfer function (F) between s_1 and s_2 in the absence of noise, in the stopped car ; this transfer function is assumed to be stationary whether noise is present or not. In a second step, the noises are present and the filter F is locked ; the new observations become : $s_1 + b_1$ and $F(s_2 + b_2) = s_2^F + b_2^F$ with $s_2^F \simeq s_1$ (if F is perfectly learned $s_2^F = s_1$) . When we use the module of the coherence function to filter the observation x_1 and set S_{max} and S_{min} respectively to 1 and 0, using $\alpha = 0.5$, the filtering of $X_1(f)$ is

$$\frac{\gamma_{s_1 s_2^F}(f)}{(\gamma_{s_1}(f) + \gamma_{b_1}(f))^{1/2}\ (\gamma_{s_2^F}(f) + \gamma_{b_2^F}(f))^{1/2}}$$

So, if $s_2^F \simeq s_1$ and $\gamma_{b_1} \simeq \gamma_{b_2^F}$, the filtering tends to the optimal Wiener filtering when one estimates s_1 from x_1. Let's justify both modifications :

- the thresholds S_{min} and S_{max} lead to a better discrimination when signal or noise is preponderant.
- the parameter α allows to obtain a filtering more selective.

Given $\hat{S}_1(f)$, we finally deduce $\hat{s}_1$ by IFFT and overlap-add. Note that, when the speech signals are really identical (on each channel), we can either work on $X_1(f)$ or on $X_2(f)$ and so estimate s_1 from both channels.

4. THE SECOND METHOD

We propose now to use the coherence function and the spectral subtraction to derive the advantages of both of them :

a) the coherence function is used to determine a speech/noise classification algorithm ; in practice, we'd better work on the MSC averaged on the whole bandwidth [0..4 kHz] for each block of 256 samples ($\overline{MSC}$) and it's represented on Figure 2 for the following sequence (noise, sentence, noise, sentence, noise).

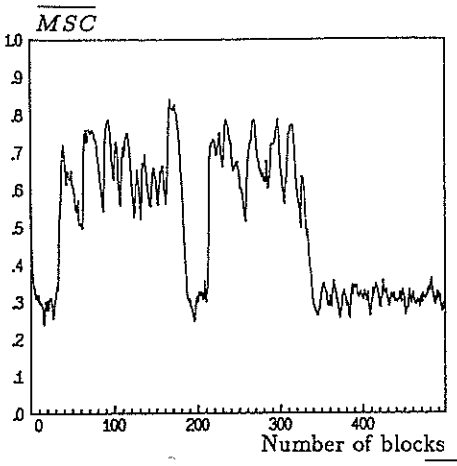


Figure 2 : Magnitude Squared Coherence $\overline{MSC}$

This figure shows that we can distinguish two parts, the first one which is relative to the silence periods and the second one relative to the sentences.

- So, if the $\overline{MSC}$ remains below a threshold (S_o), we evaluate the noise spectrum $\gamma_{B_1}(f)$; a local running count L_i is kept for the number of consecutive blocks for which the $\overline{MSC}$ falls below S_o . If L_i reaches L_{max}, we average the noise d.s.p. $\gamma_{B_1}(f)$ on the last L_{max} blocks ; $\bar{\gamma}_{B_1}(f)$ will then be used in the next spectral subtraction.
- When $\overline{MSC} > S_o$, we assume there are signal + noise.

In this step of speech detection, we must take into account the beginning and the end of each sentence ; so, some delays are introduced not to damage the speech signal, when the MSC moves from ($S_o - \varepsilon$) to ($S_o + \varepsilon$) and vice versa.

b) In the second step ($\overline{MSC} > S_o$), we perform a spectral subtraction [2] to reconstruct the original speech signal ; for each block of 256 samples, we can subtract the noise spectrum from the complete spectrum in order to get the estimated speech signal $\hat{s}_1$.

The algorithm we use for the spectral subtraction is the following :

$$\text{let } D(f) = \gamma^{\delta}_{X_1}(f) - \alpha\, \bar{\gamma}^{\delta}_{B_1}(f)$$

$$\gamma_{S_1}(f) = \begin{cases} D^{1/\delta}(f) & \text{if } D^{1/\delta}(f) > \beta\, \bar{\gamma}_{B_1}(f) \\ \beta\, \bar{\gamma}_{B_1}(f) & \text{otherwise} \end{cases}$$

with $\alpha \geqslant 1$ and $0 \leqslant \beta \ll 1$

α is the subtraction factor and choosing $\alpha > 1$ reduces the spectral noise peaks. β is the spectral floor parameter : the spectral components of $\gamma_{s_1}(f)$ are prevented from descending below the lower bound $\beta\, \bar{\gamma}_{B_1}(f)$ and for $\beta > 0$, the remnants of noise peaks are "masked" by neighboring spectral components of comparable magnitude. Speech processed by this modified method has less musical noise than speech processed by the classical spectral subtraction.

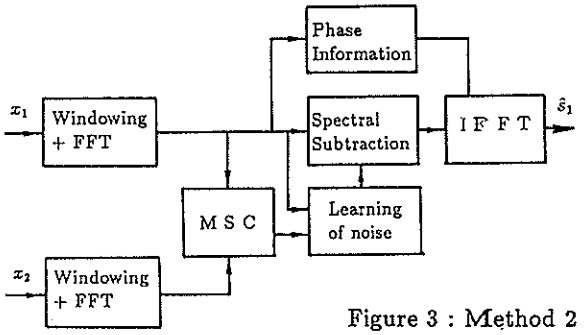


Figure 3 : Method 2

Finally, an important overlapping (50 %) is applied to form the estimated speech signal $\hat{s}_1$. The block diagram relative to this structure is represented Figure 3. The same process may be applied to x_2, so that $\hat{s}_1$ and $\hat{s}_2$ are combined to form the estimated signal.

5. EXPERIMENTAL RESULTS

In order to evaluate the performances of the two methods, two sets of recordings have been performed in different vehicles.

set 1 : it corresponds to the case presented in the paragraph 1, when speech signals and noises have been recorded simultaneously.

set 2 : the configuration is the same as previously but speech signals and noises have been recorded separately.

For the first set of recordings, only subjective tests are performed to justify the retained parameters : for all experiments, the forgetting factor a is held to a fixed value 0.65. The values of the parameters for the first method are : $S_{min} = 0.1$, $S_{max} = 0.9$, $\alpha = 2$ and for the second method $S_o = 0.35$, $\alpha = 1.5$, $\beta = 0.06$ and $\delta = 0.5$.

Even if these subjective tests tend to prove the effectiveness of our procedures (with regard to the quality, intelligibility and pleasantness), we use the second set of data to evaluate performances using some objective criteria. We first define the gain on the signal-to-noise ratio but the results are not always significative. Then we compute a distance function between the clean speech and the noisy enhanced speech : the LPC cepstral distortion measure (d_{CEP}) ; the way to compute d_{CEP} is presented here : let C_k be the cepstrum of a discrete time series, x(n) :

$$C_k = \int_{-\pi}^{\pi} \text{Log}|X(\lambda)|^2\, e^{j\lambda k}\, \frac{d\lambda}{2\pi}$$

The L_2 norm of the log spectral distortion measure between two time series, x(n) and x'(n), is given as :

$$d_{L_2}(x,x') = \int_{-\pi}^{\pi} |\mathrm{Log}|X(\lambda)|^2 - \mathrm{Log}|X'(\lambda)|^2|^2 \frac{d\lambda}{2\pi}$$

and can be equivalently represented in terms of the corresponding cepstral coefficients.
The cepstral distortion measure is a truncated version of d_{L_2} :

$$d_{CEP}(x,x') = \sum_{\ell=1}^{2p} (C_\ell - C'_\ell)^2 \ .$$

In our study, the cepstral coefficients are computed recursively from the linear predictive coefficients (a_0, a_1 .. a_p, p = 8).

$$C_o = \ell n(\sigma^2)$$

$$C_i = - a_i - \sum_{n=1}^{i-1} (1 - \frac{n}{i})\, a_n\, C_{i-n} \qquad i > 0$$

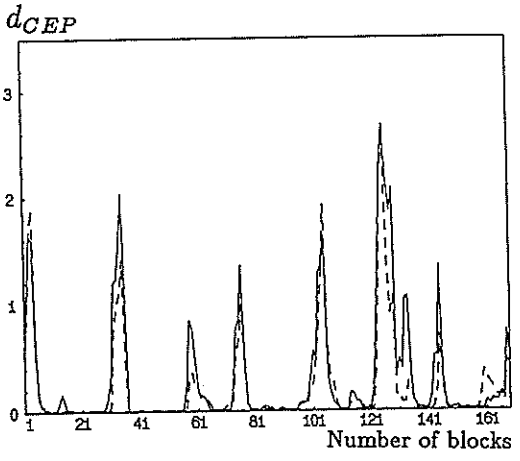


Figure 4 : Distance d_{CEP}, method 1

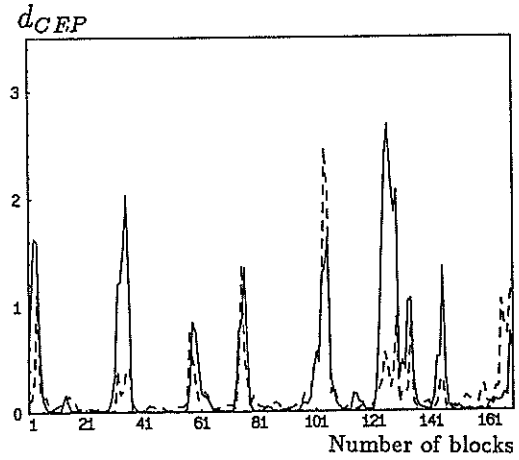


Figure 5 : Distance d_{CEP}, method 2

Figures 4 and 5 show the d_{CEP} obtained with methods 1 and 2 : the continuous lines correspond to the distance between the clean speech and the noisy speech. The dashed lines correspond to the distance between the clean speech and the enhanced speech.

To finish, we may introduce a remark which can improve the performances of the two methods we suggest : this improvement consists in using more than two observations ; we can compute the cross-correlations between the successive channels (x_i) ; all coherence functions between available observations are evaluated and averaged in the following manner :

$$\bar{\rho}(f) = \left(\sum_{i=1}^{M} \sum_{j=i+1}^{M} \rho_{x_i x_j}(f) \right) \Big/ \sum_{i=1}^{M} (M-i)$$

- The averaging allows us to estimate a better filtering in the first structure or to perform a better decision criterion in the second one.
- Moreover, when a preprocessing is made to obtain identical signals on each channel : the average $\bar{\rho}(f)$ may be used to weight $\frac{1}{M}\left(x_1 + \sum_{i=2}^{M} x_i^f\right)$ (in the frequency domain) in the first structure ; as for the second structure, the spectral subtraction may be realized on each channel. It allows to increase the SNR.

6. CONCLUSION

To conclude we may acknowledge that the use of the coherence function for both methods leads to a speech signal more intelligible and more pleasant to listen. The main advantage of these methods is the following : we don't suppose noise stationarity ; the results relative to the second method are quite encouraging since the noise spectra are well learned on short time intervals and the disturbing noise is completely removed even if it sometimes remains a very slight musical noise. The only drawback lies in the restrictive hypotheses : first of all, noises must be decorrelated if we want to obtain a good recognition system and secondly, it's necessary to get at least two channels to compute the coherence function.

Acknowledgement : The authors wish to thank the CNET Lannion A for supporting this work.

REFERENCES

[1] Allen, J.B. and Rabiner, L.R., A Unified Approach to Short-Time Fourier Analysis and Synthesis, Proceedings of the IEEE, vol. 65, n° 11, pp. 1558-1564, Nov. 1977.

[2] Berouti, M., Schwartz, R. and Makhoul, J., Enhancement of Speech Corrupted by Acoustic Noise. Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing, pp. 208-211, April 1979.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A Multiframe Spectral Weighting System for the Enhancement of Speech Signals Corrupted by Acoustic Noise

A.F. Erwood, C.S. Xydeas
Speech Processing Research Group, Department of Electrical Engineering,
University of Manchester, Manchester M13 9PL, England.

A novel acoustic noise reduction technique based on spectral weighting is presented in this paper. The proposed system employs a variable analysis interval and offers significantly better noise reduction performance than conventional spectral weighting algorithms.

1. Introduction

Speech transmission from mobile environments is usually accompanied by significant levels of accoustic noise. The effect of adding noise to a speech signal is to degrade its quality and eventually its intelligibility. Furthermore, the additive noise present with the speech signal usually impairs the performance of speech processing systems such as speech coders and speech recognisers.

In recent years, much research has been carried out into various techniques for active noise reduction. Such systems process signals from one or more microphones while aiming to preserve the quality and intelligibility of the speech signal. When one or more 'reference' signals are available in addition to the composite speech and noise signal from the main microphone, coherent adaptive noise cancellation (ANC) schemes can be employed [1]. Such a system attenuates the signal components from the main microphone which are correlated with those from each of the reference microphones. Thus a significant improvement in SNR can only be achieved when the cross-correlation between the noise components of the main microphone and those of the reference microphones are strong, while at the same time the cross-correlations between the corresponding speech components are weak. Unfortunately, this desirable condition does not prevail within mobile type environments and consequently a significant improvement in SNR cannot generally be achieved by this approach [2].

Over the last decade, noise reduction schemes based on spectral weighting (SW) have been proposed for acoustic noise reduction [3,4,5]. In these schemes, noise reduction is achieved without the need for a separate noise reference input. In each case, selective weighting of discrete lines (bins) in the successive short term spectra of the input signal is applied with the aim of attenuating noise energy while preserving the wanted speech signal. Since the bin weights must be made a function of both the short term input spectrum and the mean noise spectrum, the latter is usually derived from the input during periods of zero speech activity.

In both fixed and mobile environments, SW noise reduction schemes compare favourably with ANC schemes in terms of SNR increase, However, in traditional SW schemes performance is limited by the short analysis interval applied to the input signal and the consequential uncertainty in the observation of the noise component. In the system described here, named Multiframe Spectral Weighting Noise Reduction (MFSWNR) System, a variable length analysis interval is employed which offers superior noise reduction performance while minimising degradation of the speech signal.

2. Noise Reduction by Spectral Weighting

The input signal to a spectral weighting noise reduction system is assumed to originate from a single microphone and contains both speech and additive environmental noise. In a typical SW arrangement, the input signal is digitised and multiplied by a window function, usually a raised cosine form, which thus defines a frame of data. The discrete short-term spectrum expressed in magnitude and phase is then computed via the Fast Fourier Transform (FFT). A weighting filter separately attenuates each discrete line (bin) of the magnitude spectrum to yield the estimated speech magnitude spectrum. A common form of SW is spectral subtraction which may be described by:

$$\left.\begin{aligned} |\hat{S}_i| &= |X_i| . W_i && 1(a)\\ \hat{\theta}_i &= \Phi_i && 1(b)\\ W_i &= 1 - \frac{|\hat{\bar{N}}_i|}{|X_i|} && 1(c)\end{aligned}\right\} \quad i=0,1,\ldots,L/2$$

where $|X_i|$ and Φ_i are the magnitude and phase repectively of the i^{th} bin of the input signal, $|\hat{S}_i|$ and $\hat{\theta}_i$ are the estimated magnitude and phase of the speech signal, W_i is the spectral weighting coefficient and $|\hat{N}_i|$ is the estimated mean noise magnitude of that frequency and is

derived from $|X_i|$ during intervals when there is no speech activity. The frame size is L samples. The amount of attenuation depends on the signal to noise ratio in each bin and ranges from zero for bins containing speech energy but no noise energy up to some maximum for bins containing only noise. Further attenuation may also be applied to all bins in those frames which do not contain any speech energy. The estimated spectrum is transformed to the time domain via the inverse FFT. Inverse windowing and concatenation are applied to the resulting sequence of frame waveform segments to yield the estimate of the speech waveform.

3. Residual Noise Generation

In all spectral weighting schemes, some noise, termed the 'residual noise', is present in the speech spectral esitmate $\{|\hat{S}_i|\}$. The mechanism by which this residual noise is generated can be explained by considering, as an example, the effect of the simple weighting function of equation 1(c) operating on a sequence of frames which contain only noise energy. To simplify the discussion only the i^{th} bin will be used, though the arguments are applicable to any and all of the bins.

In the absence of speech, $|X_i| = |N_i|$ and each $|N_i|$ is a Rayleigh distributed random variable with mean value $|\bar{N}_i|$ and so the weighting coefficient W_i and estimated magnitude spectrum $|S_i|$ are given by:

$$W_i = 1 - \frac{|\hat{\bar{N}}_i|}{|X_i|} = 1 - \frac{|\hat{\bar{N}}_i|}{|N_i|} \qquad 2(a)$$

$$|\hat{S}_i| = |X_i| W_i \; . = |N_i| - |\hat{\bar{N}}_i| \qquad 2(b)$$

$\{|\hat{\bar{N}}|\}$ is the estimated mean noise magnitude spectrum, and is assumed to be close to the actual mean noise magnitude spectrum $\{|\bar{N}|\}$ of the noise process. The variance of $|N_i|$ is large compared to that of $|\hat{\bar{N}}_i|$ because the duration of the input signal from which it is derived is much shorter (typically 10 to 30 msec) than that employed for $|\hat{\bar{N}}_i|$ (depends on the duration of intervals in which there is no speech activity but typically 100 to 3000 msec). The effect of the weighting coefficient of equation 2(a) is that while $|\hat{S}_i|$ has lower mean square value than $|N_i|$, $|\hat{S}_i|$ still exhibits significant random variation from frame to frame and this is manifest as residual noise at the system output.

In those bins where the SNR is low, the variance of $|\hat{S}_i|$ can largely be attributed to that of $|N_i|$ and may be reduced by extending the duration of the analysis frame. However, in doing so the stationarity assumption for any other bins in the same frame which contain significant speech energy could be destroyed. It is thus apparent that the choice of frame duration involves a compromise between noise attenuation and the fidelity of the short term speech spectrum. This conflict exists in all single frame systems and stems from the fact that the frame duration is invariably constant and common to all bins regardless of the relative speech and noise content.

4. Multiframe Spectral Weighting

In the MFSWNR system, the effective duration of the analysis frame is chosen for each bin according to its SNR. Variation in effective frame duration is achieved by temporal averaging for each bin accross a suitable number of consecutive frames each of which has a fixed duration.

The general arrangement of the MFSWNR system is shown in figure 1. The composite speech and noise signal from the single microphone is band pass filtered, sampled, digitised and segmented into half overlapping frames each multiplied by a Hanning window. The discrete spectrum for each frame is computed via the FFT and expressed in magnitude and phase form. An odd number, K, of magnitude spectra form the input to the spectral weighting process from which the magnitude spectrum of the speech signal corresponding to the frame at position K/2 + 1 is estimated. This discrete spectrum is combined with saved phase from the corresponding frame of the input signal and transformed to the time domain via the inverse FFT. Inverse windowing is achieved by addition of corresponding samples of the overlapped portions of adjacent frames.

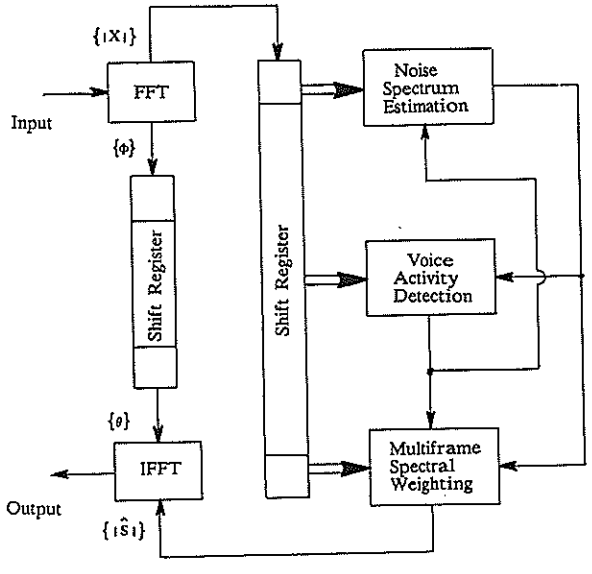


Figure 1. MFSWNR System Block Diagram

In the MFSWNR system, spectral weighting is applied to each discrete spectral line independantly of the others and comprises of several separate stages as follows:

i) SNR estimation. An estimate of the SNR of the input signal is obtained according to:

$$SNR_I = \left[\sum_{j=1}^{k} |X_{i,j}| . V_{j,M} \right] / |\hat{\bar{N}}_i| \quad i=0,1,\ldots,L/2 \qquad (3)$$

where $|X_{i,j}|$ is the magnitude spectra for each of k frames, $V_{j,m}$ is a set of M weighting functions described by:

$$V_{j,m} = \gamma_m . \exp - ((K/2 + 1 - j)^2 . 2^{(m-k)}) \qquad 4(a)$$

and γ_m determined such that:

$$\sum_{j=1}^{k} V_{j,m} = 1 \qquad 4(b)$$

j denotes the frame number, i is the number of the discrete spectral line and $|\hat{\bar{N}}_i|$ is the estimate of noise magnitude. The SNR estimate is thus of the form (signal + noise)/noise.

ii) Multiframe Averaging. The purpose of the multiframe averaging is to extend the effective frame duration according to the estimated SNR of the input. The result of the averaging process is an improved estimate of the noisy speech signal and is formed according to:

$$|\hat{X}_i| = \sum_{j=1}^{k} |X_{i,j}| . V_{j,m} \qquad (5)$$

The variable m varies from 1 (maximum temporal averaging) for low input SNR (SNRI ≈ 1) upto M (minimum temporal averaging) for high input SNR (SNRI > ≈ 3.2 ≡ 10dB). For signal to noise ratios greater than 10dB, $|X_i|$ takes the value of the centre frame, that is:

$$|\hat{X}_i| = |X_{i,K/2+1}| \qquad (6)$$

iii) Spectral Weighting. The type of weighting law applied to the noisy speech estimate is spectral subtraction so that the initial estimate of the speech signal is:

$$T_i = |\hat{X}_i| - |\hat{\bar{N}}_i| = |\hat{X}_i| . W_i \qquad 7(b)$$

where $|T_i|$ is spech estimate before post processing is applied and the weighting function, W_i, is given by:

$$W_i = 1 - \frac{|\hat{\bar{N}}_i|}{|\hat{X}_i|} \qquad 7(c)$$

In instances where the SNR is low, the estimate $|\hat{X}_i|$ tends to be closer than $|X_i|$ to the estimated noise $|\hat{\bar{N}}_i|$ and thus W_i tends to be nearer to zero and greater noise attenuation is achieved.

5. Post Processing

In order to further improve the subjective quality of speech signal, residual noise attenuation and comfort noise addition are applied after spectral weighting.

i) Residual Noise Attenuation. It has been established through infomal listening tests that the speech signal possesses a masking property in that the residual noise is less apparent when speech is present. Thus during periods when there is no speech activity, attenuation of the residual noise signal can be applied advantageously. This is achieved by weighting each bin in the estimated speech spectrum according to:

$$|\hat{S}_i| = |T_i| \ (\alpha . \beta + (1-\alpha)) \qquad (8)$$

where α is a constant which controls the level of attenuation and β is the output of a first order low-pass filter with a time constant of one half the frame duration and having a binary level input (0 or 1) derived from the voice activity detector (VAD).

ii) Comfort Noise Addition. Depending on the nature of the noise at the system input and on the level of residual noise attenuation, listening tests have indicated that the addition of some low level 'comfort noise' can be subjectively beneficial. In the MFSWNR system, the comfort noise is derived by filtering a zero mean, gaussian white noise source to have the same mean spectral shape as the input noise and adding at a suitable level it to the estimated signal after residual noise attenuation.

6. Voice Activity Detection

Voice activity detection is required for two reasons: i) it determines those frames which may be used to update the estimate of the average noise magnitude spectrum and ii) it enables residual noise attentuation to be applied when there is no speech present. A detailed description of the VAD is beyond the scope of this paper, however, its essential properties can be summarised as follows:

i) Binary result for each frame.
ii) Decision based on comparisons of the frame energy with several dynamic thresholds.
iii) Noise reduction enhancement for each frame making use of the mean noise magnitude spectral estimate prior to energy calculation.
iv) Short term energy histogram used to set thresholds.
v) VAD result biased towards indicating speech present.

7. Noise Spectrum Estimation.

The estimate of the noise spectrum is derived from the spectrum of the input signal during periods when there is no speech activity. In order to generate the required estimate, the Noise Spectrum Estimator (NSE) makes use of the sequence of magnitude spectra of the frames of the input signal and also the VAD output. When the VAD indicates a speech activity for a particular frame, the discrete spectrum of that frame is input to a band of non-linear first order low-pass filters, the outputs being the required estimate of the average noise spectrum. This filtering process is described by:

$$|\hat{\bar{N}}_i| = |\hat{\bar{N}}_i|.c_i + |X_i|(1-c_i) \quad i=0,1,\ldots,L/2 \tag{9}$$

where $\{|N_i|\}$ is the required estimate, $\{|X_i|\}$ is the input spectrum and $\{c_i\}$ is the set of filter coefficients. When the VAD output indicates the presence of speech, $|N_i|$ is not updated but is held until the VAD again indicates that no speech is present. The use of a bank of first order filters provides a convenient way to optimise the noise spectrum estimate to obtain the best bpossible residual noise performance from the spectral weighting process when the additive noise spectrum is non-stationary. The objective is always to obtain the smoothest (in a temporal sense) possible noise spectral estimate consistent with that estimate tracking changes in the actual noise spectrum. The required time constant for the filters is determined taking each bin in turn according to the normalised error between the noise spectral estimate and the average of the spectra of several frames adjacent to and including that indicated by the current VAD output.

8. Results and Conclusion

The MFSWNR system has been evaluated by computer simulation. Results based on informal listening tests indicate that this system achieves significantly better performance than single frame schemes in terms of noise reduction for given recovered speech quality. Graphical results (figures 2,3) also suggest that a significant level of noise reduction is achieved while preserving important spectral detail of the speech signal.

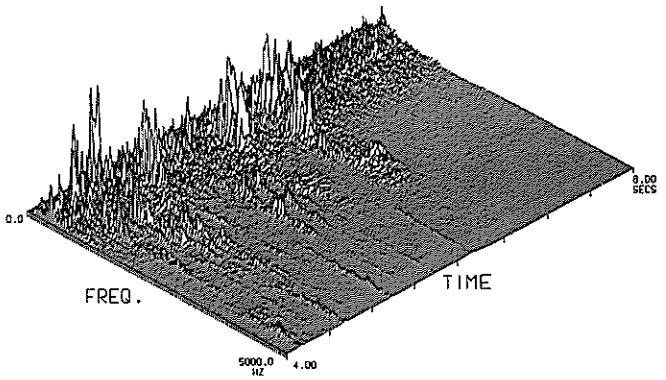


Figure 2. Magnitude spectrum of the speech and automobile noise input

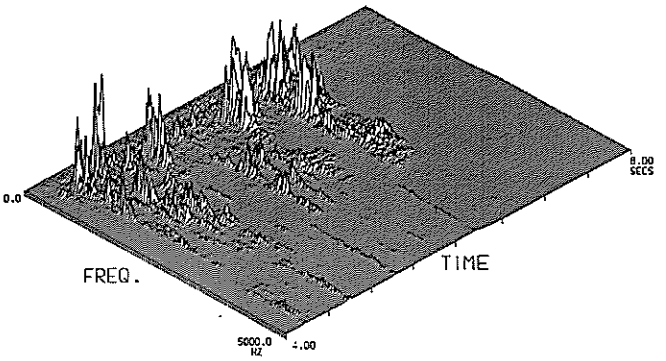


Figure 3. Magnitude spectrum of MFSWNR system output.

References

1. 'Adaptive Noise Cancelling: Principles and Applications' Widrow, B. et al., Proceedings of the IEEE. Vol.63, No.12, Dec. 1975. pp 1692-1719.

2. 'Adaptive Noise Cancellation with Reference Input-Possible Applications and Theoretical Limits', Armbruster, W., Czarnach. R., Vary. P., Eurasip. Signal Processing III. 1986 pp 391-394.

3. 'Suppression of Acoustic Noise in Speech Using Spectral Spectral Subtraction', Boll, S.F., IEEE Trans. ASSP. Vol. ASSP-27 No.2, April 1979. pp 113-120.

4. 'Reduction of Acoustic Noise in Speech Signals by Pre-Processing and Spectral Subtraction', Erwood, A.F., Xydeas, C.S., IERE Int. Conf. Dig. Proc. Sig. April 1981 pp 347-355.

5. 'Noise Suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits', Vary, P., Sig. Processing. No.8. 1985, pp. 387-400.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Trainable noise substraction filters for speech enhancement in the car

Laurent BARBIER* **,Chafic MOKBEL*, Gérard CHOLLET*

*ENST Dept. SIGNAL, CNRS URA-820 46 rue barrault, 75634 PARIS cedex 13, FRANCE
**EDF DER dept. AMV,1 avenue du général De Gaulle, 92141 Clamart, FRANCE
(phone : 33 1 45 81 75 47 ; fax : 33 1 45 89 79 06 ; e-mail : barbier@sig.enst.fr)

Abstract: Two different non linear approaches: neural network filtering and adaptive Kalman filtering are compared in order to achieve noise reduction of speech signals in the car.

1: Introduction.

Mobile radio and general purpose computers are available in cars. For safety purpose, the driver may need to use speech input to control the new devices using a microphone located on the dash-board. This microphone gets a very noisy signal with a S/N ratio varying between 10 and -10dB. Our goal is to filter out the noise. Two different techniques are compared : adaptive Kalman filtering and neural network filtering. In order to train the neural network, a close-talking reference microphone is being recorded synchronously with the dash-board one.

2: Analysis of the signals:

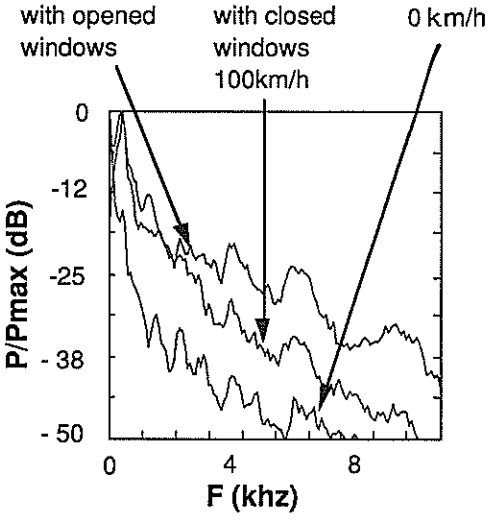
2.1: Speech database:

The speech database consists of 4 isolated words repeated 3 times in the following conditions:
BMA : 0 km/h, engine running.
BFFO: 90 km/h, ring road, opened windows.
BFF1: 100 km/h, motorway, closed windows.

2.2: analysis of the noise:

To evaluate the degradation of these signals, we have done a spectral analysis in the different noise conditions. The average spectrogram is computed using 125 frames of 256 samples (with an overlapping part of 128 points and a 16 khz sampling frequency that's about 1 second of noise). In order to plot all these curves in the same graphic, each one has been normalised with his maximum value.

Fig 1: power distribution of noise



We can see in all these spectra that we have a peak in low frequencies (about 150 hz) which corresponds to the fundamental frequency of the engine. The spectrum then decreases quasi exponentially, and after 4 kHz the power is very low. If fact, when the speed increases or when we open the windows, the noise power in higher frequencies (above 150 hz) increases and also the S/N ratio decreases (from 15 dBs in the BMA case to -10 dBs in the BFFO case).

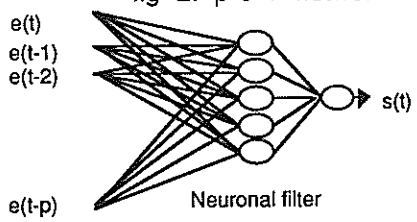
3: Neural Network Filtering:

3.1:Principle:

A neural network is being used to filter out the noise. Earlier studies has been made on spectrograms using architectures capable to realize a mapping function [3][5]. For that

application, we use also a multilayered feed-forward network composed of one or two levels allowing us to use the well known Back-Propagation Algorithm [4]. Several improvements have been done on this algorithm: adaptation of the learning parameters [1] and New error [2]. In our experiments, we use the time sampled signal. The input is composed of p+1 successive samples of noisy speech signal from time t-p to time t, and the ouput is the noise-free speech signal at time t . At time t+1, this window is moved of one sample. So, if N is the number of samples (ex N=10000), we need about N*P computations to get the total gradient.

fig 2: p*5*1 network



3.2: Selection of the network:

The optimal number of input nodes was first investigated. Our criteria were the error rate after convergence and the computation time. P=100 was found a good compromise. Several networks structures were then studied. Different trials have been made: (100*1), (100*2*1), (100*4*1), (100*8*1). After convergence, these four networks trained on the same data (four seconds of speech and noise) gave approximatively the same error rate. We also tried a (30*30*30*1) network and found a higher error rate. So, for our experiments, we chose the (100*1) network.

3.3: Experiments:

A network was trained for each word using 3 repetitions of this word for learning base and about 15 iterations until convergence. So we get 4 networks (1 per word). The results are obtained by filtering the first set of the 4 words N1 using these 4 networks.

3.4: Results:

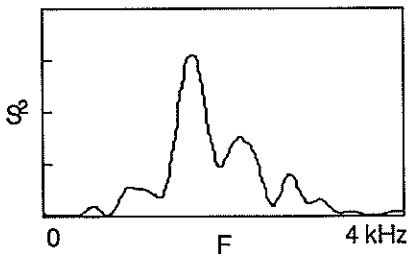
One way to evaluate the performance of this method is to use a measure of S/N ratios. Thse were computed as R=(PS-PN)/PN were PS is the power in the speech segment and PN the power of the preceeding noise segment.

table 1: S/N ratios in dBs

signals \ words	1	2	3	4	glob.	impr.
S1	45.1	55.2	30.5	56.0	46.7	
N1	13.6	6.3	13.0	17.4	12.6	
F1	25.8	27.5	22.9	33.3	27.4	14.8
F2	25.1	31.6	22.4	31.1	27.6	15.0
F3	26.6	29.0	23.5	28.7	26.9	14.3
F4	25.3	29.3	24.9	26.1	26.4	13.8

In table 1, the first row concerns the reference signal S1 and the second the noisy signal N1. Then, each row corresponds to the filtering of N1 by the 4 different networks trained on different words. We get for all cases an improvement of 13 dBs. The generalization seems to be good because filters trained on one word obtained identical results on the other words. To analyse how our filters work, we have computed the spectral response of filter 1. Peaks may correspond to speech formants. The frequencies below 1 kHz are attenuated. The network may learn the longterm spectral characteristics of both noise and speech.

Fig 3: frequency response of F1 filter



4:Adaptive Kalman Filtering:

4.1:Introduction:

In the second part of this paper we develop a more traditional approach to the problem of speech enhancement. To solve it several techniques are used : EM algorithm, MAP estimation, Kalman filtering...[7-10]. Numerous studies on the application of Kalman filtering in speech enhancement involves additive white or colored noise[8-10]. Tests done on simulated data were reported satisfactory. Here we present the results of some tests done on real data recorded in the car environment.

4.2:Principle:

The filter presented has been described several times in the literature[8][9]. In the following, we make the following assumptions :

i) Noise is additive and stationnary in a word context.

ii) Noise is colored and could be represented by an AR model (order 4).

iii)Speech signal could be represented by an AR model of order 10.

Assuming these hypothesis, the adaptive filter is defined as follows:

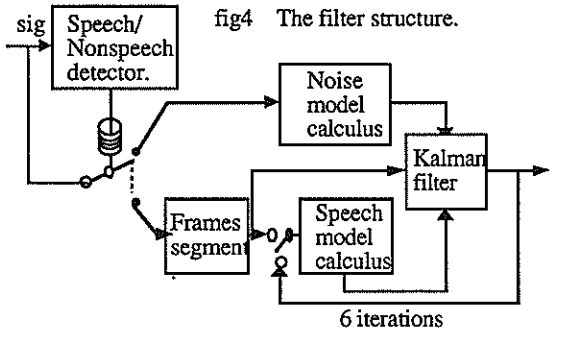


fig4 The filter structure.

A speech/Nonspeech detector defines the noise regions where a noise model could be computed. In the speech region a frame by frame analysis is performed. For each frame a speech model is first calculated. It will be used to filter this frame by the Kalman algorithm using the noise model. When filtered, a new estimation of the speech frame model is performed which will be used to filter the initial frame. This is done iteratively until the residuel of the speech model becomes less than a certain threshold. To discriminate Speech/Nonspeech regions, the Loglikelihood between progressing and sliding windows is calculated using ladder algorithms[6]. When this Likelihood comes over a certain value a rupture detection is retained. Endpoints ruptures are located using temporal constraints. This method has given good results and we have been able to detect speech words when S/N ≈ -5dB.

To filter, the algorithm presented in [8] is computed assuming colored noise hypothesis. The state equations could be written as :

$$\begin{bmatrix} x(n) \\ v(n) \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} x(n-1) \\ v(n-1) \end{bmatrix} + \begin{bmatrix} c & 0 \\ 0 & c_v \end{bmatrix} \begin{bmatrix} w(n) \\ \beta(n) \end{bmatrix}$$

$$s(n) = [\, d^T \; d_v^T] \begin{bmatrix} x(n) \\ v(n) \end{bmatrix}$$

Where s(n),v(n) and x(n) are respectively the noisy signal, the noise and the clean signal and A and B the prediction matrixes.

These equations could be presented :

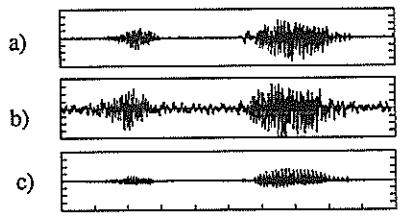
$$X(n) = A\,X(n-1) + C\,W(n)$$

$$s(n) = D^T X(n)$$

The solution of this canonical model is given by the Kalman algorithm. In each iteration step, A is updated using the filtered frame x(n) and the initial frame s(n) is refiltered.

4.3:Results:

The database described in part 2 is used in the tests. Fig 5 shows the results of filtering the word "Lipka". In this figure, are represented from top to bottom: the filtered word, the noisy word and the reference word recorded by the close-taking microphone. The first results show that this filter eliminates all the noise in the silence zones. Unfortunutaly, a lot of distorsions are introduced to the signal. This aspect is discussed in the next chapter. It was also noticed that after about 8 iterations the speech frame converges to a perfect AR model.



a) filtered signal. b)noisy signal.
c) clean signal. The word here is "LIPKA"
fig 5

5 : Comparisons between the two methods:

If we look at the signals in terms of S/N ratio, the two methods tested show a good ability to filter out the noise. But, because we are interested in speech, their real performance must be evaluated in terms of spectral distorsion or recognition rates. To compare them, we have computed the Log Spectral Deviation defined as the L2 of the Log of the normalized smoothed spectra. The results in table 2 were obtained.

noisy signal	Kalman Technique				Neural approach
	1° Iter.	2° Iter.	3° Iter.	4° Iter.	
13.8	24.2	14.23	26.31	18.53	19.81

Table 2. Evaluation of distorsions respect to the clean references recorded by the close talking micro.

This table shows clearly that both techniques introduce a lot a distorsion on the speech signal. The frequency response of the neural network approach explains what append in that case (fig 3). All frequencies under 1 kHz are cut off. That corresponds to the first two formants of speech. For the Kalman technique, approximating the speech model by a AR one and considering the noise as additive are probably the main sources of distorsions. The algorithm used here maximizes at every iteration the joint probability density (JPD) of the speech and the noise knowing a model of the noise and the noisy frames. Starting the iterations with a model extracted from the noisy signal makes the filter converge to a local maximum of the JPD. Starting with a more robust estimation of the speech frame model consists a better solution.

6: Conclusions:

The goal was to enhance the speech signal from noise. Two different approaches have been compared in the context of real application. Althought similar techniques were tested successfully on synthetic signals, the test on real signals are unsatisfactory. Despite significant improvement in terms of S/N ratio, the two techniques introduce large spectral distorsions wich make them unsuitable in this form for coding and recognition applications.

References :

[1] L.W Chan, F. Fallside. "An adaptive algorithm for back propagation networks", Computer Speech and Language, 1987.
[2] P. Haffner, A. Waibel, H. Sawai and K. Shikano. "Fast Back-Propagation Learning Methods for Neural Networks in Speech.
[3] R. P Lippman, "introduction to computing with neural nets",IEE ASSP magazine pp 4-22, April 1987.
[4 D.E Rumelhart, J. Mc Clelland and the PDP Research Group, Parallel Distributed Processing, Vol 1, Chap 8, MIT press, 1986.
[5] S. I Tamura and A. Waibel, "Noise reduction using connectionist models", IEEE 1988.
[6] Brandt A.v., "Detecting and estimating parameter jumps using ladder algorithms and likelihood ratiotests.", IEEE-ICASSP, pp 1017-1020, 1983.
[7] Feder M. and Oppenheim A.V., "Methods for noise cancellation based on the EM algorithm", IEEE-ICASSP, Vol S, pp 557-560, 1988.
[8] Koo B., Gibson J.D. and Gray S.D., "Filtering of colored noise for speech enhancement and coding", IEEE-ICASSP, Vol S, pp 349-352, 1989.
[9] Morikawa H. and Fujisaki H., " Noise reduction of speech by adaptive Kalman filtering", Traitement du signal, Vol 22, pp 53-68, 1988.
[10] Paliwal K.K. and Basu A., " A speech enhancement method based on Kalman filtering", IEEE-ICASSP, Vol 1, pp 177-180, 1987.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

MISSING PACKET RECOVERY OF LOW-BIT-RATE CODED SPEECH USING A NOVEL PACKET-BASED EMBEDDED CODER

M. M. Lara-Barron, G. B. Lockhart

Department of Electronic and Electrical Engineering,
University of Leeds, Leeds, LS2 9JT, England.

Missing packets due to network congestion can seriously degrade the quality of received speech in voice packet communication. A variety of techniques exist which attempt to reconstruct the missing speech segments but these methods cannot be directly applied to low bit rate encoded speech such as ADPCM where the decoded output depends not only on current transmitted codewords but also on the past history of the decoder. In this case decoder parameters at the receiver lose track of encoder parameters after a missing packet leading to propagation of the decoding error into the following packet. A novel packet-based embedded coding scheme will be described which mitigates the effects of parameter mistracking caused by packet loss. The method may be applied to any low bit rate coder that utilises past information to decode current samples but specific results are given for its application to the CCITT 32 kbit/s ADPCM coding standard.

1. INTRODUCTION

Packet technology has much potential for the integration of different type of services in a single communication network [1]. While information integrity is normally the most important performance criterion for data services, real-time voice communication requires packets to be delivered to the receiver with minimum overall end-to-end delay. In this application some reliability may be sacrificed provided speech quality is not significantly degraded. Total delay in packet networks comprises a fixed component due to packetisation and propagation, and a variable delay due to buffering at network nodes. Under heavy load conditions, queues at the network nodes can build up to the extent that some packets must be purposely discarded in order to reduce mean packet delay across the network to within acceptable limits [2]. When a speech packet is discarded or is otherwise absent from the receiver, some strategy must be invoked to fill the gap left by the missing packet.

A number of missing packet recovery strategies for PCM encoded speech have been investigated and include, reading out silence, repeating the previous received packet, or deriving a substitute waveform from previous correctly-received speech [3]. Such strategies, however, cannot be applied without modification to low-bit rate encoded speech such as ADPCM or to predictive schemes in general where the decoded output depends not only on current transmitted codewords but also on the past history of the decoder. In this case, decoder parameters at the receiver lose track of encoder parameters after a missing packet and the decoding error propagates for a number of samples into the following packet [4,5]. Parameter mistracking may be prevented by sending parameter information at the beginning of each packet but this requires additional channel capacity with a consequent reduction in network efficiency. Another approach is the use of variable-rate coding as a means of matching the bit rate of each speech source to the prevailing network capacity. This may be an adequate strategy for small networks where the encoder can rapidly detect and respond to transient overloads but the long delays involved in the feedback of network state information prevents its use in larger networks.

A novel packet-based embedded encoding scheme will be described which mitigates the effects of parameter mistracking. No parameter information is transmitted and no network state feedback is necessary.

2. PACKET-BASED EMBEDDED CODING

Fig. 1(a) shows a segment of voiced speech with a single 16 ms packet missing at the receiver. With PCM encoding, a substitute for the missing packet can be obtained from the previous correctly-received packet (packet 1) by one of a number of waveform substitution methods [3,6] as illustrated in Fig. 1(b). The following received packet (packet 3) can then be decoded normally. If differential or predictive coding is used packets can be reconstructed in the same way from past decoded samples but, in the absence of correctly updated parameter information, decoder parameters at the receiver will lose track of encoder parameters leading to propagation of the decoding error into the following packet. This is illustrated in Fig. 1(c) for 32 kbit/s ADPCM coded speech, where parameter mismatch at the end of the reconstructed packet (packet 2) produces distortion in packet 3 due to error propagation, although packet 3 is correctly received.

Embedded encoding allows a bit rate reduction at any point in a network without affecting the essential operation of encoder and decoder. For example, an embedded version of DPCM [7] generates two distinct bit streams: one "essential" and the other "sup-

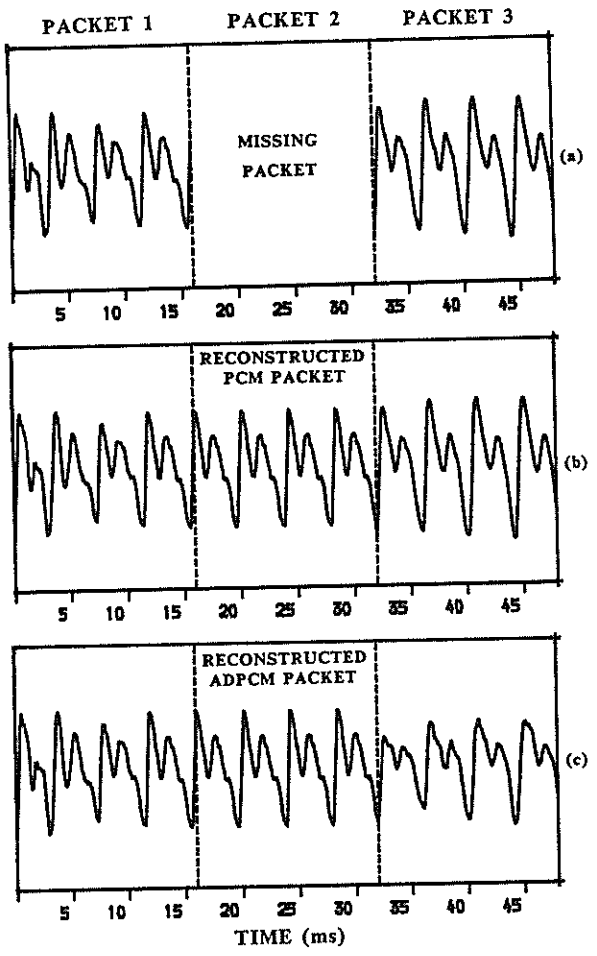


Figure 1 Missing packet reconstruction with PCM and ADPCM coded speech.

plementary". Provided the essential stream is correctly received the decoder will operate correctly with speech quality commensurate with DPCM at the bit rate of the essential stream. However, quality is improved if part or all of the supplementary bit stream is received. This type of embedded strategy has been successfully applied to information discarding in packet networks [2,4,8] but unless essential and supplementary bits are transmitted in separate packets, incurring additional packetisation delay, relatively complex packet disassembly/assembly operations are required at an overloaded node in order to delete the supplementary bits.

In the proposed scheme, illustrated in Fig. 2, the principles of embedded coding are applied at a packet level so that in the event of node overload, whole packets may be discarded instead of individual bits within the packet. Both transmitter and receiver attempt to predict the next packet on the basis of previously decoded speech. At the transmitter, the current packet is compared with its prediction to determine the degradation which would result from reconstruction at the receiver and labelled with a "reconstruction index", R [9]. The reconstruction index is then compared to a fixed threshold for classification of packets as "essential" (difficult to reconstruct) or "supplementary" (easy to reconstruct). Switches SW1 and SW2 in Fig. 2 are held in position "E" or "S" after an essential or supplementary packet respectively. Both types of packets are low bit rate encoded and transmitted through the network in the normal way but after a supplementary packet the encoder parameters are updated by locally encoding the predicted packet instead of the original packet (i.e. SW2 is closed). In this case, the predicted packet is also fedback into the packet predictor in order to predict the next packet. In the network, supplementary packets are preferentially discarded by overloaded nodes. At the receiver, essential packets are decoded normally. Switches SW3 and SW4 are held in position "E" when an essential packet is received and in position "S" when a

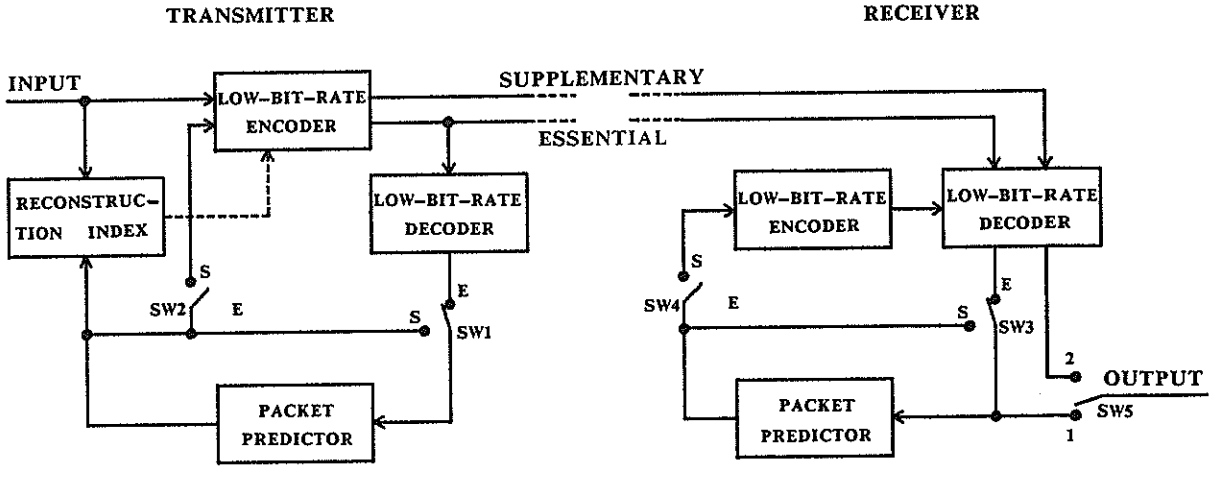


Figure 2 Packet-based embedded encoding of low-bit-rate coded speech.

missing packet occurs or a supplementary packet is received. Switch SW5 remains in position "1" unless a supplementary packet is present at the decoder in which case it is set to position "2". A local encoder is necessary at the receiver for parameter updating. When a missing packet occurs, the predicted packet is output but it is also locally encoded to update the decoder parameters. If a supplementary packet is received it is decoded and output normally but decoder parameters for the next packet are kept in step with encoder parameters at the transmitter by locally encoding the reconstructed packet as if the supplementary packet was missing.

Supplementary packets therefore play the same role as the variable rate stream in a conventional embedded coding system. Parameter conditions at the receiver and transmitter are kept in step provided only supplementary packets are discarded in the network but when possible, supplementary packets are used at the decoder to increase output speech quality. There is a penalty due to less accurate sample prediction in the encoder following a supplementary packet but efficient voice packet classification at the transmitter will ensure that the effects of parameter mismatch following a supplementary packet are small compared with severe mistracking errors which arise when packet prediction is performed only at the receiver. It is important to note that parameter mismatch will occur after every supplementary packet regardless of whether the packet is or is not actually lost. Such a tradeoff is a well-known property of embedded encoding schemes [7]; a small penalty is tolerated at maximum information rates in order to improve performance when information is lost.

3. SIMULATION OF A PACKET-BASED EMBEDDED ADPCM CODER

The system illustrated in Fig. 2 was simulated in order to assess the performance of a packet-based embedded encoder. The input speech consisted of 12 sec. of recorded speech comprising 6 sentences spoken by male and female speakers. The speech is sampled at 8 kHz, initially converted into 12-bit linear PCM codewords and packetised in 128 samples corresponding to 16 ms speech segments. The encoder and decoder operate according to the CCITT 32 kbit/s ADPCM standard [10]. The current packet is predicted at both the transmitter and the receiver based on past ADPCM decoded samples using a pitch replication waveform substitution method [6]. At the transmitter, the current and predicted packets are compared in order to determine the reconstruction index R and the current packet is labelled accordingly. R is calculated according to the signal to distortion ratio (SDR) computed on a mean square basis. In order to reduce the occurrence of consecutive missing packets every other packet is unconditionally labelled with a low value of R [9].

The reconstruction index R of each packet is compared to a SDR threshold at the transmitter which differentiates between essential and supplementary packets. If R is above threshold the packet is classified as supplementary: otherwise as essential. Both types of packets are ADPCM encoded and sent to the network for transmission. Packet discarding at network nodes was simulated by removing all packets with labels exceeding specific values of R, and determining the corresponding average SDRs and missing packet rates. Reception of essential packets is guaranteed when this value of R is larger than the SDR threshold. Fig. 3 shows the performance of packet-based embedded ADPCM coder for different values of SDR threshold in comparison with random discarding of ADPCM encoded packets simulated by discarding packets at random for fixed probabilities between 0 and 0.5.

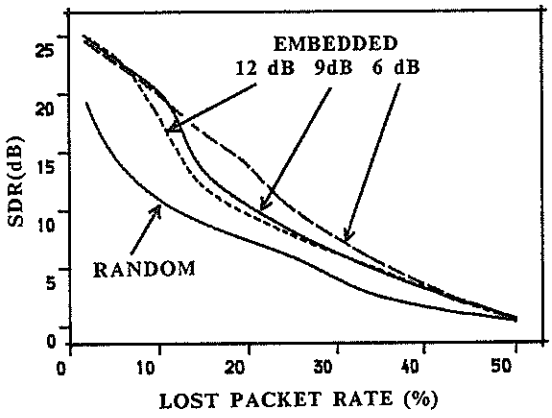


Figure 3 SDR for packet-based embedded ADPCM encoding.

The three curves of Fig. 3 corresponding to SDR thresholds of 12 dB, 9 dB and 6 dB respectively show a significant improvement over random discarding. With a 12 dB threshold the ratio of supplementary to essential packets is 1:15 and the SDR for zero or low lost packet rates is relatively high. In this case the only significant distortion derives from the reconstruction of the missing packets at the receiver since the effects of encoding a small number of easily reconstructed packets with sub-optimum parameters is not significant. However, if lost packet rates increase beyond 6%, the 12 dB curve drops more rapidly since supplementary packets have been exhausted and the loss of some essential packets produces mistracking error in addition to the normal reconstruction error. With 9 dB and 6 dB thresholds the ratio of supplementary to essential packets becomes 1:10 and 1:5 respectively, and the onset of mistracking error is deferred to missing packets rates of about 10% and 20% respectively. The penalty for improved performance at higher lost packets rates is therefore an increase in the distortion at zero and low rates due to encoding a larger number of less easily predicted packets with sub-optimum parameters. Thus a tradeoff exists between performance at low and high missing packet rates which may be modified by adjusting the SDR threshold at the transmitter according to the overload characteristics of the network.

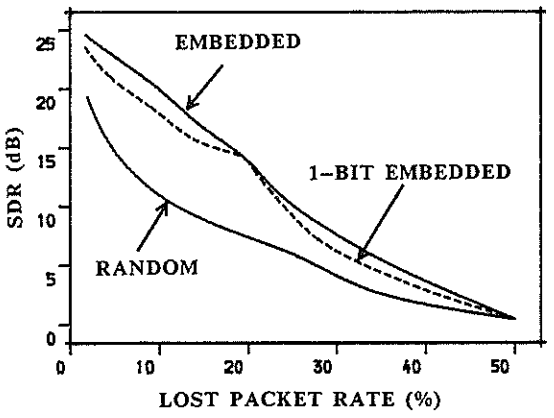


Figure 4 SDR for simplified packet-based embedded ADPCM encoding.

Subjectively, speech quality is always superior using packet-based embedded coding rather than random discarding. While the reconstruction of randomly discarded ADPCM coded speech packets permit up to about 5% of packets to be discarded with minimum impact on voice quality, packet-embedding extends this to about 13% when a SDR threshold of 6 dB is used.

4. SIMPLIFIED PACKET EMBEDDED CODER

In the packet-embedded scheme described prediction indices of packets in an overflowing node queue must be read to determine which packet to discard. It is desirable to allocate only a small number of bits to the value of R in order to minimise overhead information and simplify processing at the nodes. We assessed the performance of the packet-based embedded ADPCM coder when R is represented by 1 bit, differentiating only between essential and supplementary packets. (An additional bit is required to distinguish every second packet; however this information will already be available in the header if a sequence number is included for re-ordering of packets at the receiver.) Fig. 4 shows the performance of this scheme for a 6 dB SDR threshold.

Simulation was carried out by deleting packets with R = 1 at random with fixed probabilities between 0 and 1 and calculating the corresponding lost packet rates and average SDR values. For higher lost packet rates all packets with R = 1 are discarded and, in addition, packets with R = 0 are selected at random with fixed probabilities. The performance of the simplified scheme is only slightly inferior to the same scheme with unquantised R; equivalence occurs when essential packets have been discarded with probability 1. Although a closer performance is obtained by increasing the number of bits allocated to R, our experience suggests that no more than 2 or 3 bits are required to obtain comparable speech quality.

5. ACKNOWLEDGEMENTS

M. M. Lara-Barron acknowledges the financial support of the "Consejo Nacional de Ciencia y Tecnologia", Mexico.

REFERENCES

[1] Turner, J.S., "New Directions in Communications," *Proceedings of the International Seminar on Digital Communications*, Zurich, Switzerland, March 1986, pp. 25-32.

[2] Yin, N., Li, S.Q., and Stern, T.E., "Congestion Control for Packet Voice by Selective Packet Discarding," *Proceedings of the IEEE Global Communications Conference*, Tokyo, Japan, Vol. 3, November 1987, pp. 1782-1786

[3] Goodman, D.J., Lockhart, G.B., Wasem, O.J., and Wong, W.C., "Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-34, No. 6, December 1986, pp. 1440-1448.

[4] Suzuki, J., and Taka, M., "Missing Packet Recovery Techniques for Low-Bit- Rate Coded Speech," *IEEE Journal on Selected Areas in Communications*, Vol. SAC-7, No. 5, June 1989, pp. 707-717.

[5] Petr, D.W., DaSilva, L.A., Jr., and Frost, V.S., "Priority Discarding of Speech in Integrated Packet Networks," *IEEE Journal on Selected Areas in Communications*, Vol. SAC-7, No. 5, June 1989, pp. 644-656.

[6] Lara-Barron, M.M., and Lockhart, G.B., "Correlation Techniques for Reconstructing Lost Speech Packets," *Proc. of IERE Conf. on Digital Processing of Signals in Communications*, Loughborough University, September 1988, pp. 197-202.

[7] Goodman, D.J., "Embedded DPCM for Variable Bit Rate Transmission," *IEEE Transactions on Communications*, Vol. COM-28, No. 7, July 1980, pp. 1040-1046.

[8] Bially, T., Gold, B., and Seneff, S., "A Technique for Adaptive Voice Flow Control in Integrated Packet Networks," *IEEE Transactions on Communications*, Vol. COM-28, No. 3, March 1980, pp. 325-333.

[9] Lara-Barron, M.M., and Lockhart, G.B., "Selective Discarding Procedure for Improved Tolerance to Missing Voice Packets," *Electronics Letters*, Vol. 25, No. 19, September 1989, pp. 1269-1271.

[10] CCITT Recommendation G.721, "32 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)," *CCITT Red Book*, Fascicle III.3, October 1984, pp. 125-159.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A TEXT-TO-SPEECH SYSTEM FOR DANISH

Bent Bagger-Sørensen*, Otto Bertelsen, Preben Dømler
Carsten Henriksen*, Peter Holtse, Peter Molbæk Hansen
Henrik Nielsen*, Niels Reinholt Petersen, and Jørgen Rischel

Institute of General and Applied Linguistics
Faculty of Humanities, University of Copenhagen
Njalsgade 80, DK-2300 Copenhagen S
Denmark

* Telecommunications Research Laboratory
Lyngsø Alle 2, DK-2970 Hørsholm
Denmark

The paper presents a text-to-speech system for Danish. This system converts Danish orthographical text into speech. Conceptually, the conversion process is carried out in three stages, higher-level linguistic processing, phonetic-level processing, and synthesization. The structure and function of each of the corresponding three system modules are described. The entire system is implemented in software to ensure maximal flexibility in the development phase, and it runs on a personal computer.

Introduction

Research in speech synthesis has been going on for some time at the Institute of General and Applied Linguistics (IGAL), University of Copenhagen and at the Telecommunications Research Laboratory (TFL).

In the late sixties a terminal analog synthesizer, with formant circuits arranged in parallel and controlled by an analog function generator, was constructed and used at IGAL as a research tool.

Exploratory work on synthesis of Danish was carried out in the early seventies by researchers from IGAL and TFL. This work was based on a version of the Swedish Ove-III synthesizer.

Since the mid-eighties researchers from the two institutions have formed a research group with the goal of developing a system converting standard Danish orthography into speech. At the present stage of development the entire system, including a simulated formant-coded terminal analogue synthesizer, is running on an IBM PS/2-70 personal computer.

The purpose of this paper is to give an outline of the main components of the system.

Text Parsing and Structure Transformation

Any text-to-speech system must cope with the problem of assigning an adequate phonetic representation to the input text. The various methods of solving this problem reflect - among other things - varying levels of linguistic ambition, the traditional systems with letter-to-phoneme rules representing the lowest level of ambition, and systems with all known sorts of linguistic (including semantic and pragmatic) knowledge the highest level of ambition.

In our text-to-speech system the linguistic component currently represents a medium level in this respect. The linguistic component is (conceptually) divided into two subcomponents: the analysis part and the transformer part.

1. The analysis part takes care of identification of the sequence of linguistic elements represented by the text and assignment of linguistic structure to the identified elements.

2. The transformer part carries out a transformation of the linguistic structure determined by the analysis part to an appropriate phonetic representation.

The interface between the two parts is a sort of linear morphophonemic representation.

The analysis part is concerned with parsing in a broad sense, i.e. with the structural interpretation resulting from combining grammatical and lexical knowledge of the system with a specific input text. This subcomponent is currently endowed with syntactic, morphological, and lexical knowledge. Conceptually, this knowledge is declared to the system in the form of a phrase structure grammar and a morph-lexicon, and the structural interpretation is carried out by a parser module which "understands" the formalism of the grammar and lexicon. The declarative formalism and the parser are described in Molbæk Hansen, 1989.

In this system no formal distinction is made between morphology and syntax, but the formalism allows the user to introduce and use such a distinction if he sees fit, and the

word, which is to all appearances an important unit in phonology, is recognizable in the actual morphophonemic interface format. In fact the grammatical formalism is related to X-bar theory with the category word representing level zero (cf. Selkirk 1982).

The transformer part reformats the morphophonemic strings at the interface level into phonetic representations using a set of ordered context-sensitive rules. This component is rather trivial, since its task is of a rather mechanical nature.

The real challenges lie in the analysis subcomponent. The integration of as much linguistic knowledge as possible is particularly important in Danish (and other Scandinavian languages), because sentential phonology is characterized by quite regular stress reduction phenomena (unit accentuation, see e.g. Rischel 1982) which are conditioned by syntactico-semantic properties of the constituents, cf. *han solgte huse* "he sold (was selling; was a seller of) houses", with stress on the last word only, vs. *han solgte husene* "he sold the houses" with stress on the two last words. These phenomena have nothing to do with emphasis for contrast or the like (which text-to-speech systems cannot even hope to cope with in a systematic way yet), but the information needed to identify them are in many cases extractable from a lexicon/grammar based analysis of text, and since they are very important for the naturalness of Danish sentence prosody, they must be integrated in any high-quality text-to-speech system for this language.

Phonetic-Level Transformations

As the next step in the conversion process the fairly broad phonetic transcription which is output from the higher-level linguistic processing is tranformed into a string of tables containing parameter values (such as formant frequencies etc.) and time references by which, following an interpolation routine, the synthesizer is ultimately controlled.

The development of the phonetic component takes place within the framework of the Synthesis Programming Language (SPL), a custom computer language specified to suit the needs of linguistic rule formulation (see Holtse and Olsen, 1985).

The basic unit to which the rules refer is the phone. All phones are listed in a library, where each is associated with a matrix of distinctive features uniquely defining the phone, and with a table of control parameters (formant frequencies, bandwidths etc.) for the synthesizer. Each parameter value is accompanied by a specification of transition times (internal and external to the phone), which control the change over time between the values of that parameter in adjacent phones. The transition time specifications as well as the parameter values may be modified by rule.

The phonetic rules are stated as context-sensitive rules in the general format A / (B) ___ (C) → D (i.e. unit A between optional B and C changes to D). This notation allows the phonetic rules to be rendered in linguistically meaningful terms. Conceptually the rules group into three categories in accordance with their particular domain of application. Thus, we may distinguish between insertion rules, segmental rules and suprasegmental rules.

The insertion rules are considered higher-level phonetic rules. Since the input from the morphology component may not necessarily be a full-fledged segmental form from the point of view of acoustic phonetics, the input string is accommodated to the low-level context by the replacement or insertion of relevant phone units, e.g. stop consonant explosions and aspirations. After the application of these rules the string has its fullest segmental shape. The segmental rules primarily handle assimilation phenomena between segments, and lastly, the suprasegmental rules control segment durations and fundamental frequency movements on the basis of prosodic information (e.g. stress and boundary markers) in the input string.

The phonetic component is currently developing, and in this phase there are, within reasonable limits, no *a priori* restrictions as to the number of units defined, or to the number and complexity of the rules. This principle ensures the flexibility and adaptability essential to a development system.

The speech synthesizer

The speech synthesizer converts control parameter specifications into actual sound.

The design and implementation of the synthesizer has been governed by the following basic requirements. It should be capable of realizing the widest possible range of phonetically relevant distinctions, also such distinctions which may perhaps not be specifically relevant for Danish. Furthermore, it should be easily adaptible to modifications elsewhere in the text-to-speech system, and it should be hardware independent.

As a consequence, the synthesizer is implemented entirely as a software simulation model written in a high-level programming language (C), and is running in a UNIX environment.

The synthesizer is formant coded, and in order to meet the above phonetic requirements, we have chosen a hybrid design with two resonator branches, one cascade, the other parallel (similar to the design given in Klatt, 1980). A block diagram of the synthesizer is shown in figure 1. In addition to the conventional voice and noise sources the present design includes a single-pulse generator, which has proved useful in the synthesis of stop consonant releases. Voicing and noise may be fed to either of the

two branches, but it is also possible, if required, to feed the sources to both branches simultaneously. Normally vowels and sonorant consonants are synthesized in the cascade branch and other consonants in the parallel branch. In order to improve the quality of voiceless fricatives (particularly *s*) an elliptic high-pass filter with variable cutoff frequency has been introduced between the noise source and the parallel branch. At the time of writing the synthesizer is controlled by 46 parameters which are updated every 10 miliseconds.

All calculations in the speech synthesizer are performed in floating point format to avoid overflow and truncation errors. Unfortunately this demand for precision adversely affects the real-time factor. Thus, with the present hardware (an IBM PS/2 personal computer equipped with a 80386, 25MHz processor) the real-time factor is approximately 5.

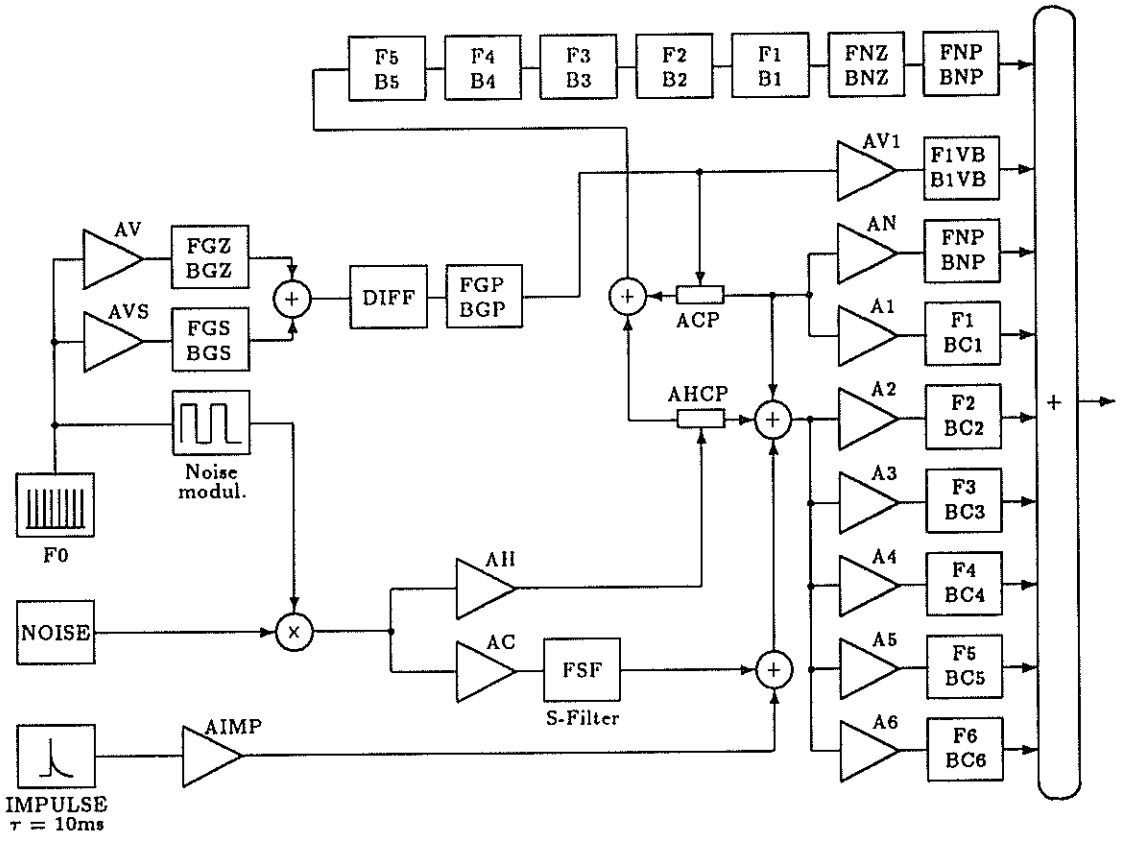


Figure 1: *Schematic diagram of the IGAL/TFL synthesizer*

AV/AVS:	*Ampl. of voiced source*	*F1 - F6:*	*Formant frequencies*
AH/AC:	*Ampl. of aspiration/frication noise*	*B1 - B6:*	*Formant bandwidth (cascade)*
AIMP:	*Ampl. of single pulse (plosive sounds)*	*BC1 - BC6:*	*Formant bandwidth (parallel)*
AV1:	*Ampl. of voice bar formant*	*FGx/BGx:*	*Glottal resonators*
AN:	*Ampl. of nasal formant*	*FNZ/BNZ:*	*Nasal resonator 1*
A1 - A6:	*Ampl. of formants in parallel resonators*	*FNP/BNP:*	*Nasal resonator 2*
ACP/AHCP:	*Cascade/Parallel distribution*	*F1VB/B1VB:*	*Voice bar resonator*
F0:	*Fundamental frequency of voice source*	*FSF:*	*Cutoff frequency for S-filter*

The output data are converted to a 12 bit fixed format (corresponding to a 72 dB dynamic range), and sent in two-byte chunks to the parallel port, which communicates with a custom-built speech buffer and D/A converter. Since the parallel port provides a standard I/O interface the need for special-purpose analogue I/O cards is eliminated.

The synthesizer as described above is, of course, subject to further modifications. Currently work is undertaken to improve the voicing source. As a first step a slightly more natural, though still not satisfactory voice quality has been obtained by introducing small-scale perturbations in intensity and fundamental frequency.

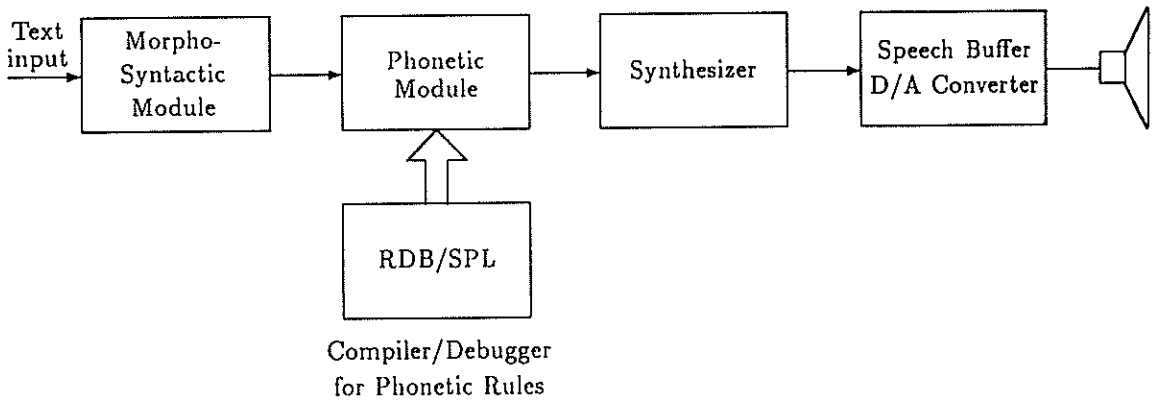


Figure 2: *Block diagram of the IGAL/TFL Synthesis System*

Concluding Remarks

For the time being the phonetic module receives limited sentence-syntactic information from the morpho-syntacic module. Therefore the system is only capable of handling short, syntactically simple sentences. Work is currently directed towards incorporation of more detailed syntactic information.

Although the primary aim of our work is high-quality synthesis, the quality of the generated speech still sounds unmistakably synthetic. This reflects the fact that neither the phonetic rules system nor the synthesizer itself have as yet, in our opinion, reached an adequate level of refinement.

References

Holtse, P. and Olsen, A. 1985: "SPL: A speech synthesis programming language", *Annual Report of the Institute of Phonetics, University of Copenhagen 19,* p. 1-42.

Klatt, D.H. 1980: "Software for a cascade/parallel formant synthesizer", *J. Acoust. Soc. Am. 67,* p. 971-95.

Molbæk Hansen, P. 1989: "Syntax, morphology, and phonology in text-to-speech systems", *Annual Report of the Institute of Phonetics, University of Copenhagen 23,* p. 119-52.

Rischel, J. 1982: "On unit accentuation in Danish", *Annual Report of the Institute of Phonetics, University of Copenhagen 16,* p. 191-240.

Selkirk, E.O. 1982: *The Syntax of Words,* MIT Press.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

INTONATION SYNTHESIS FOR MANDARIN SPEECH

JAVAID MIRZA

Department of Physics, National University of Singapore, Singapore 0511

An algorithm is developed and tested for generating intonation for statement type Mandarin Chinese sentences to be synthesized on text-to-speech synthesizer. The algorithm may be applied for any tone language providing the phonemes' durations and tonal slope values for the language are previously known.

Mandarin Chinese has four tones whose slope values for normally spoken speech were determined from a sufficient data. The phoneme duration data was taken from previously published literature. Thes informations are supplied to the algorithm which generates the intonation for a text-to-speech synthesizer.

For a large sentence containing many clauses the continuation from one clause to the following clause poses some problem as the continuation modifies the slope(s) of the opening tone(s) of the following clause. This continuative effect was also determined and formulated for all the four tones and the information utilized in the algorithm for the synthesis of many -clauses sentences. Minute details as consonantal effect on the consecutive tones have been ignored. Mandarin Chinese, synthesized with thus generated intonation, is intelligible and fair in sounding.

1. INTRODUCTION

Speech synthesis based on phonetic concatenation requires (1) the knowledge of duration for each phoneme and (2) the intonation contour of each sentence to produce a reasonably natural sounding speech. In this paper we will discuss an algorithm for automatic generation of intonation contour given the string of phonemes with known durations.

This algorithm is exclusively applicable to tone languages such as Mandarin and requires a previous knowledge of tonal slopes of all the tones used in the language. Unlike French and German etc. which are non-tone languages the incorporation of tonal slopes in intonation contour of a tone language is imperative not only for the production of a naturally sounding speech, but also for the intelligibility of the synthesized speech. First we give here the results of our investigation on the average slope values of the four Mandarin tones and then describe an algorithm which constructs the intonation contour of a Mandarin phonetic string.

2. CALCULATION OF MANDARIN TONES' SLOPES

Mandarin has four tones; their names, descriptions, characteristic pitch values and representative discritical marks are shown in Table 1 [Chao, 1948].

TABLE 1 : Mandarin Chinese tones

Tone	Chinese Name	Description	Pitch	Graph
1	Inpyng-sheng	high-level	55	
2	Yangpyng-sheng	high-rising	35	
3	Shaangsheng	low-dipping	214	
4	Chiusheng	high-falling	51	

Pitch values are symbolic in the sense that the digits represent starting and end points of the tones pitch range. The entire pitch range is divided in five equal intervals [1]. Unfortunately the tone slopes cannot be calculated from the Chao's pitch data associated with the starting and end points of tones because of lack of information on tone durations. We calculated the tones' slopes by recording 32 Mandarin sentences; each sentence was spoken by five native Mandarin speakers. These 160 sentences were analyzed for their intonation contours or pitch tracks with autocorrelation method using ILS package [2]. The resulting bar graphs of pitch versus time were examined with original speech signal. These displays easily yielded the identity of the tones whose slopes were to be measured.

The slope values were calculated from reasonably long straight central portions of the tones, which ranged between 50 to 150 msec long depending upon the duration of the tones. Mostly

TABLE 2 : Average slope values in neutral and continuation position of Mandarin Chinese tones. Slopes are given in Hz/msec.

Tone	Average slope in neutral position	Conjunction	Average tone slope in precon-junction word	% Change from the slope of column 2
1	0.026	and*	0.015	- 41.4
		and**	0.008	- 70.1
		and then	0.039	48.8
		but	0.030	15.2
2	0.221	and*	0.283	28.6
		and**	0.256	16.6
		and then	0.310	41.0
		but	0.290	32.1
3	0.362	and*	0.434	20.5
		and**	0.439	22.2
		and then	0.428	18.8
		but	0.467	29.8
4	- 0.431	and*	- 0.318	25.0
		and**	- 0.393	8.6
		and then	- 0.376	12.5
		but	- 0.370	14.0

* One verb sentences
** Two verbs sentences

the slopes were calculated using endpoint pitch values of the straight portions; when wavy parts existed in the tones, slopes were calculated using least square fit method. The average slope values for the Mandarin tones are given in Table 2. It must be pointed out here that only the tones that existed in the first two words of each spoken sentence were analysed for slope calculation so that the intonational effect of interclause continuation, sentence termination and punctuation etc. are nil or minimum on these tones. Thus the average values given in the second column of the Table 2 belong to the tones which are free of any intonation effects.

Another important effect which was studied was the effect of linguistic continuation on tones' slopes. Continuation is described as a lingual act of continuing in speech from one clause or phrase to another clause or phrase using conjunctions like "and", "but" etc. It has long been maintained that speakers tend to modify their pitch just before conjunction to manifest the continuity act from one clause or phrase to another, and that in most languages the said pitch modification is localized to a word or a syllable immediately preceding the conjunction. We observed that in Mandarin such a phenomenon also holds and that pitch modification is localized to immediately preceding word and is a function of the conjunction used. In Mandarin it is observed that the interclause continuity pushes up the slope from tones 2 to 4. Modification of tone 1 is inconsistent and it is not taken seriously because its slope is nearly zero and any change in its value would reflect a big change in percentage. Columns 4 and 5 of Table 2 show the modifications in tones' slopes which are conjunction dependent.

3. ALGORITHM TO GENERATE INTONATION CONTOUR

The algorithm to generate intonation contour was designed for a parallel formant synthesizer which was developed at the British Joint Speech Research Unit [3]. The synthesizer comprises 4 variable resonators and two high frequency fixed resonators which are used to generate formants. The resonators' outputs are passed via fixed shaping filters to the summing unit and fixed output filter [3]. Excitation is supplied independently to each resonator. It can be totally voiced, totally unvoiced or mixed. The synthesizer software supplies a list of phonemes as well as tables of formants and corresponding amplitudes for each phoneme. When synthesizing a sentence, the user has to key in a string of phonemes which constitute the sentence as well as the pitch value and duration for each phoneme. A rule system incorporated in the synthesizer software generates continuous tracks for formants and their amplitudes by reading target values of formants and amplitudes from supplied tables of each successive phoneme. Formants and amplitude values of consecutive phonetic elements are smoothly joined by rules taking into account their durations and then these values are supplied as control parameters to resonators to generate sound.

Exactly the same way the rule constructs the pitch track or intonation contour from the user-supplied constant pitch value for each phoneme, and this track is used as control parameter.

Unfortunately the intonation contour developed by the above rule does not work for Mandarin because Mandarin is a tone language and its every vowel in a sentence carries a distinct tone, not a constant pitch as in French and German. Our algorithm is exclusively meant for tone language. It accepts the phonetic strings as JSRU synthesizer software does but with tone number preceding each vowel element as shown below in example; Consonants are not given tone numbers.

M 3AI N 4AR Y 3ER NG J 1EE H 2ER Y 1AR

D1 D2.....D13

(Meina yang ji he ya)

(May Nah rears chickens and ducks)

Durations, D, are provided underneath each phonetic element in centiseconds. Normally durations can be fixed through repeated editing and audition or can be reasonably predicted through experience. After receiving entire duration information the algorithm prompts for starting pitch value for all the tones; these values are written in a table by the algorithm.

Reading the slope value of tone from a table, the supplied duration D and the starting pitch value the algorithm computes pitch track for all the tones. A rule then linearly joins the end points of consecutive segments to fill the gaps, these gaps in fact correspond to the consonants in phonetic strings.

The starting pitch values can be changed if desired to produce a different voice. This algorithm was tried on JSRU synthesizer along with the supplied software for the synthesizer to generate intonation contour for 8 Mandarin sentences. The intelligibility and sounding of the synthesized Mandarin sentences was certainly better than when their intonation contour was prepared with the software supplied by the supplier of the synthesizer.

REFERENCES

[1] Chao, Y.R., Mandarin Primer (Harvard University Press, Cambridge, 1948).
[2] ILS V6.0, User's Guide (Signal Technology Inc. California, 1987).
[3] Rye, J.M. and Holmes, J.N., A Versatile Software Parallel-Formant Speech Synthesizer, JSRU Report No. 1016, Cheltenham, England (1982).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A STATISTICAL MODEL OF DURATION CONTROL FOR SPEECH SYNTHESIS

Karl Huber
Gruppe fuer Sprachverarbeitung, Institut fuer Elektronik, ETH Zuerich, Switzerland

Abstract

We describe the modelling of segmental duration in speech for use with speech synthesis of German. As the synthesis is based on diphones, we consequently use a diphone inventory as a reference for semi-automatic duration measurement in natural utterances. The measurement procedure is based on a prediction of phonetic variants that are likely to occur and a prediction of duration itself. Thus duration prediction is used not only for synthesis, but also in the context of duration measurement rendering the measurement procedure more secure and efficient. The class of *generalized linear models* is found to be well suited for modelling duration, which depends on multiple factors. Model selection and estimation of parameters is eased by the use of the computer program *GLIM*. The example of data analysis presented is based on a sample of 30 rather lengthy sentences spoken by a single speaker, representing a variety of duration effects.

Introduction

Segmental duration refers to the timing structure in speech, which on a perceptual level manifests itself in a heterogeneous way as rhythm, accentuation, phrasing, speaking fluency, speaking rate. Speech duration is by no means just a matter of appearance but plays a major role in intelligibility.

For a long period, research on duration in speech has primarily been linguistically motivated. In the seventies, when the first text-to-speech systems came up, there was some activity in finding rules for duration control in speech synthesis systems, with D. Klatt probably being the predominant researcher. In the eighties some projects emerged that were concerned with duration for use in speech recognition [1,2]. The focus in research is slightly shifted depending on its motivation. Linguistically motivated research seems to be primarily interested in particular effects such as influence of voicing or tensity on duration, and experiments are designed carefully to make these effects evident. In speech synthesis, duration models have to account for all effects at the same time. But models are possibly restricted to a single speaker, including speaker dependant features. In continuous speech recognition, speaker *in*dependant features in duration are of much more interest.

In speech synthesis based on diphones, there arise some problems as the synthesis units do not correspond anymore to phonetic units. Different levels have to be distinguished and one level has to be designated as duration control level.

Model building starts from a sample of duration data, measured in an appropriate way. A model structure (a class of models) is chosen, and the best model is selected during an iterative process of fitting a sequence of models to the data. Statistics offers various methods for fitting models to data. The class of statistical models we are using are *generalized linear models.*

The model used by Klatt [3] is a rule system with a basically multiplicative function of the form $y = d_0 + d_1 \cdot f_1 \cdot f_2 \cdots$ which is nonlinear and difficult to estimate the parameters for. Omitting the additive factor and taking the logarithm we have: $\log(y) = \log(d_1) + \log(f_1) + log(f_2) \ldots$ being a purely linear function of the parameters. By applying a suitable transformation, we may retain a multiplicative model and still profit from the benefits of linear models.

It has been argued that simple additive or multiplicative models may not account for interactions between factors [2]. Linear models may in principle also account for interactions. But the combination of factor values grows rapidly, and the amount of data necessary to model all these interactions may soon become very extensive.

Factors affecting duration

Most approaches to segmental duration were based on phonemes [3,1,4,5,2] measuring the corresponding segments in speech. An exception is made by Campbell [6], who uses syllables as basic units. Whatever the basic duration units may be, it seems to be clear that on a segmental level phonemes may be regarded as factors having some impact on duration of the corresponding speech units. As other influential factors are concerned, a distinction is usually made between vowels and consonants as they react somewhat differently to factors like stress. Other factors that are commonly mentioned are the following. For vowels: postvocalic context, voicing, manner of articulation, accentuation, monosyllabic words versus polysyllabic words, syntactic boundaries, position of vowel relative to word, phrase and clause, content word versus function word. For consonants: clustering of consonants, initial, medial or final position in word.

All this information found in literature is very valuable in designing a model for segmental duration. But for use with a statistical model avoiding a rule system, a comprehensive scheme has to be found including possible factors in a unified way.

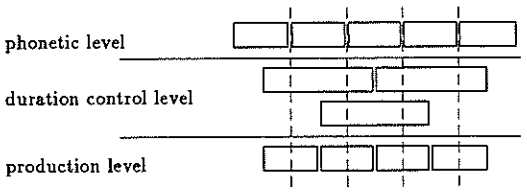


Figure 1: *Stratification of units*

We set up a 3-level stratification as shown in figure 1. On the phonetic level we have the realizations of phonemes, being a rather close description of an actual utterance that may include speaker dependant elements. The phonetic level is not a mere linear string, but there is some structure made up by syllables, words, feet, and (prosodic) phrases. On the lowest level - we call it *production level* - we do have the basic units to be measured in the speech signal. In our case, these units are diphones generally representing the transition between two speech sounds. As diphones seem not to be well suited for duration control, we use the sum of two adjacent diphones to form a new *duration control unit.* From figure 1 we see that such a unit may be attributed to three speech sounds. The medial one, supposed to have the most importance, the initial and final one, which are only partly involved. On a segmental level we use three factors to represent a *duration control unit.* If these three sounds are consonants, the combination of these three factors also accounts for the clustering effect. We also include transition information on the segmental level. There may be a sound transition, a syllable transition, a word transition, a phrase transition or a sentence transition. Then we assign the foot level information that consists of the syllable position within a foot and the number of syllables in a particular foot. As the first syllable in a foot is always stressed and the successors are unstressed, stress information is included in the foot information. Apart from accounting for speech rhythm, this might also account to some extent for the effect of polysyllabic shortening and possibly the effect of shortening in function words. Furthermore, position of a foot relative to phrase and sentence will be attributed to a *duration control unit.* A foot may be labelled phrase-initial, phrase-medial, phrase-final, sentence-initial, or sentence-final.

Figure 1 shows that the *duration control units* are overlapping. This is found to be necessary for the synthesis step, where duration of the synthesis units has to be reconstructed from *duration control units.*

Measurement procedure

Duration measurement has usually been done by visually inspecting spectrograms. An automatic or semiautomatic segmentation procedure is of course welcomed, provided it is accurate enough and flexible in use.

We use a complete diphone inventory [7] as a reference for duration measurements, i.e. test utterances will be segmented refering to the appropriate diphone elements. We start from a phonetic transcription of the test utterance which may include some allophonic variations, as we do not know exactly what the realization will actually be. For example, a schwa may be elided or not, a glottal stop may be realized or not, an unaccented /i/ may be realized as a lax [ɪ]. The phonetic transcription will be converted to a representation of *production elements*, the diphones.

Moreover, each element is assigned a predicted duration and an uncertainty interval. The actual duration is assumed to lie within this interval. The prediction is of course done by the duration model that will be described below.

The duration values are rastered such that a *dynamic programming (DP)* technique may be employed. The raster step is chosen as 2 ms for durations below 20 ms, 5 ms for durations below 100 ms, and 10 ms for durations above 100 ms, assuming that resolution is relative to duration.

All information on *production elements* and variants and assumed duration ranges is represented by a *lattice* (figure 2). A *lattice* is a *directed acyclic graph* with a *partial order*

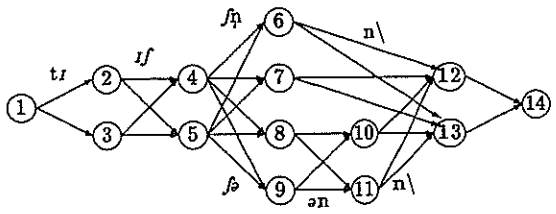


Figure 2: *Example of a simplified lattice representing [tɪʃən] and the variant [tɪʃn̩] from the word "britischen"*

on the set of its vertices, with a single root and a single end vertex. The *partial order* property just means that the vertices are numbered from left to right according to a relative position in a linear string. This is essential for the use of the *DP* technique. Duration measurement is now equivalent to mapping the *lattice* to the test utterance and finding the path with the best score. The score is a function of the local distances computed for each edge in the *lattice* with respect to the signal segment mapped to that edge. The local distance function measures the discrepancy between the test segment and the reference element expected. The form of the distance function depends on the element that is expected. With a speech sound expected, a spectral distance is calculated, with a pause expected, the intensity distance is calculated.

DP technique is used to do the optimization. Thompson [8] points out that *DP* may be implemented in the general context of a *chart parser.* The implementation will not be described here, but some comments will be made on the performance of the measurement procedure.

The procedure described above relies heavily on the information on elements possibly occuring and the duration range. The process may fail or perform badly if elements of the test utterance are not included in the *lattice* or if the actual duration is not included in the duration range. A proper working prediction model is a prerequisite for a good performance.

The process is based on reference elements and always expects an element that fits. This may not always be the case as the number of reference elements is limited.

Thus it is still necessary to inspect the results and correct the input information if any faults are detected.

Because duration is measured, but at the same time also predicted (based on estimates made from measurements) this is a kind of bootstrap procedure.

A statistical model

What are our goals in modelling duration? First of all, a model should help us in determining what factors do actually affect segmental duration and which do not. Clearly,

apart from the systematic component, there will always be a random component that may not be explained any further, since there is natural variation in speech. Therefore, reducing the residual variance by including a wealth of parameters may not be the one and only goal. There is in fact some danger in overinterpreting data, especially in small data sets such that the particular data are modelled, but not the underlying process. Thus we expect the model to tell us what the random component is like.

For the purposes of speech synthesis we ought to have a prediction model that assigns an actual duration value to a combination of given factors. A descriptive model does not suffice. And if possible, we would also like to have an idea of the prediction error or an interval within which a duration value will be likely to reside. All this demands for a comprehensive statistical model depending on multiple factors, allowing for a simultaneous estimation of its parameters.

We know that a *normal* probability distribution is not adequate for duration data (cf. [1,6]). The distribution of duration data has two important properties. First, duration is always positive valued, and second, the variance depends heavily on the mean value. Campbell [6] showed that a *log-normal* distribution fits duration data quite well. In fact, with a *log-normal* distribution of the random variable Y we have a variance

$$var(Y) = \mu^2\phi$$

where μ is the mean and ϕ is the *dispersion* parameter. A linear model has been fitted to log-transformed duration data and the residuals have been calculated. Figure 3 shows a *normal quantile-quantile plot* of the residuals exhibiting a rather straight line. This indicates that the *log-normal* distribution may be acceptable.

The *gamma* distribution with parameters μ and ν is a possible alternative. It is defined for positive real values, and the variance function equally depends on the mean value, $var(Y) = \mu^2/\nu$. As we found that the *gamma* distribution performs somewhat better, it will be used here.

Instead of using classical linear models, we will be using *generalized linear models* [9]. Roughly speaking, *generalized linear models* allow for modelling of random variables with distributions from the exponential family (including e.g. the *gamma* distribution). Moreover, a *link function* g may be introduced such that

$$\eta = g(\mu)$$

with μ being the expectation of the random variable Y, $\mu = E(Y)$ and η being the linear predictor,

$$\eta = \sum_i x_i\beta_i$$

The x_i's are refered to as the *covariates* or explanatory variables and the β_i's are the parameters of the model. The *covariates* may be categorical data (factors) or continuous valued data. For details on *generalized linear models* refer to McCullagh and Nelder [9].

Using the logarithm as the *link function* ($\eta = \ln(\mu)$) implies two things: First, this suggests that the underlying process is a multiplicative one and second, the model being used for prediction will yield positive duration values for arbitrary β's.

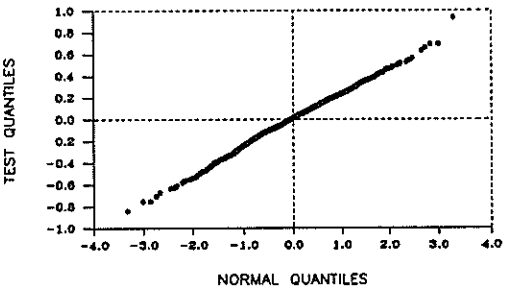


Figure 3: *Normal quantile-quantile plot for log-transformed duration residuals. The standard deviation is 0.063*

factor	dev	df	mean dev	red	df
ic	120.24	1150	0.1046		
c1	104.76	1144	0.0916	15.48	6
c2	91.533	1134	0.0807	13.23	10
c0	81.823	1124	0.0728	9.709	10
fl	74.817	1104	0.0677	7.007	20
po	72.14	1099	0.0656	2.677	5
tr	67.629	1095	0.0618	4.511	4

Table 1: *Example of deviance analysis (dev: residual deviance, df: degrees of freedom, red: reduction of residual deviance)*

Parameter estimation and model development is now carried out using the software package *GLIM377* [10]. Model development is an iterative process consisting of adding factors and interactions of factors to the model, analysing the *deviance* reduction, and removing factors not contributing significantly to *deviance* reduction.

Data analysis

As the units of segmental duration are not phonemes, the results presented may not be directly comparable with results in literature, so we will concentrate on the relative role that the factors play. Table 1 gives an example of analysis of duration based on units described above. On the first line we fit the *intercept* (*ic*) giving a mean *deviance* of 0.105. Then we add factors *c1* (class of medial phoneme), *c2* (class of phoneme to the right), *c0* (class of preceding phoneme), *fl* (foot level information), *po* (position of foot in phrase or sentence) and *tr* (transition information). Each factor added reduces the residual *deviance*. The mean *deviance* finally reduces to 0.062. For a predicted duration of 100 ms this yields a standard deviation of about 25 ms. Modelling short vowels separately (including the same factors) yields a mean *deviance* of 0.045 (s.d. of 21 ms for a duration of 100 ms). *Deviance* may be reduced further by substituting phonemes for phoneme classes.

Table 2 shows the parameter estimates arranged one factor per column. The foot level estimates are shown in table 3. It may be noticed that the first parameter in each column equals zero. So the *intercept* of 4.57 applies to the combination of the top factor values. The prediction for a specific combination of factors may now be constructed by summing

c1	c2	c0	po	tr
la 0.0	na 0.0	sh 0.0	si 0.0	wo 0.0
na -0.054	la 0.11	na 0.21	me 0.066	so -0.11
vo 0.094	vo 0.12	vo 0.20	pf 0.18	ph 0.22
vf 0.007	vf 0.20	vf 0.13	pi 0.025	sy -0.08
lv 0.35	lv 0.27	lv 0.23	sf 0.17	se 0.30
uf -0.15	uf 0.24	uf 0.072	mo 0.15	
sh -0.11	sh 0.036	la 0.15		
	vs -0.076	vs 0.13		
	us 0.022	us 0.10		
	pa 0.17	pa -0.29		
	gs -0.020	gs -0.11		

Table 2: *Estimated model parameters with an intercept of 4.57*

	1 syl	2 syl	3 syl	4 syl	5 syl
pos 1	0.092	0.098	0.11	0.007	0.147
pos 2	–	0.0	-0.024	0.016	0.096
pos 3	–	–	-0.044	-0.036	0.008
pos 4	–	–	–	-0.063	-0.006
pos 5	–	–	–	–	-0.065

Table 3: *Estimated foot level parameters*

the parameters for the factors that apply (including the *intercept*) and exponentiating it.

From the first column it may be seen that long medial vowels (*lv*) do have the greatest duration, that laterals (*la*), short vowels (*vo*) and voiced fricatives (*vf*) are about the same, and that unvoiced fricatives (*uf*) and schwa (*sh*) are shortest. From the third column it may be seen that initial pauses (*pa*) and glottal stops (*gs*) tend to shorten the segment. The position information performs exactly as one might expect. The longest are sentence and phrase final segments (*sf,pf*) and in isolated position (*mo*); sentence initial (*si*), phrase initial (*pi*) and medial (*me*) do not differ much. Transition information is significant as well. Segments including a sentence transition (*se*) are longest, phrase transitions (*ph*) are long, word transitions (*wo*) medium, and syllable (*sy*) and sound transitions (*so*) are shortest.

Table 3 shows the parameters for the foot factor. Rows correspond to the position of the syllable in the foot. Columns correspond to the number of syllables per foot. 'pos 1' indicates the stressed syllable with the greatest duration as expected. (1,4) is an exception that lacks an explanation. There is a tendency for the last syllable in the foot to be shorter with increasing number of syllables per foot. Also duration seems to decrease with the position in the foot. This tendency is supported by other data sets even though the values are not very significant.

Conclusions

The components of the duration model presented have been designed with a view to speech synthesis based on diphones. Consequently, the measurement units were chosen to be diphones. The measurement procedure is based on a knowledge component that accounts for possible allophonic realizations and predicts the segment duration in a particular context. Modelling duration may be a process that possibly has to be repeated for different speakers, different speaking styles, different languages. It is useful to have at one's disposal instruments that ease modelling, instruments for duration measurement, and instruments for the fitting of statistical models.

Acknowledgement

This work was financed by a ETH research grant. I would like to thank my colleagues of the speech processing group for their support in preparing this paper.

References

[1] Th.H. Crystal and A.S. House. Segmental durations in connected-speech signals: Current results. *J. Acoust. Soc. Am.*, 83(4):1553–1573, 1988.

[2] J.F. Pitrelli and V.W. Zue. A hierarchical model for phoneme duration in American English. In *Eurospeech 89*, pages 324–327, 1989. European Conference on Speech Communication and Technology, Vol 2.

[3] D.H. Klatt. Synthesis by rule of segmental durations in English sentences. In Lindblom B. and Öhman S., editors, *Frontiers of Speech Communication Research*, pages 287–301, Academic Press, 1979.

[4] D. O'Shaughnessy. A multispeaker analysis of durations in read French paragraphs. *J. Acoust. Soc. Am.*, 76(6):1646–1672, 1984.

[5] K. Bartkova and Ch. Sorin. A model of segmental duration for speech synthesis in French. *Speech Communication*, 6:245–260, 1987.

[6] W.N. Campbell. Syllable-level duration determination. In *Eurospeech 89*, pages 698–701, 1989. European Conference on Speech Communication and Technology, Vol 2.

[7] H. Kaeslin. A systematic approach to the extraction of diphone elements from natural speech. *IEEE Transactions on ASSP*, 34(2):264–271, April 1986.

[8] H.S. Thompson. A chart parsing realisation of dynamic programming with best-first enumeration of paths in a lattice. In *Eurospeech 89*, pages 378–381, 1989. European Conference on Speech Communication and Technology, Vol 1.

[9] P. McCullagh and J.A. Nelder. *Generalized Linear Models.* Chapman and Hall, second edition, 1989.

[10] M. Aitkin, D. Anderson, B. Francis, and J. Hinde. *Statistical Modelling in GLIM. Oxford statistical science series*, Clarendon Press, 1989.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ARTICULATORY SPEECH SYNTHESIS USING A TIME-DOMAIN MODEL

G.T.H. WRIGHT & F.J. OWENS

Department of Electrical & Electronic Engineering
University of Ulster

Abstract

This paper describes an articulatory speech synthesiser based on a time-domain acoustic-tube model of the vocal tract. The system consists of an interactive graphics editor, a model of the speech articulators, a time-domain simulation of a lossy acoustic tube model of the vocal tract and a two-mass model of the vocal cords. The graphics editor allows the visual editing of ten articulatory variables which provide a two-dimensional geometrical description of the vocal tract in the mid-sagittal plane. The vocal tract model is based on the Kelly-Lochbaum transmission-line model. The basic Kelly-Lochbaum structure has been extended to permit the modelling of viscous friction losses, yielding wall losses, radiation from the mouth, variable vocal tract length, aspiration and frication.

1 INTRODUCTION

The basic aim of most speech synthesis systems is to model the speech production process in one way or another. Although many of the most recent speech sythesis systems produce highly intelligible speech, most sound very mechanical, which makes them unacceptable for prolonged listening. One of the reasons for this lack of naturalness is that they are based on a source-system model of the speech production process, in which it is assumed that the source, i.e. the modulated airflow, is linearly separable from the system, i.e. the acoustic characteristics of the vocal tract. This is a fairly gross approximation, since it is well known that the generation and propagation of acoustic waves inside the vocal tract is governed by a single set of acoustic equations. In addition, in most current speech synthesis systems, such natural speech processes as co-articulation can only be simulated approximately in the acoustic domain. It may be expected, therefore, that a speech synthesis system in which the motion of the speech articulators, the acoustic and mechanical properties of the vocal tract and the vocal-cord/vocal-tract interaction, are taken into account, should be capable of producing much more natural-sounding synthetic speech.

This paper describes a time-domain articulatory speech synthesis model, programmed in PASCAL on a VAX minicomputer, which models directly the motion of the speech articulators and the generation and propagation of sound inside the vocal tract. A block diagram of the system is given in Figure 1. It consists essentially of an interactive graphics editor, an articulatory model for converting articulatory positions to a vocal tract area function, a time-domain acoustic tube model of the vocal tract and a two-mass model of the vocal cords.

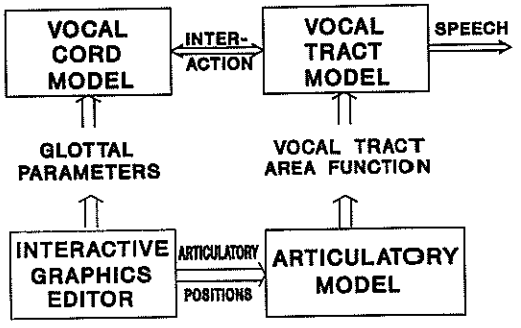


Figure 1: Articulatory Speech Synthesiser

2 ARTICULATORY MODEL

What we perceive as natural-sounding speech is generated by precisely-timed movements of the physiological mechanisms of the human speech production process. Below we describe the articulatory model used to represent the vocal tract configuration and the interactive graphics editor designed to make the process of generating and evaluating the configuration faster and easier.

At present, the articulatory model is essentially that due to Mermelstein [1973]. This model represents the vocal tract outline as a function of ten variables which specify the position of the jaw, tongue, lips, velum and hyoid. As shown in Figure 2, this provides a two-dimensional outline of the vocal tract shape, viewed in the mid-sagittal plane. The position of the jaw and hyoid are expressed directly in a fixed co-ordinate system. Lip and tongue body positions are specified with respect to the moving jaw. The tongue body and tongue blade are considered as separate but interdependent articulators with the tongue tip being specified relative to the tongue body. The tongue body outline is modelled as a circle with a moving centre and fixed radius. The lips are allowed to open (close) or protrude (retract) relative to the jaw and maxilla.

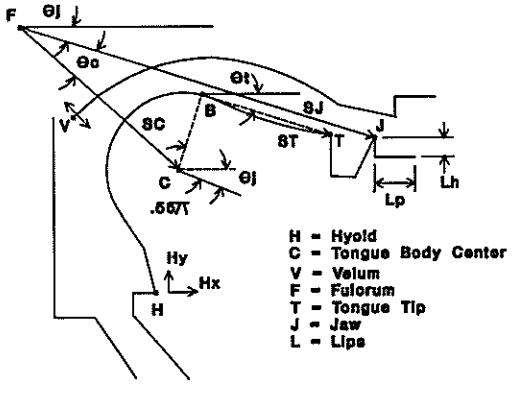


Figure 2: Articulatory Model and Parameters

The interactive graphics editor is designed for any Raster type graphics terminal. The articulatory model is manipulated by using a mouse to graphically adjust the relative amplitude of ten bars, where each bar represents a different articulatory parameter. Therefore the desired articulatory configuration can be rapidly defined. Each time any of the parameters is altered, the system automatically adjusts and displays the corresponding mid-sagittal outline and the vocal-tract area-function as shown in Figure 3.

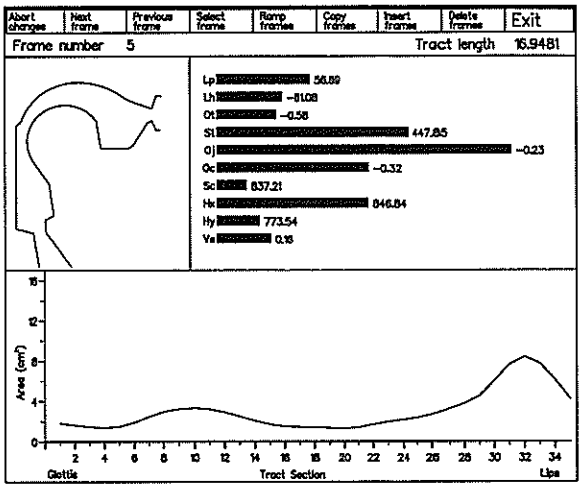


Figure 3: Interactive Graphics Editor

3 ACOUSTIC TUBE MODEL

The vocal tract is represented by a time domain simulation of a discrete, acoustic-tube model [Owens, Wright and Ramsey, 1989]. The propagation of sound within the vocal tract is modelled by computing the forward and backward travelling sound pressure waves [Kelly and Lochbaum, 1962]. Laminar flow losses due to viscous friction between the air and the vocal tract walls are lumped at the input and output of adjacent lossless tubes as shown in Figure 4. The complete vocal tract is modelled as a series of 35 lossless sections interconnected with lumped viscous loss resistances.

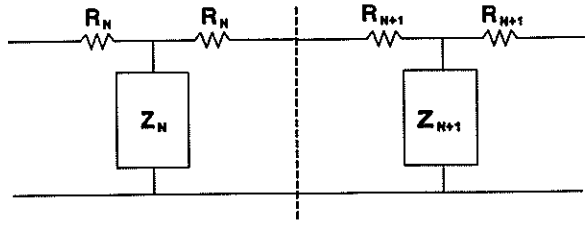


Figure 4: Modelling of Viscous Loss

Losses due to the vibration of the non-rigid vocal tract walls are incorporated in the model by assuming inde-

pendent motion of the wall of each section of vocal tract. The wall motion is modelled by a second order system and is used to modulate the pressure in each section. The motion and hence the degree of modulation depends on the mass, damping and compliance of the vocal-tract walls.

The impedance to sound radiation from the mouth is modelled as an equivalent parallel R-L circuit, where the values of R and L are those for a circular vibrating piston in an infinite plane baffle.

The nasal tract is modelled in much the same way as the oral tract. It is 12·5 cm in length and is coupled to the oral tract at a point 8 cm from the glottis. Most sections, except for a few in the vicinity of the velum, have a fixed cross-sectional area. A single sinus cavity model is also incorporated in a similar way. This cavity model is coupled 7 cm from the velum and is 4·5 cm in length and is of fixed cross-sectional area.

The vocal cord model used in the system is the two-mass model originally developed by Ishizaka and Flanagan [1972].

Turbulent noise models for the automatic production of aspiration and frication have been incorporated into the structure already described. The frication model is based on the research of Shirai [1975], and allows for automatic determination of the location of the turbulence as well as the frequency characteristics of the noise source, both of which are flow-dependent. For the automatic production of aspiration, a similar procedure is used except that a noise pressure source is added to the pressure difference across the glottis.

One disadvantage of the basic Kelly-Lochbaum reflection line simulation of the vocal tract is that the tract length is fixed. If the synthesiser is to be considered a complete model of the human speech production process then it must allow variation of the vocal tract length. In normal speech production the vocal-tract length varies between about 16 cm and 19 cm. This is simulated in the model by continuously varying the sample rate, ie a continuous variation of the vocal tract length is achieved by a continuous variation of the tube section length which leads to a related variation of the sampling frequency [Wu et al, 1987]. Therefore a system has been implemented to convert the signal sampled with a varying sample rate to a signal with a constant sample rate. The most straightforward way to perform this conversion is to reconstruct the signal, or the lowpass filtered version of it, from the samples and then resample it at the new sample rate [Crochiere and Rabiner, 1983]. The normal method of sample rate conversion, decimation and interpolation, could not be used since the two signals have incommensurate sampling rates.

The conversion is realised by a time-varying lowpass digital filter, implemented as a Finite Impulse Response filter and designed by windowing the impulse response of an ideal lowpass filter. The expression for the filter, including sampling at the new frequency, is

$$y(m) = \sum_{n=N_1}^{N_2} x(n) W(mT_o - nT_i) \frac{sin(\pi(mT_o - nT_i)/T_i)}{\pi(mT_o - nT_i)/T_i} \quad (1)$$

where

- $x(n)$ is the input signal sampled with frequency $f_i = 1/T_i$,
- $y(m)$ is the output signal sampled with frequency $f_o = 1/T_o$,
- $W(mT_o - nT_i)$ is the windowing function,
- N_1 and N_2 denote the minimum and maximum values of n involved in the computation of y(m) and
- the SINC function is the impulse response of an ideal lowpass filter.

The values of the impulse response $\hat{h}(t)$ that are used to give y(m) are spaced T_i apart in time and the centre of $\hat{h}(t)$ corresponds with the output sample to be calculated. Therefore $\hat{h}(t)$ will be shifted by T_o each time the next output sample is to be calculated. As the input and output sample periods are not the same, different values of $\hat{h}(t)$ will be used to calculate each output sample. Effectively the signal x(n) samples and weights $\hat{h}(t)$ to give y(m). The set of samples x(n) involved in the determination of y(m) are a function of the sampling periods T_i and T_o, the endpoints of the filter t_1 and t_2, and the output sample m being computed.

Tests were performed in the frequency and time domain to determine the optimum type and length of the function used to window the impulse response of the lowpass filter. The window functions tested were Rectangular, Triangular, Hamming, Hanning, Blackman and 4-term Blackman-Harris [Harris, 1978]. The tests were carried out on a large range of synthetic speech and figures of merit were calculated from the results, which can be seen in Table 1. These values represent the overall average values obtained. The time-domain values represent the overall average signal to noise ratios obtained and the frequency-domain values represent the overall average degree of spectral match.

Window function	Value of merit (dB)	
	Time Domain	Frequency Domain
Rectangle	29·0	26·7
Triangle	28·6	26·6
Hamming	47·7	52·2
Hanning	57·3	49·3
Blackman	58·0	53·3
4-term Blackman-Harris	59·5	53·8

Table 1: Values of merit for different window functions

The tests clearly showed that, excluding the triangular window function, a significant reduction in the error introduced by the sample-rate conversion process could be gained by using a soft window function instead of using a rectangular window function. This conclusion differs from that of Wu et al [1987] who concluded that no advantage could be gained by using a Hamming window function instead of a Rectangular window function in a similar investigation. This may be due to the small amount of test data used. Also, the distortion introduced by the linear interpolation of cepstrally-smoothed spectra may have contributed to their not being able to determine the improvement gained when using a soft window function.

The tests clearly showed that the smallest errors introduced by the sample-rate conversion process were obtained when the 4-term Blackman-Harris window function was used. They also showed that an acceptable conversion could be obtained when a Blackman, Hanning or Hamming window function was used. All the tests distinctly showed that the optimum number of window points was eight.

4 CONCLUSIONS

The system which has been described has the potential for producing natural-sounding synthetic speech when driven by time-varying articulatory or vocal tract area functions and an appropriate pitch contour. So far we have used the system to synthesise vowels, nasalised vowels, diphthongs, plosives, fricatives and some short words. Work is in progress to optimise the performance of the model by obtaining improved estimates of physiological parameters and to incorporate a model of articulatory dynamics.

ACKNOWLEDGEMENTS

Much of the work described in this paper was carried out with the support of a CAST Award from British Telecom Research Laboratories (BTRL). The development of the articulatory model was carried out by Neville Ramsey, with financial support from BTRL. Also, the help and advice of Dr John Local, Dr John Kelly and Dr John Coleman at the University of York is also gratefully acknowledged.

REFERENCES

Crochiere, R.E., Rabiner, L.R., Multirate Digital Signal Processing, Prentice-Hall, 1983.

Harris, F.J., On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform, Proc. of the IEEE, Vol 66, No 1, 1978, pp. 51-83.

Ishizaka, K. and Flanagan, J.L., Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords, Bell Syst. Tech. J., Vol 51, No 6, 1972, pp. 1233-1268.

Kelly, J.L. and Lochbaum, C., Speech Synthesis, Proc. Stockholm Communications, R.F.T., Stockholm, Sweden, September, 1962.

Mermelstein, P., Articulatory Model for the Study of Speech Production, J. Acoust. Soc. Amer., Vol 53, No 4, 1973, pp 1070-1082.

Owens, F.J., Wright, G.T.H., Ramsey, N.W., A Time-Domain Articulatory Speech Synthesiser, Proc. of the ESCA European Conference on Speech Communication and Technology, Eurospeech 89, 1989, pp. 176-179.

Shirai, K., Fujisawa, H. and Koyama Y., Modelling of the Generation of the Fricative Consonants, Elect. and Comms. in Japan, Vol 58-A No 6, 1975, pp 31-39.

Wu, H.Y., Badin, P., Cheng, Y.M. and Guerin, B., Vocal tract simulation : Implementation of continuous variations of the length in a Kelly-Lochbaum model, effects of area function spatial sampling, I.C.A.S.S.P., 1987, pp. 1.4.1-1.4.4.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Modelling Prosody Parameters for Declarative English Sentence Structures

Michael Wagner, Bob McKay, Santha Sampath, David Slater
Department of Computer Science
University College, University of NSW

A set of 144 declarative sentences with a subject-verb-object structure is drawn from a vocabulary of monosyllabic and disyllabic English words. Syllable timing, fundamental frequency contours and energy contours of the sentence set are analysed with respect to the syllabic structure of the sentences. Results on syllable timing are compared with the theory of stress timing. Multivariate correlation analysis provides predictions for syllable timing, average fundamental frequency of syllables and average energy of syllables.

1. INTRODUCTION

The accurate modelling of the suprasegmental parameters of speech, namely the timing of speech events and the contours of energy and fundamental frequency, is important for the synthesis of high-quality speech. It is also considered an essential component of any future general-purpose continuous speech recognition system.

A number of studies show that the fundamental frequency contour of declarative sentences has a falling tendency; yet, there is much controversy as to how the F0 decline may be described quantitatively. It appears that vowel quality, consonantal context, word stress and sentence stress all play some role in determining the F0 contour [1,2].

The energy contour of an utterance is not only a function of the intrinsic intensities of individual speech sounds but is also obviously influenced by word and sentence stress. Moreover, the aerodynamics of human speech production predicts a correlation between the energy and fundamental frequency contours [3,4].

This paper reports the first results of a study of Australian English prosody. The speech material is described in the following Section 2 of the paper. The syllable timing and the timing of stress intervals are discussed in Section 3. Section 4 looks at the dependency of fundamental frequency and energy contours on the syllable structure of the sentence and Section 6 presents the conclusions.

2. SPEECH DATA

The sentences examined were restricted to simple transitive sentences, with one-word subject noun phrases followed by one-word verbs and one-word object noun phrases. The words chosen are either monosyllabic or disyllabic, the latter class having a stress on either the first or second syllable. Some care was taken to ensure that both relatively high and relatively low vowels were represented in the words chosen in order to balance the intrinsic effects of vowel quality on F0.

Four words were chosen for each of the monosyllabic (11-words), disyllabic with stressed first syllable (12-words) and disyllabic with stressed second syllable (22-words) classes, for both nouns and verbs. This gives a total of twelve nouns and twelve verbs. The words were chosen so that half of the words in each class had relatively high vowels and the other half had relatively low vowels. The words chosen are listed in Table 1.

Each of the twelve nouns was used as a subject noun phrase with each of the twelve verbs. This produced a set of 144 sentences. Each of the twelve nouns was also used twelve times as an object noun phrase. The order of the sentences, and the object noun phrase for each sentence were determined pseudo-randomly. A selection-with-replacement algorithm was used which allowed the same noun to be both the subject and object in a single sentence, and this in fact occurred in five of the sentences.

	Nouns		Verbs	
	High vowel	Low vowel	High vowel	Low vowel
11-words	Greeks kings	mobs tarts	beat leave	mark rob
12-words	teachers students	farmers robbers	nibble tutor	charter honour
22-words	cadets patrols	Pathans Malays	admit respect	garotte retard

Table 1. Word selection for the sentence generator.

The sentences were recorded by a male speaker of Australian English and digitised at 8000 samples/s and 12 bits/sample. The signal energy was determined for each 16ms frame and the voicing and fundamental frequency parameters were determined every 16ms with a centre-clipping autocorrelation algorithm based on a frame size of 48ms [5].

Syllables were marked automatically with some manual corrections. For each syllable, maximum, average and total energies, average and central fundamental frequencies, zero-crossing rate and first linear-prediction coefficient were determined. In addition, the maximum-energy frame for each syllable was recorded for the determination of syllable and stress timing.

These acoustic and timing parameters were then correlated with the different structures of the recorded sentences.

3. SYLLABLE TIMING

The first part of the investigation examined the relationship between sentence syllabic structure and the timing of syllables. As each sentence of the recorded material consists of subject, verb and object, and each of the three words can be monosyllabic or disyllabic with word stress on either the first or the second syllable, the recorded material is divided into 27 groups of sentences where each group has a characteristic syllabic structure.

The three different word structures are denoted as "11" for a monosyllable, "12" for a disyllable with a word stress on the

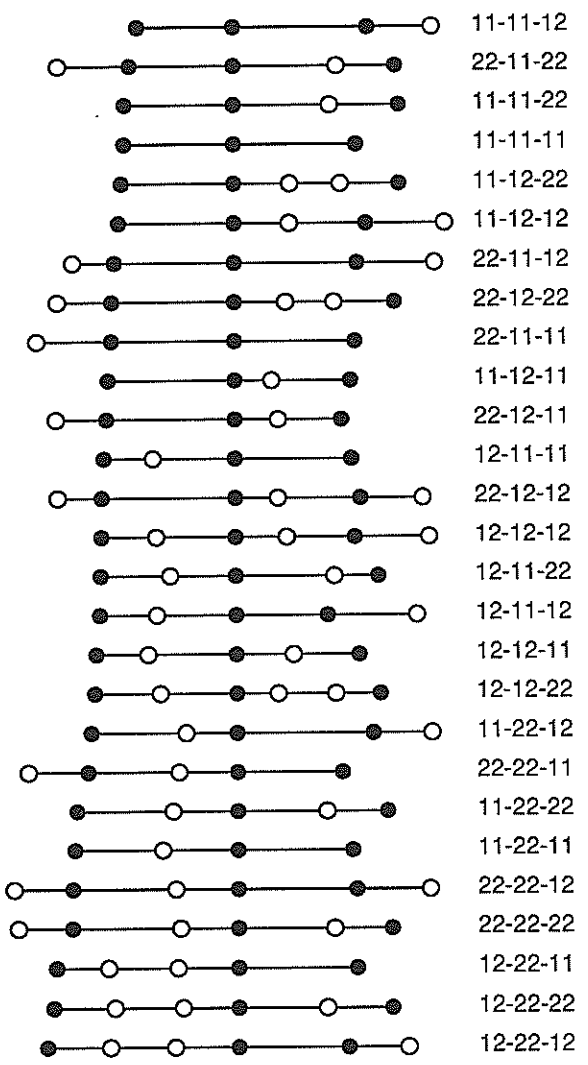


Figure 1. Stress intervals vs syllabic structure sorted by 1st interval duration (stressed=fat, unstressed=hollow).

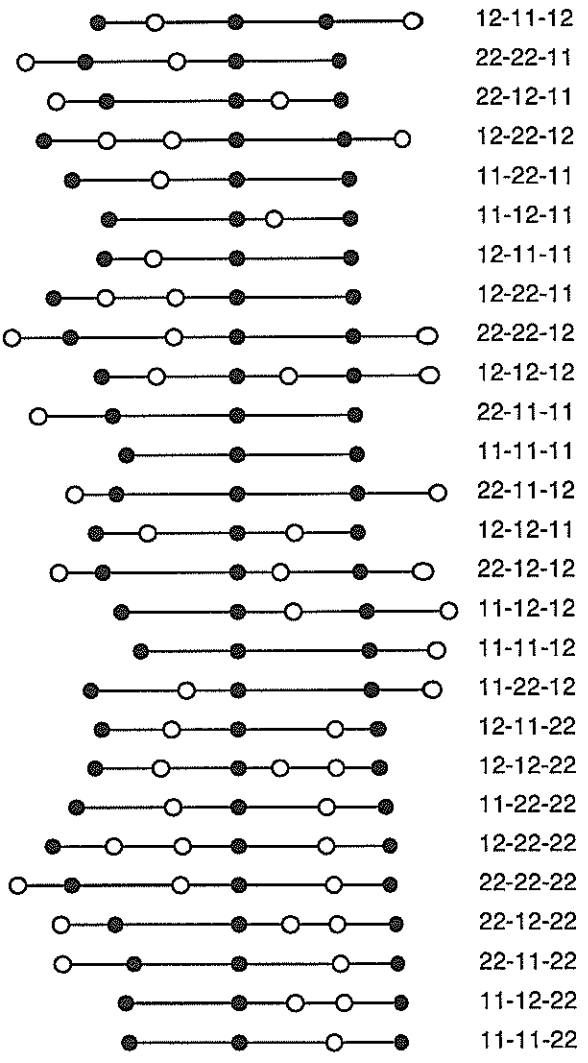


Figure 2. Stress intervals vs syllabic structure sorted by 2nd interval duration (stressed=fat, unstressed=hollow).

first syllable, and "22" for a disyllable with word stress on the second syllable. This notation is extended to sentence syllabic structures in the obvious way. For example, "11-12-22" denotes the syllabic structure of the sentence "Kings honour patrols", etc.

Since English is often described as a stress-timed language [6], the remainder of this section analyses the distributions of the lengths of the stress intervals for the recorded material.

Figure 1 shows the average size of the first stress interval, that is the time difference between the energy maxima of the first and second stressed syllables against the 27 different sentence syllabic structures. It is clear that the stress timing theory offers only a very rough explanation of syllable timing as the first stress interval varies from a minimum of 270ms for the 11-11-12 structure to a maximum of 538ms for the 12-22-12 structure.

The number of unstressed syllables within the first stress interval clearly influences the length of the interval as all (but 1) of the structures containing no intervening unstressed syllables appear at the top of Figure 1, followed by the structures with one intervening unstressed syllable and finally the 3 structures with two intervening unstressed syllables producing the longest intervals.

In the case of one intervening unstressed syllable, the average first stress interval is longer where the unstressed syllable belongs to the second word (11-22-xx and 22-22-xx structures) as compared with those intervals where the unstressed syllable belongs to the first word (12-11-xx and 12-12-xx structures).

Figure 2, which presents the average size of the second stress interval as a function of the sentence syllabic structure, shows firstly that there is no strong correlation between the length of the first stress interval and the length of the second stress interval, thus throwing further doubt on the general validity of the stress timing theory. In fact, the 3 longest average second-stress intervals are produced by 3 of the structures

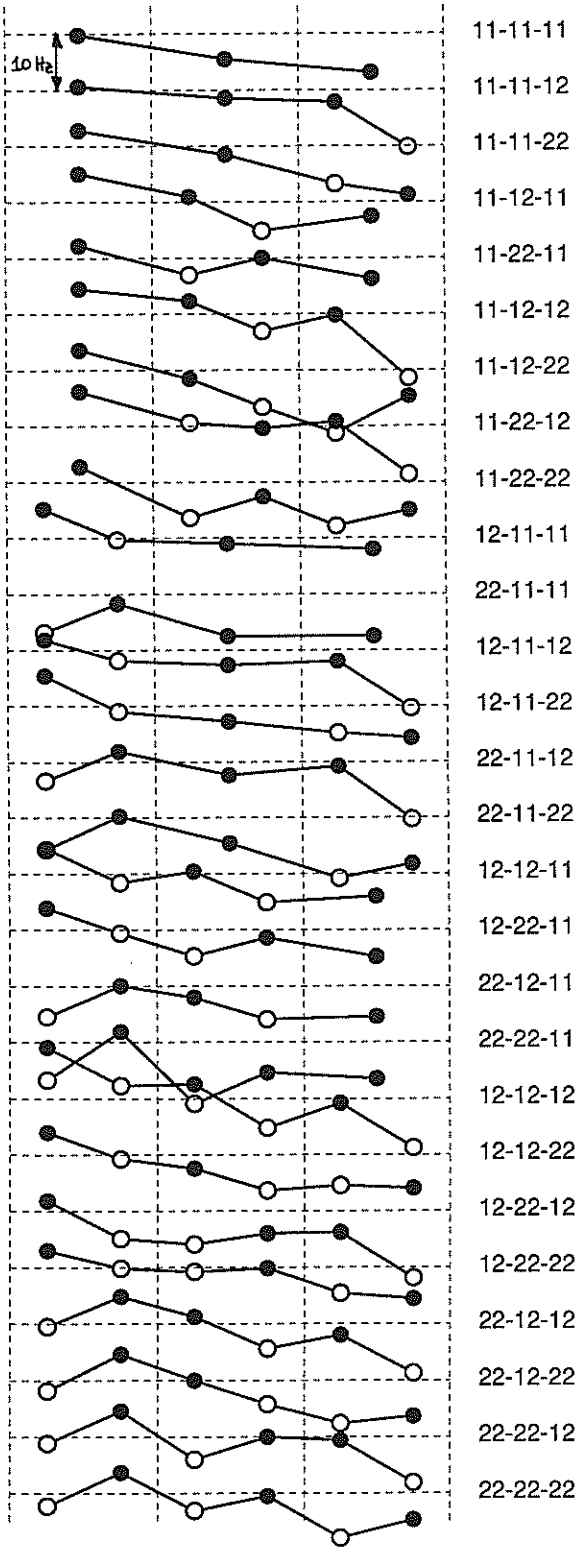


Figure 3. Average syllable F0 vs syllabic structure.

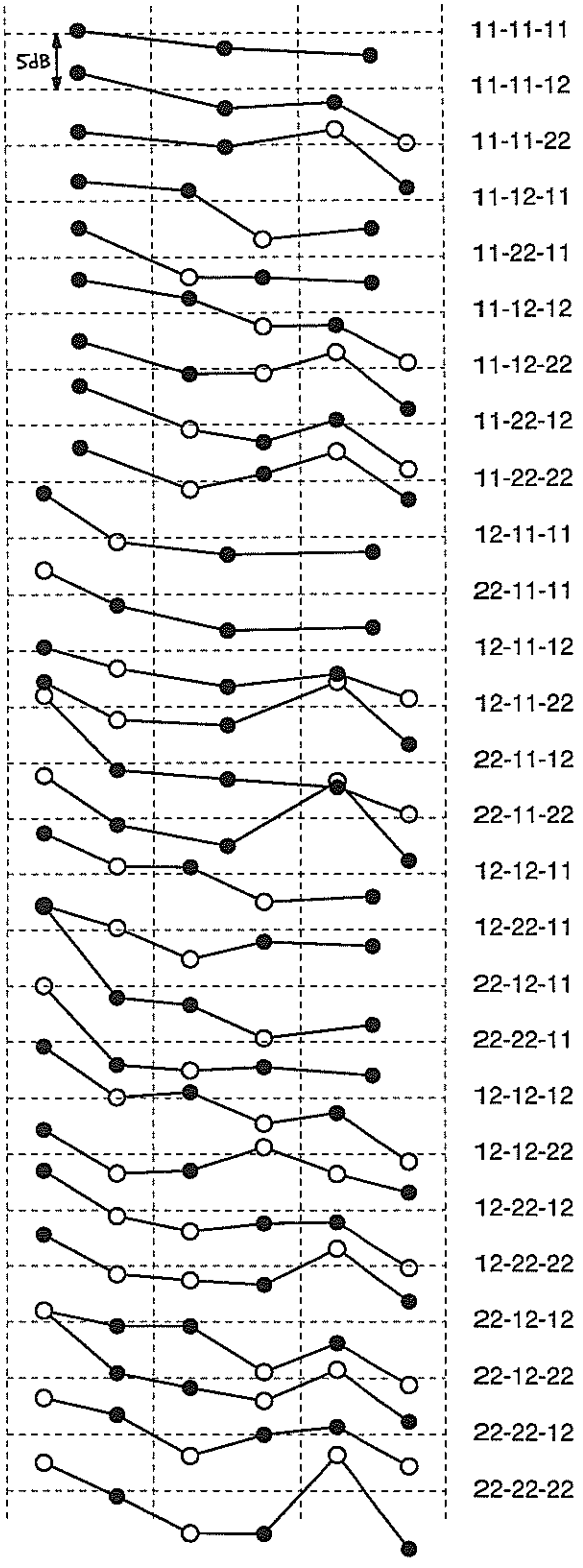


Figure 4. Average syllable energy vs syllabic structure.

with the shortest average first-stress intervals. Figure 2 also shows that the dominant influence on the 2nd stress interval is the structure of the final word with a 22-structure producing consistently longer intervals.

Multivariate correlation analysis was performed on the first and second stress intervals in order to obtain a quantitative description of the dependencies and a synthesis model for the intervals.

Introducing the variables x_1 for the number of unstressed syllables during the first stress interval (x_1=0,1 or 2), and x_2 for the direction in which a single unstressed syllable belongs (x_2=-1 if it belongs to the first word, x_2=1 if it belongs to the second word, x_2=0 if there are 0 or 2 intervening syllables), the first stress interval T_1 may be expressed as

$$T_1 = 324ms + x_1 * 96ms + x_2 * 35ms$$

with x_1 accounting for 49% and x_2 for 6% of the variance of T_1 and a total correlation of R=0.74.

Introducing the variables x_3 for the structure of the last word (x_3=1 if it is 22, x_3=0 otherwise) and x_4 for the size of the first stress interval in ms, the second stress interval T_2 may be expressed as

$$T_2 = 428ms + x_3 * 97ms - 0.24 * x_4$$

with x_3 accounting for 37% and x_4 for 10% of the variance of T_2 and a total correlation of R=0.67. It is interesting to note that there is a small negative correlation between the two stress intervals which suggests that instead of following a constant rhythm, the speaker possibly uses the second stress interval to compensate for an above-average or below-average duration of the first stress interval.

4. ENERGY AND FUNDAMENTAL FREQUENCY CONTOURS

Results of the analysis of the dependency of fundamental frequency contour on the sentence syllabic structure are shown in Figure 3. The horizontal axis represents the sequence of stressed (fat dots) and unstressed (hollow dots) syllables for the 3 words. The vertical axis represents the average fundamental frequency for the given syllable over the group of sentences with the given structure.

For most of the 27 different sentence structures, the falling tendency of a fundamental frequency contour from the first to the second word is easily observed. From the second to the third word, the speaker tends to raise the fundamental frequency very slightly. Within disyllabic words, most transitions between the two syllables are marked by a distinct fall in fundamental frequency.

Multivariate correlation analysis shows the following correlations between fundamental frequency, position of word in the sentence and syllable structure of the word:

Introducing the variables x_1 for the position of the word in the sentence (x_1=1,2 or 3), and x_2 for the word structure (x_2=0 if structure=11, x_2=-1 if structure=12, x_2=1 if structure=22) the average fundamental frequency for the syllable may be expressed as

$$F0_{ave} = 106.4Hz - x_1 * 3.7Hz + x_2 * 0.7Hz$$

with x_1 accounting for 18% and x_2 for 1% of the variance of $F0_{ave}$ and a total correlation of R=0.44.

The corresponding average energy contours for the 27 sentence categories are shown in Figure 4. In this figure, the vertical axis represents the average energy over the sonorant part of the syllable measured in dB.

The figure shows a decline of syllable energy over the sequence of words which corresponds to the falling tendency of the fundamental frequency shown in Figure 3. The other dominating influence on the syllable energy which is clearly shown is of course the higher energy for stressed syllables as compared with the unstressed syllables.

Modelling the energy contour on the variables x_1 as defined before and x_3 for the word stress (x_3=1 for a stressed syllable, x_3=0 for an unstressed syllable), results in the model for the average syllable energy of

$$E_{ave} = 43.4dB - x_1 * 1.7dB + x_2 * 2.3dB$$

with x_1 accounting for 25% and x_2 for 15% of the variance of E_{ave} and a total correlation of R=0.63.

5. CONCLUSION

Multivariate correlation analysis has been used to model the syllable timing, fundamental frequency and energy contours for simple declarative sentences. It was found that the strict application of the stress timing theory to the synthesis of English sentences is not valid and that both word stress and the number of unstressed syllables within a stress interval influence the timing of the stressed syllables. Models were derived for the fundamental frequency and energy contours which take into account the word stress and position within the sentence of the given word.

6. REFERENCES

[1] Lieberman, P. et al., Measures of the sentence intonation of read and spontaneous speech in American English, J.Acoust.Soc.Am. 77 (1985) 649-657.

[2] t'Hart, J., Declination has not been defeated - a reply to Lieberman et al., J.Acoust.Soc.Am. 80 (1986) 1838-1842.

[3] Waibel, A., Recognition of lexical stress in a continuous speech understanding system - a pattern recognition approach, Proc. ICASSP (1986), 2287-2290.

[4] Millar, J.B. and Wagner, M., The automatic analysis of acoustic variance in speech, Lang. & Sp., 26 (1983) 145-158.

[5] Rabiner, L.R. and Schafer, R.W., Digital Processing of Speech Signals (Englewood Cliffs, 1978).

[6] Ladefoged, P., A course in phonetics (San Diego, 1982).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A DTW-BASED APPROACH TO THE AUTOMATIC LABELING OF SPEECH ACCORDING TO THE PHONETIC TRANSCRIPTION

Daniele FALAVIGNA, Maurizio OMOLOGO

I.R.S.T.- Istituto per la Ricerca Scientifica e Tecnologica - Loc. Pante' di Povo, 38100 Trento - Italy

An approach to the automatic labeling of speech according to the phonetic transcription of the input utterance is described. Basically, a spectral variation function is evaluated, which allows the detection of fast spectral changes as well as slow transitions by using a single level representation. As a result, the peaks of such a function correspond either to phonetic unit boundaries or to acoustically significant subphoneme transitions. As a second step, the input utterance is time-aligned with a template generated by the concatenation of unit prototypes by using both a dynamic programming algorithm and the previously determined spectral variation function. A further refinement step is then applied to the obtained boundaries by analyzing the local energy contour. Performance has been evaluated, in terms of the discrepancy between automatically and manually determined boundaries, on a database of 265 Italian words: as a result, a boundary accuracy of 88.5% has been attained with a tolerance of 20ms.

1. INTRODUCTION

For both speech synthesis and speech recognition purposes, labeling continuous speech material according to the phonetic transcription represents a fundamental task. Especially when the goal is the manual definition of phoneme-like unit boundaries, this work becomes tedious, time consuming and error prone; furthermore, even if an expert phonetician performs the segmentation, the decisions are subjective and not reproducible.
In order to move toward the acquisition of large databases, both for a speech synthesizer and for a speech recognizer [1], we are developing a tool which combines semi-automatic labeling with facilities for hand-labeled correction.
Different automatic methods have been proposed in the literature to overcome the problem of automatic labeling. Assuming that a set of reference templates has been manually obtained, a good labeling on repetitions of the same set can be automatically determined by using a time alignment procedure [2].
In the development of a unit-based speech recognizer a way both to select a consistent unit set and to obtain the corresponding segment boundaries is to perform a forced recognition on the same material used to train the units themselves; as an example, an iterative training procedure has been recently described in [3].
Alternative methods have been proposed, which need no preliminary hand labeling. In [4] speech signal is preprocessed by an auditory model and successively transformed into a multi-level description: this output can be successively exploited to get the desired level of segmentation. In [5] a dynamic programming algorithm is applied to the incoming signal in order to minimize a cost function, which takes into account both a spectral variation parameter and a measure of similarity with the reference phones; in this case no insertion or deletion of units is generated.
The approach proposed in this paper can be classified with the latter type of methods described above.
A preliminary step is required to define a reference set of LPC unit prototypes. The segmentation and labeling system consists of the following steps:
- a grapheme to unit conversion is applied to the orthographic representation of the incoming utterance;
- LPC(Linear Predictive Coding) analysis is carried out on the corresponding speech waveform;
- a spectral variation function is computed to emphasize the clearest acoustic events;
- a DTW(Dynamic Time Warping)-based algorithm aligns the speech waveform with a concatenation of LPC prototypes representing the speech units to be bounded;
- by analyzing the local energy contour some of the obtained resulting boundaries are modified.

In this work the speech units correspond to a set of allophones of the Italian language. However, the flexibility of the proposed method allows the definition of other phonetically or acoustically significant units.

2. SYSTEM DESCRIPTION

The purpose of the system which is being developed is to identify the boundaries of the acoustic-phonetic units, which are related to the phonetic transcription of a given input utterance. Hence the input consists of both a speech signal and its corresponding text.
The system consists of the following steps: 1)unit definition, 2) signal analysis, 3) spectral variation evaluation and 4) time alignment. In this section we describe each of these steps in some detail.

2.1. Unit definition

As mentioned previously, the system requires an inventory of reference units. In a first version, the system was based on the following set of 38 speech units:
a- the phonemes of the Italian language;
b- three special allophones of /n/ (e.g. "/n/ versus fricative sounds", "/n/ versus vowels", "/n/ versus plosive and affricate sounds");
c- allophones of "vowel versus silence".

Further units have been successively included taking into account the relation between articulatory properties and the dynamic spectral evidence. For future experiments an extended set of 53 units is envisaged, which includes:
d- closure and release of stops;
e- different realizations of liquids (e.g. "/l/ versus silence", "/r/ in the right context of a plosive sound" etc.);
f- geminate consonants (nasals, liquids, affricates).

Each unit has been selected accurately from a speech database, starting from both the spectrogram and a spectral variation function, which is described in the next section. Even if these prototypes are used in the broad segmentation stage, care must be taken in this step to

provide a consistent definition of left and right context for each unit; performance of the segmentation system has resulted strongly dependent on this preliminary task.
To determine a synthetic prototype of the utterance, to be used in the time alignment stage, a grapheme to unit conversion is desirable: for this purpose a grapheme to phoneme transcription rule system, used for a speech synthesis task, is included in the system. This conversion is followed by a unit selection from the inventory of reference units.
The successive prototype concatenation step is performed by joining the selected units; each unit is represented through a set of LSP (Line Spectrum Pair) parameters determined by the LPC analysis. A prototype of silence is imposed both at the head and at the tail of the resulting reference.

2.2. Signal analysis

LPC coefficients and their transformations (e.g. LPC cepstral coefficients, LSP coefficients and so on) model the spectral envelope well and are widely used in speech processing. We are using the derived LSP parametrization, since it shows interesting time interpolation properties and a simple relationship with the LPC power spectrum. However, other spectral representations, such as the auditory model described in [6,7], could be more suitable to our purpose: we are investigating this choice, even if it requires a great computational load.
At the moment the input utterance is sampled at 15 kHz. Preemphasis of the digitized speech is accomplished by a first order digital filter whose transfer function is: $H(z)=1-0.95z^{-1}$. A 20 ms Hamming window, spaced every 10 ms, is used to determine the input signal to the LPC analysis. The LPC coefficients are obtained by using the Durbin recursive procedure. The predictor order is fixed to 16. The LPC coefficients are then converted into the LSP representation (Fig. 3a) by using a fast procedure described in [8].

2.3. Spectral variation function

A spectral variation function, which takes into account the spectral evolutions corresponding to significant acoustic changes, is used in the DTW-based broad segmentation stage.
As a preliminary study, a class of frame by frame spectral variation functions was investigated [8]: however, we believe that the definition of such type of functions should start from the consideration that the whole spectral structure, over the time, can be summarized in a few segments, which can be named local acoustic events. Each segment can be distinguished from its neighbours, starting from its mean spectrum. When the signal displays rapid characteristic changes, corresponding to quick consecutive local events (transient sounds), only the starting and ending boundaries of this collection of transition phenomena can be well captured, even in a manual segmentation task. A spectral variation function is desired to make evident these local acoustic events.
The function here described [9] is similar to the one defined in [10]; in our preliminary experiments it turned out that the latter function shows slightly better contours in speech versus silence and silence versus speech transitions, but depicts less prominent peaks in transitions between spectrally similar phones.
Let us denote as x_k the spectral representation for the k-th frame of a given utterance; x_k can be an LPC vector or any other appropriate spectral representation of a given order p. Let us consider the normalized inner product between the two vectors x_i' and x_j' as:

$$s_{ij} = (x_i')\, x_j' \,/\, \| x_i' \| \, \| x_j' \| \qquad (1)$$

where: $x_i' = x_i - m$ (2)

and m is the mean vector evaluated in an interval that includes both i-th and j-th frames. Given the left and right matrices $C_l=[x_{k-L}.....x_k]$ and $C_r=[x_k.....x_{k+L}]$ referred to the (2L+1)-frame interval [k-L ,....., k+L], we evaluate an (L+1)x(L+1) scatter matrix defined as follows:

$$S_{L,k}=\begin{bmatrix} s_{k,k} & \cdots & s_{k,k+L} \\ \cdots & \cdots & \cdots \\ s_{k-L,k} & \cdots & s_{k-L,k+L} \end{bmatrix} \qquad (3)$$

We define the spectral variation function of order L at the k-th frame as follows:

$$V_L(k)=1-(\Sigma S_{L,k})/(L+1)^2 \qquad (4)$$

where $\Sigma S_{L,k}$ denotes the sum of the elements of (3).
The normalized inner product has already been considered as an appropriate similarity function for pattern classification tasks. The resulting spectral variation function corresponds, in other words, to the average of the angles between each pair ($x_i \in C_l$, $x_j \in C_r$). With a suitable choice of L, this feature is expected to be a good measure of long term as well as short term correlations among vectors of C_l and C_r. In the following of this work, results are referred to the case L=6.
Its computation can be well organized in a recursive fashion: at each frame only 2L+1 elements of the scatter matrix $S_{L,k}$ have to be evaluated. It is worth noting that an equivalent definition can be derived by assigning the middle vector x_k either to C_l or to C_r: in this case the spectral variation function should be referred to transitions instead of frames.
The above defined function shows a low number of prominent peaks (Fig. 3c), which can be considered as anchor points for the time alignment step[9]. On the other hand, its valleys correspond to steady state segments and could be usefully exploited to identify the middle of a phoneme (e.g. for diphone selection).

2.4. Time alignment

The objective of this stage is to provide a preliminary assignment of the unit boundaries, starting from:
- the LSP parameters of the synthetic reference determined by joining the unit prototypes (reference pattern) selected in 2.1;
- the LSP parameters computed in the signal analysis step (test pattern);
- the corresponding spectral variation function.

Different DTW based approaches can be considered to time align the reference pattern to the test pattern [2].
In our experience the application of a standard DTW algorithm can generate gross misalignments; therefore, a constraint has been introduced, which can be summarized as follows: the optimal warping path has to go through the points of a grid, determined by both the boundaries between prototype units in the reference and the frames corresponding to spectral variation peaks of the test (Fig. 3d). In the test utterance, start and end boundaries are fixed. No further constraints (e.g. slope) have been included in the optimal path computation.
The distorsion measure, on which the minimization problem is based, is the rms log spectral measure, that is the L_2-norm of the spectral difference between two given log LPC-spectra.
The alignment algorithm described above can produce

unsatisfactory hypotheses, especially on boundaries corresponding to short events. A corrective procedure has been introduced, which is based on the energy contour of the input utterance: it is applied only on some transition classes, as we discuss in the next section. Broadly speaking, the log-energy derivative is evaluated in a short segment (approximately 80 ms) centered on the hypothesized boundary. This boundary is then moved to the position corresponding to the derivative maximum.
Furthermore, a time domain analysis, applied near each obtained boundary, is under study. We consider this step useful to correct boundaries corresponding to events for which a low tolerance is requested(e.g. plosive burst). In order to get this improvement, we are employing some features which can provide, sample by sample, an hypothesis of the class to which the surrounding context belongs.

3. SYSTEM PERFORMANCE

3.1. Speech material

System performance has been evaluated on a phonetically consistent database of 265 Italian words (spoken by a male speaker). The database can be considered rather difficult for a segmentation task due to several critical phoneme sequences that have been included. It contains 1229 phoneme to phoneme transitions and 530 word boundaries, but results refer to phoneme transition boundaries.

3.2. Evaluation criteria

Even if the manual labeling has been carried out by the authors (not expert phoneticians), at the moment performance of the proposed system can be given only by comparison with hand labeling accuracy.
However, this criterium could be considered inadequate to our final purpose. Tolerance should depend on the phoneme categories involved in a given transition; furthermore, the speech material should be manually labeled by the same expert phonetician. Otherwise, performance evaluation could be inconsistent.
On the other hand, a subjective assessment of automatically determined phoneme boundaries (made by expert phoneticians), may be more suitable in evaluating the performance of such a system. We are investigating this choice.

3.3. Results

The first version of the system was based on the use of the reduced set of 38 units described in 2.1. This version did not take into account the energy contour (the correction procedure was not included). Performance of this version is described in Fig. 1a. Correct boundary rates are reported as a function of an error threshold, which represents the tolerance on the automatically determined boundary, with respect to the manual one.
Further investigation on these preliminary results, for an error threshold of 20ms, has shown that:
- most of the errors in the "vowel to vowel" transitions concern diphthongs at the end of a word. The segmentation of two or more vowels is arguable: such sounds should probably be represented as a single acoustic-phonetic unit;
- errors in the "vowel to stop" transitions are due to a broad manual segmentation: even if stops are characterized by a single unit, the boundary marking the onset of the closure can generally be fixed with low accuracy;

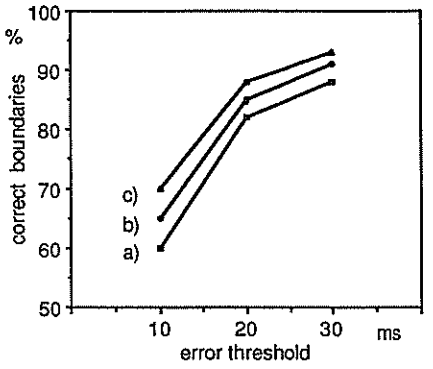


Fig.1 a-b-c. Performance of the system, for three different versions, as a function of the error threshold .

- gross errors are generally referred to transitions from liquid to vowel and viceversa; these errors probably depend on the similarity of the formantic structure of the sounds involved and could be reduced by exploring the local energy contour;
- secondary peaks of spectral variation function are often present inside vowels that are: lateralized, nasalized or at the end of a word. The last case is probably related to the fading of higher formants that characterizes these sounds.
Taking into account these observations, a second version has been defined, which includes the energy correction procedure applied to: "plosive to vowel", "vowel to plosive", "vowel to fricative" transitions (Fig. 1b). Note that the application of this procedure on each transition does not produce any benefit.
At last, a new set of 48 units has been considered which includes both different realizations of liquids and geminate nasal consonants. In this case the errors are consistently reduced (Fig. 1c) with respect to the first version: given the tolerances of 10, 20, 30 ms, the system provides correct boundary rates of 68%, 88.5%, 93%, respectively.
In Fig.2 correct boundary results are organized in a confusion matrix with respect to transitions between categories defined as follows: vowels (including glides), fricatives (weak and strong fricatives, affricates), liquids, plosives, nasals.

	vowel	plosive	fricative	nasal	liquid
vowel	48 / 70	139 / 146	130 / 143	68 / 83	54 / 70
plosive	252 / 268				4 / 4
fricative	205 / 213	18 / 20	3 / 4	0 / 1	
nasal	86 / 96	4 / 6	5 / 7		
liquid	60 / 86	2 / 3	8 / 8	0 / 1	

Fig. 2- Confusion matrix results for transitions between different phoneme classes: the number of estimated boundaries within 20 ms of the manual determined ones and the number of occurrences are given in the upper and lower corner of each box, respectively.

4. CONCLUSIONS AND FUTURE WORK

A DTW-based approach to the automatic labeling of speech according to the phonetic transcription has been described. The resulting system has been shown to provide satisfactory hypotheses of phoneme boundaries. However, we believe that further improvements, at different levels, can be obtained.
Future work will be focused on both the definition of an acoustically more consistent set of units and on the use of alternative speech analysis parameters, such as the synchrony output of an auditory model. Furthermore, a refinement step is being developed to provide better performance in transitions where a high accuracy is requested.

REFERENCES

[1] R.De Mori, R. Gretter, R. Kuhn, G.Lazzari, L.Stringa, "Modelling Operator-Robot Oral Dialogue for Applications in Telerobotics", accepted at 10th ICPR, June '90.
[2] H. D. Hohne, C. C. Coker, S. E. Levinson, L. R. Rabiner, "On Temporal Alignment of Sentences of Natural and Synthetic Speech", IEEE Trans. ASSP vol. ASSP 31, n°4, pp. 807-813: August 1983.
[3] R. Pieraccini, A.E.Rosemberg, "Automatic Generation of Speech Units for Continuous Speech Recognition", Proc. ICASSP '89, pp.623-626.
[4] J.R.Glass, V.W.Zue, "Multi level acoustic segmentation of continuous speech", Proc.ICASSP '88, pp. 429-432.
[5] B. Van Coile, "Computer Aided Segmentation of Spoken Words Given Their Orthographic Representation", Proc. European Conference on Speech Technology '87, vol. 1, pp. 277-280.
[6] S. Seneff, "A Joint Synchrony Mean-Rate Model of Auditory Speech Processing", Journal of Phonetics, 1988, 16, pp.55-76.
[7] R. De Mori, Y.Bengio, P.Cosi," On the Generalization Capability of Multi-Layered Networks in the Extraction of Speech Properties", Proc.IJCAI '89, pp.1531-1536.
[8] M. Omologo, "The Computation and Some Spectral Considerations on Line Spectrum Pairs ", Proc. EUROSPEECH '89, vol.2, pp. 352-355.
[9] D.Falavigna, M. Omologo, "A Spectral Variation Function for Acoustic Speech Segmentation", Proc. VERBA '90, pp. 365-372.
[10] J.G. Wilpon, B.H. Juang, L.R.Rabiner, "An Investigation on the Use of Acoustic Sub-word Units for Automatic Speech Recognition", Proc. ICASSP '87, pp.821-824.

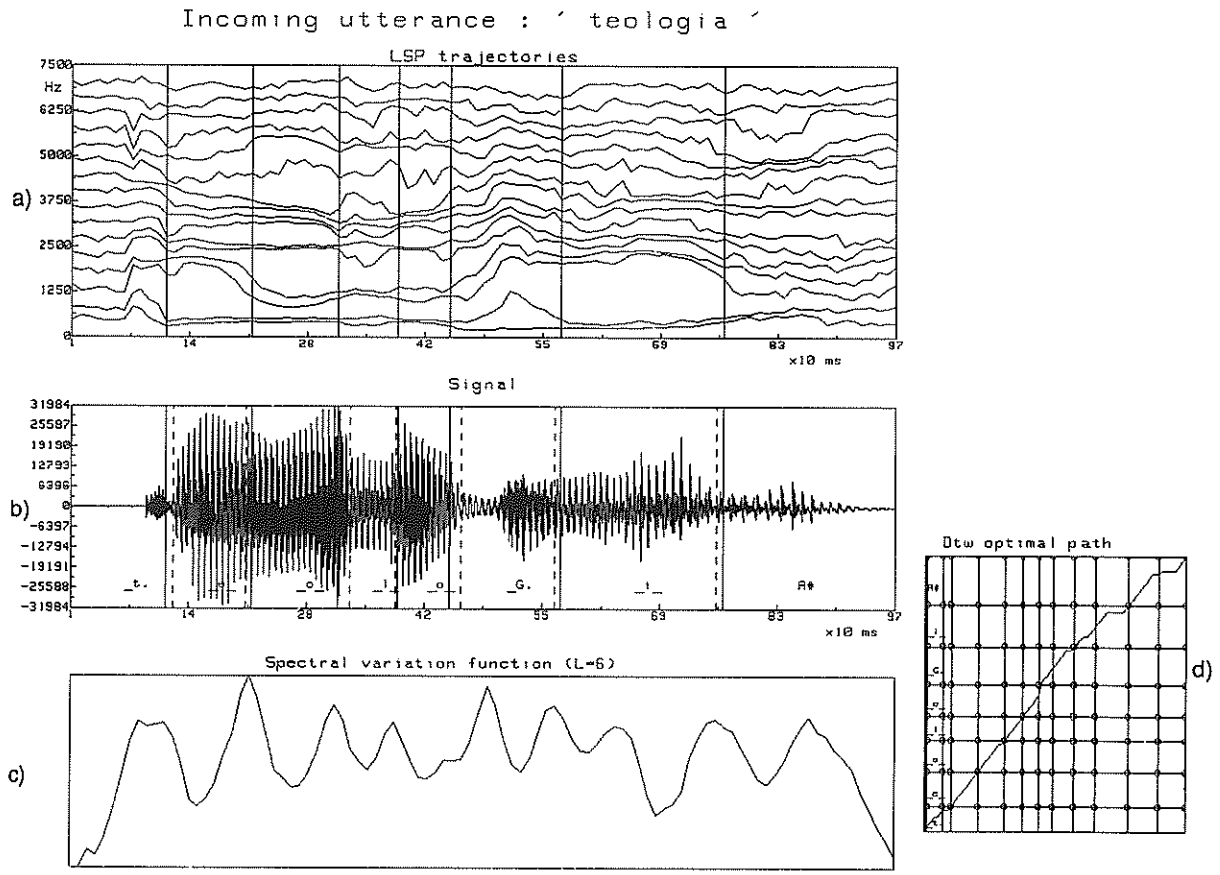


Fig.3. Example of the segmentation and labeling system output for the input word "teologia" (theology).
a) LSP trajectories; b) input speech: dashed lines and solid lines indicate manually and automatically determined boundaries respectively; c) spectral variation function; d) the DTW optimal path and the corresponding constraint point grid.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

SPEECH SYNTHESIS ON THE BASIS OF ACOUSTICAL TUBE MODELS FOR VOCAL AND NASAL TRACT

P. Köhler and A. Lacroix

Institut für Angewandte Physik, University of Frankfurt
D-6000 Frankfurt 1, Robert-Mayer-Str. 2-4, FRG

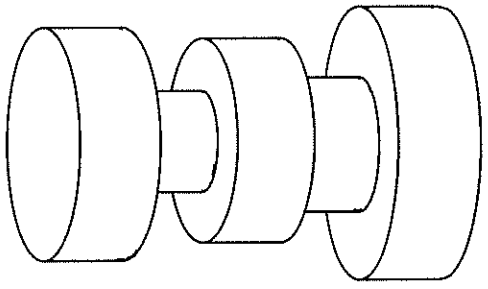
This paper describes the extended multi-tube-model of the vocal tract, in the sense that a nasal tract is coupled to it via a three-port-adaptor. This is a physically more consistent representation of the human voice production system than the traditional LPC representation. Beginning with the well established LPC model of speech production the effects of resonators and lossless tube models in terms of the pole-zero diagram of the transfer function are investigated. Then the enhanced model is described, the transfer function is derived and for some classes of tube configurations the effect of nasal coupling is demonstrated.

1. INTRODUCTION

An established method for speech synthesis is derived from LPC speech processing, where the receiver is used for speech synthesis [1,4,5,6]. One parameter set of the LPC-model, called reflection coefficients, has a close relationship to an acoustical tube model consisting of uniform tubes of identical length. The cross-sectional discontinuities are described by the reflection coefficients. LPC is commonly assumed to be in a close relationship to the human speech production system, which is necessary to adequately reproduce the variations of natural speech (about 1000 different phoneme transitions in the German language; differing intonation). But speech synthesis based on standard LPC method implies some disadvantages that evolve from inconsistencies in the representation of the human speech production system: (A) for the generation of plosives the frame length is too long. (B) The excitation of the vocal tract in case of fricatives takes place between the velum and the lips at varying locations depending on the phoneme instead of the glottis (LPC-synthesis implies that the tubes are always excited in front of the first tube). (C) The total tube length varies as a consequence of lip formation and tongue shape. (D) Changes from nasals to non-nasals imply an alternation of the meaning of the LPC-parameters and therefore a loss of information about the vocal tract. (E) LPC-synthesis is based on an all-pole model which cannot handle antiresonances that appear in nasal sounds. It is our aim, to improve the Multi-Tube-Model by coupling a nasal tract to it with the aid of a time-varying three-port-adaptor, which is the equivalent to the velum in the human vocal tract and therefore it is a trial to solve the problems (D) and (E) of speech synthesis based on standard LPC models, because this model is a more consistent representation of the human speech production system and it contains zeroes in its transfer function, which will be explained later.

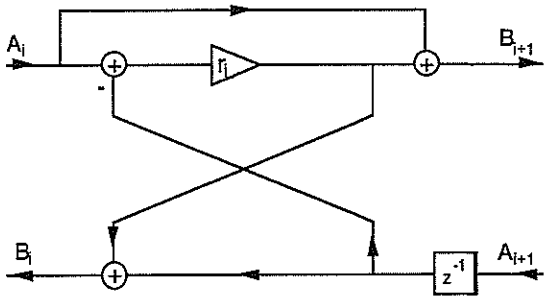
2. THE MULTITUBE MODEL

The Multi-Tube-Model (MTM) has two representations: 1.) The acoustical representation in which tubes of equal length but different cross-sectional areas are coupled together (see figure 1). The excitation takes place in front of the first tube and the sound wave is radiated at the end of the last tube. The jumps in the cross-sectional areas cause reflections of the acoustical waves.



- figure 1: Multi-Tube-Model consisting of five tubes -

2.) Figure 2 shows the discrete-time equivalent of a cross-sectional discontinuity and one tube-unit at the right. Complete MTM's emerge from concatenating this structure.



- figure 2: two-port-adaptor/acoustic tube combination -

The connection between both representations is that the coefficients of the two-port-adaptors can be calculated out of the

cross-sectional areas A_i and A_{i+1} [5].

$$r_i = \frac{A_i - A_{i+1}}{A_i + A_{i+1}} \tag{1}$$

The transfer function can be calculated from the scattering transfer matrices, which describe the transmissional behaviour of the lattice-structure

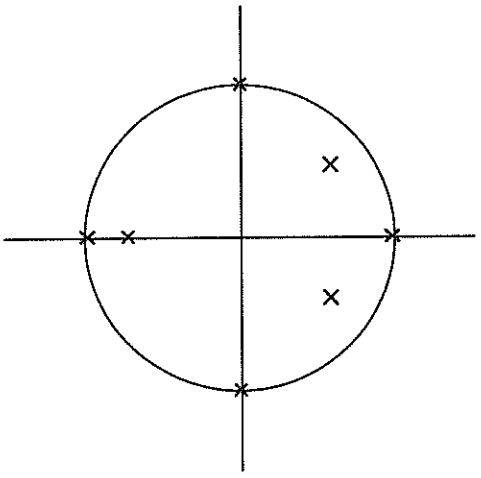
$$T_i = \frac{1}{1+r_i}\begin{pmatrix} z^{-1} & r_i \\ r_i z^{-1} & 1 \end{pmatrix} \tag{2}$$

with

$$\begin{pmatrix} B_i \\ A_i \end{pmatrix} = T_i \begin{pmatrix} A_{i+1} \\ B_{i+1} \end{pmatrix} \tag{3}$$

2.1. Resonators

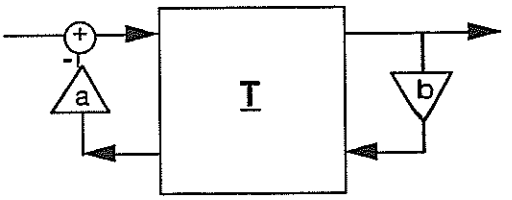
It can be shown that if there are major jumps in the cross-sectional area ($|r_i| \approx 1$) the transfer function tends to factorize in the transfer function before and after that jump [3]. Resonators are characterized by two major jumps, $r_k \approx -1$ and $r_{k+n} \approx 1$, with the reflection coefficients in between close to zero. Therefore the transfer function is factorized into three parts, where the part of the resonator can be calculated as $z^{-n}-1$, a cyclotomic polynomial. In figure 3 it can be seen, that a resonator of order 4 with $r_i=-.99$ and $r_{i+4}=.99$ has 4 equally distributed poles almost on the unit circle (besides other poles).



- figure 3: pole-zero-diagram of a MTM including a resonator -

2.2. Lossless Termination

If a Multi-Tube-Model is terminated with reduced two-port-adaptors whose reflection coefficients are either plus or minus one,



-figure 4: terminated MTM -

for example a=1 and b=-1 in figure 4, then the transfer function is determined by

$$\frac{Y}{X} = \frac{1}{(1\ 1)\ T \begin{pmatrix} -1 \\ 1 \end{pmatrix}} \tag{4}$$

with

$$T = \prod_{i=1}^{n} T_i. \tag{5}$$

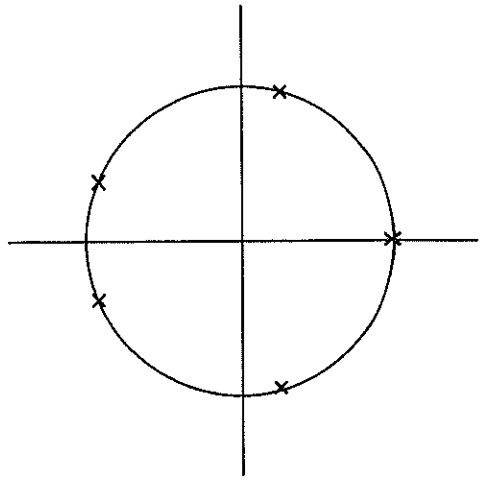
Now it can be shown by complete induction [3], that

$$(1\ \ 1)\,T \tag{6}$$

can be expressed as:

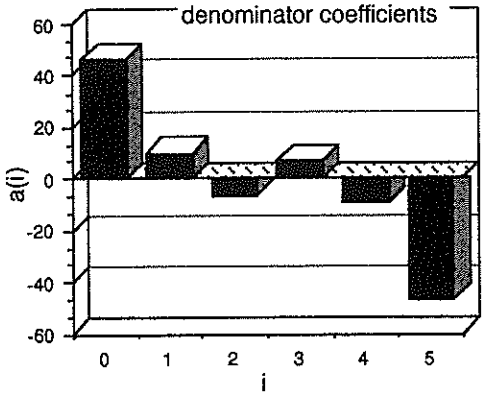
$$\left(\sum_{i=1}^{n} a_i z^{-i} \ , \ \sum_{i=0}^{n-1} a_{n-i} z^{-i} \right) \tag{7}$$

Out of this form the conclusion can easily be drawn that the resulting denominator polynomial is symmetric if b=+1 and antisymmetric if b=-1. These polynomials have no poles outside the unit circle because $|r_i| \leq 1$ ([5,6]) and therefore all poles must be located as complex conjugate pairs on the unit circle.



- figure 5: pole-zero-diagram of a lossless terminated MTM -

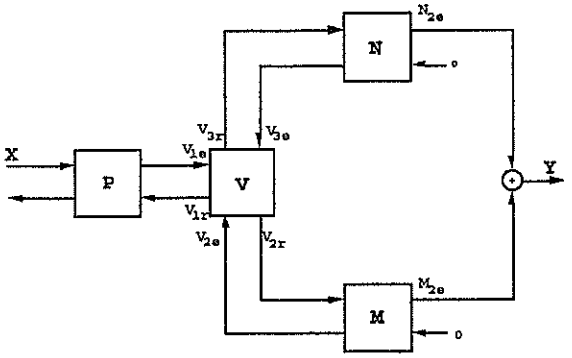
Figure 5 shows the pole-zero-diagram of the transfer function of a lossless terminated multi-tube-model and figure 6 shows the antisymmetric coefficients of its denominator polynomial.



- figure 6: denominator coefficients of a lossless terminated MTM -

3. NASAL TRACT COUPLING

Our approach for speech synthesis is to enhance the standard MTM by an additional nasal tract interconnected by a time varying three-port-adaptor (see figure 7). This is the equivalent of the velum in the human speech production system, which controls the dynamic coupling of the fixed nasal tract. All of the reflections (at the glottis, the lips and nostrils) are represented inside the lattice structures P (pharynx), N (nose) and M (mouth).



- figure 7: the extended Multi-Tube-Model -

It is assumed, that no acoustic energy from the nostrils is coupled in the mouth tract at the lips and vice versa.

3.1. Calculating The Transfer Function

It is impossible to calculate the transfer function in the normal way; i.e. to express port variables on one side of an adaptor by those on another side via scattering transfer matrices because they are not defined for three-port-adaptors as V. But it is possible to calculate a scattering matrix out of the continuity equations for pressure and flow. With S_i as the three cross-sectional areas of the pharynx, mouth and nose part of the velum, the scattering matrix is:

$$S_v = \frac{1}{\sum_{i=1}^{3} S_i} \begin{pmatrix} S_1-S_2-S_3 & 2\,S_1 & 2\,S_1 \\ 2\,S_2\,z^{-1} & (S_2-S_1-S_3)\,z^{-1} & 2\,S_2\,z^{-1} \\ 2\,S_3\,z^{-1} & 2\,S_3\,z^{-1} & (S_3-S_1-S_2)\,z^{-1} \end{pmatrix} \quad (8)$$

with

$$\begin{pmatrix} V_{1r} \\ V_{2r} \\ V_{3r} \end{pmatrix} = S_v \begin{pmatrix} V_{1e} \\ V_{2e} \\ V_{3e} \end{pmatrix} \quad (9)$$

These three equations can be transformed into two equations, which connect V_{1e} and V_{1r} with V_{2e}, V_{2r}, V_{3e} and V_{3r}:

$$T_v = \frac{1}{S_2} \begin{pmatrix} S_2 & 0 & S_1+S_2-\frac{1}{2} & S_1-\frac{1}{2} \\ -S_2 & 0 & \frac{1}{2}-S_2 & \frac{1}{2} \end{pmatrix} \quad (10)$$

with

$$\begin{pmatrix} V_{1r} \\ V_{1e} \end{pmatrix} = T_v \begin{pmatrix} V_{3r} \\ V_{3e} \\ V_{2r} \\ V_{2e} \end{pmatrix} \quad (11)$$

and an equation, which expresses the pressure continuity in the form of acoustic flow-waves:

$$\begin{pmatrix} S_3 & S_3 & -S_2 & -S_2 \end{pmatrix} \begin{pmatrix} V_{3r} \\ V_{3e} \\ V_{2r} \\ V_{2e} \end{pmatrix} = 0 \quad (12)$$

The next step is to express the (4x1)-vector in equation (11) through the radiation of lips and nostrils. This is achieved by a (4x4)-matrix, in which the scattering transfer matrices of the mouth and the nasal parts are included:

$$\begin{pmatrix} V_{3r} \\ V_{3e} \\ V_{2r} \\ V_{2e} \end{pmatrix} = \begin{pmatrix} N & 0 \\ 0 & M \end{pmatrix} \begin{pmatrix} 0 \\ N_{2r} \\ 0 \\ M_{2r} \end{pmatrix} \quad (13)$$

whereby

$$N = \begin{pmatrix} z^{-1} & 0 \\ 0 & 1 \end{pmatrix} \prod_{i=1}^{nmax} \frac{1}{1-r_{ni}} \begin{pmatrix} z^{-1} & r_{ni} \\ r_{ni}z^{-1} & 1 \end{pmatrix} \quad (14)$$

It should be noted that the first term in equation (14) (like in the definition of **M**) represents the scattering transfer matrix of a two-port which is contained in the three-port-adaptor V.
Out of the equations $Y = N_{2e} + M_{2e}$, (12) and (13) a (4x1) vector **A** can be generated, which connects the output at the lips and the nostrils with their sum Y

$$A = \frac{1}{S_2(N_{12}+N_{22})+S_3(M_{12}+M_{22})} \begin{pmatrix} 0 \\ S_3(M_{12}+M_{22}) \\ 0 \\ S_2(N_{12}+N_{22}) \end{pmatrix} \quad (15)$$

with M_{ij} denoting the polynomial in row i and column j of the matrix $\tilde{\mathbf{M}}$ so that

$$\begin{pmatrix} 0 \\ N_{2r} \\ 0 \\ M_{2r} \end{pmatrix} = AY. \quad (16)$$

With the definitions above the transfer function can be expressed as

$$\frac{Y}{X} = \frac{1}{(0\ 1)\ P\ T_V \begin{pmatrix} N & 0 \\ 0 & M \end{pmatrix} A} \quad (17)$$

where **P** is the scattering transfer matrix of the pharingeal tract.

3.2. Properties of the Transfer Function

First it is obvious that if the nasal tract is not coupled to the vocal tract ($S_3 = 0$) or if the mouth tract equals the nasal tract (**M=N**) the model tends to become an all-pole-model, because the numerator (the denominator of **A**) can be reduced by canceling against the denominator polynomial. It can also be proven that poles from resonators in the mouth or nasal tract show up too in the transfer function of the whole system. The results from chapter 2.2. can also be used to analyze the structure of the numerator polynomial (the denominator of **A**), because the open terminations at lips and nostrils can be transformed in MTM's terminated by reduced two-port-adaptors with reflection coefficients close to -1. So the numerator polynomial can be approximated as

$$S_3(1\ 1)\,M\begin{pmatrix} 0 \\ 1 \end{pmatrix} + S_2(1\ 1)\,N\begin{pmatrix} 0 \\ 1 \end{pmatrix} \approx$$
$$S_3(1\ 1)\,\tilde{M}\begin{pmatrix} -1 \\ 1 \end{pmatrix} + S_2(1\ 1)\,\tilde{N}\begin{pmatrix} -1 \\ 1 \end{pmatrix} \quad (18)$$

if $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{N}}$ represent the scattering transfer matrices of the mouth and the nasal tract without the last reflection at the nostrils or the lips respectively. So the numerator is the sum of two antisymmetric polynomials. If nose and mouth tract have the same length the resulting numerator polynomial can be approximated by an antisymmetric polynomial as well and the experiences we gained in modelling the vocal tract even showed that the zeros of the transfer function are placed on the unit circle of the complex plane, but we could not prove the validity or the falsity of this statement.

4. PERSPECTIVES

This approach may lead to improved speech synthesizers, which are able to handle nasals better than the established ones. The physical proximity of this model may allow to reproduce the phoneme transitions e.g. with the tongue by length-varying resonators or with the jaws more adequatly then established methods. A parameter generating structure for this model has been suggested by Rahim and Goodyear in [2].

REFERENCES

[1] Kelly, J. and Lochbaum, C.: "Speech Synthesis", 4th Int. Congr. Acoust., paper G42, 1962.
[2] Rahim, M.G. and Goodyear, C.C.: "Articulatory Synthesis with the Aid of a Neural Net", IEEE Conf., ICASSP 1989, pp. 227-230.
[3] Köhler, P.: "Untersuchungen von gekoppelten Mehrrohr-Systemen als Modelle des Sprechtrakts", Diploma Thesis, Frankfurt 1989.
[4] Frank, W. and Lacroix, A.: "Multi-Tube Models for Speech Synthesis", Proceedings Eusipco-86, The Hague 1986, pp. 373-376.
[5] Makhoul, J.: "Stable and Efficient Lattice Methods for Linear Prediction", IEEE Trans. on Acoustic, Speech and Signal Processing, ASSP-25 (1977), pp. 423-428.
[6] Markel, J.D. and Gray, A.H. Jr.: "Linear Prediction of Speech", Berlin / Heidelberg / New York: Springer 1976.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ERGODIC HIDDEN MARKOV MODELS FOR SPEECH SYNTHESIS

Piero Pierucci (*), Alessandro Falaschi (**)

(*)IBM Italy, Rome Scientific Center, Via del Giorgione 159, 00147 Rome, ITALY
(**)Univ.di Roma 'La Sapienza', INFOCOM Dpt., Via Eudossiana 18, 00184 Rome, ITALY

This paper presents applications of Ergodic Hidden Markov Models in speech synthesis. EHMM using autoregressive gaussian continuous densities as state observation densities are defined. The ability of such models to represent phonotactical constraints of the language is investigated through the analysis of the transition matrix structure and with some ergodic speech speech synthesis experiments. Applications of EHMM models in speech synthesis units segmentation and phone-like models statistical units representation are presented.

1. INTRODUCTION

Hidden Markov Models has become an increasingly popular technique in speech processing although their use has been mainly focused on speech recognition. Recently a number of contributions have appeared in technical literature about the application of Ergodic HMM to speech processing. In [1] a completely connected HMM was first proposed for time series modeling; when applied to the speech signal the model showed potential segmentation properties in terms of broad phonetic classes, and promising performances where obtained in speaker recognition applications. In that paper a phisical interpretation of the model as a vocal tract model was proposed, capable of only K (K = 5) different configurations that alternate according to a statistical transition matrix (global model). Once in a state the model possesses the characteristics of a stationary process described by an all pole filter (local model). Following this approach the constraints among possible spectral sequences produced in natural speech can be represented. In [2] a larger EHMM model (K = 64) was proposed for low bit rate speech coding applications. The characteristics of EHMM allowed straightforward coding schemes, and superior speech coding performances were obtained versus classical Vector Quantization schemes. These results were confirmed in [3], where a still larger model (K = 256) has been trained over speech data uttered by 10 speakers; informal listening tests showed the consistency of the approach and the perceptual significance that underlies the EHMM framework. Further studies [4] investigated the use of EHMM as a spectral tracking tool, obtaining smoother spectral estimates of speech signals if compared to classical LPC spectral analysis. In [5] a first investigation on the use of HMM in speech synthesis was carried out; the combined use of EHMM and classical left to right HMM Models was experimented to provide statistical word models for speech synthesis applications. A set of different utterances of the same word are first represented in terms of state sequences of a K = 64 states EHMM. Then a 5 states LTR HMM model is trained on this data in order to obtain a compact word model accounting for spectral sequence representation and duration modeling. In this paper we make a further assessment of EHMM model's ability to represent phonotactical constraints of language, introducing an EHMM model which is able to synthesize speech, rather than encode it. It is expected that utterances generated by this model will have no linguistic identity but that they retain crucial informations for speech synthesis systems, concerning allowed speech sound transitions of the language used during the model's parameter estimation.

The rest of the paper is organized as follows. Section 2 describes the realization of the EHMM. Section 3 reports some of the EHMM features, e.g. transition matrix characteristics and information theory considerations about the resulting information source entropy. Section 4 describes the implementation of an Ergodic speech synthesizer and shows some example of synthetic speech produced by the model. Section 5 presents applications of the model in speech synthesis.

2. MODEL DEFINITION

As foreworded in the introduction an EHMM for speech consists of a set of K states, a set of K observation densities (for instance we adopted autoregressive gaussian observation densities), an K x K transition probabilities matrix, allowing every state pair sequence, and an initial probability vector, giving the a priori probability of being in each state at the beginning of the encoding process. The experiments here reported refers to a state cardinality of K = 64,128,256,512. The choice of LPC spectral representation is motivated by the fact that such local model is very well suited to represent either the multivariate continuous observation densities utilized by Hidden Markov Modeling of speech, as well as the parameters needed to drive a speech synthesizer.

To each state is associated a set of M prediction coefficients and a voicing parameter derived as in [11], constituting the control parameters for the synthesis filter, and the prediction coefficients autocorrelation function, needed for the re-estimation of the HMM densities parameters. The local model is thus defined by the set of continuous probability density functions defined, for an observation O_t at time t, as :

$$(1)\ f_i(O_t) = \frac{1}{2\pi^{\frac{M}{2}}} \left(\prod_{i=0}^{M-1} \frac{\beta_i}{\sigma^2}\right) e^{\left(\frac{-1}{2}\delta(R_{t,i})\right)} \frac{1}{\sqrt{(2\pi\sigma^2_{V,i})}} \cdot e^{-\frac{(V_t - mV_i)^2}{(2\sigma V_i)}},$$

whith

$$(2)\quad \delta(R_{t,i}) = r^{vi}(0)r^t(0) + 2\sum_{j=1}^{M} r^{vi}(j)\ r^t(j)$$

$r^{qi}(j)$ is the autocorrelation of LPC coefficients of state i, $r^{i}(j)$ is the autocorrelation of input speech frame,β_i are the eigenvalues of autocorrelation matrix, and m_{Vi}, σ_{Vi} are the parameters defining the gaussian voicing probability distribution for state i, which is supposed to be independent from the spectral density. The global model is completely defined by a set of K states $Q \equiv \{q_i\}$, a set of initial probability values $\Pi \equiv \{\pi_i = Prob(q_i^{t=0})\}$, a stochastic transition matrix $A \equiv \{a_{i,j} = Prob(q_i^{t-1}, q_i^{t})\}$, $1 \leq i,j \leq K$.

Model parameters are estimated by means of the Baum algorithm [6], on about 700 phonemically compact words [7] pronounced by a reference speaker. When model size approaches values as K = 64 or more, some modification to the classical Baum algorithm are required to obtain reliable estimates; we adopted a scaling technique following the approach described in [13]. Speech is sampled at a frequency of 10 KHz; linear prediction analysis has been performed on a frame length of 32 ms., with an 8 ms. frame displacement. The initial EHMM parameters are computed by means of the binary-splitting version of the Lloyd-Max vector quantization algorithm [8], using the Likelihood Ratio distortion measure, in order to be consistent with pdfs used in the stochastic model [9] as observation densities. The transition matrix is initialized by means of a smoothed co-occurrence count statistics of the VQ labels collected for the same training data.

3. MODEL FEATURES

Figure 1 reports, in a graphical form, the resulting transition probabiliy matrix A after 4 cycles of the Baum reestimation algorithm, for a K = 64 states model. The periodicity of the resulting transition probability estimates results from the binary splitting method adopted for the VQ initialization of the EHMM.

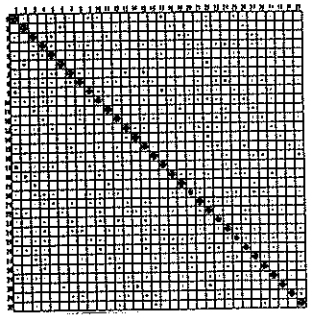


Figure 1 : Behaviour of transition probability matrix A; first 32x32 entries; dot size is proportional to the corresponding probability value (K = 64).

As the number of states of the model grows (k = 64 --> 512) the transition matrix becomes more sparse, as can be seen from table I, but the model preserves erodicity, i.e. each state can be reached form all the others. This can be verified multiplyng the transition matrix by itself $\log_2 K - 1$ times, and checking for the absence of all-zero columns. This assures that at least one path exists between all the states, whose maximum length, (i.e. the number of states visited along the path) is K.

Table 1.	
EHMM model size	**% number of transitions < > 0**
K = 512	7.06
K = 256	12.44
K = 128	27.92
K = 64	36.88

Table I : Number of non null entries in transition matrix A for different model's sizes.

One of the main transition matrix characteristics is that most of the possible state transitions possess very low probability values. If a phonetically compact set of utterances is used to obtain the parameter estimates, the transition matrix can thus be used to represent non allowed spectral transitions in natural speech. This can be seen, from a statistical point of view, as a method to represent the degree of order of the acoustic data. The statistical framework used allows us to define an absolute H^0 and conditional H^1 intrinsic entropies, represented by

$$(3)\ H^0 = -\sum_{i=1}^{K} \pi_i \log_2 \pi_i$$

$$(4)\ H^1 = -\sum_{j=1}^{K}\sum_{i=1}^{K} \pi_j\, a_{i,j} \log_2 a_{i,j}$$

where π_i and $a_{i,j}$ are respectively the absolute and conditional probability of state q_i . The true entropy evaluation theoretically requires knowledges of an infinite numerable order statistics, but yet a simple one-memory Markov model allows an adequate reduction of the source information rate : for a model with K = 256 we have $H^0 = 7.40$, $H^1 = 1.02$.

Finally we can observe, from figures 2 and 3 showing the number of allowed output and input transitions per state, that states with a small number of output transitions have a small number of input transitions. This can be interpreted as a confirmation of the model ability to represent acoustic constraints of the speech signal.

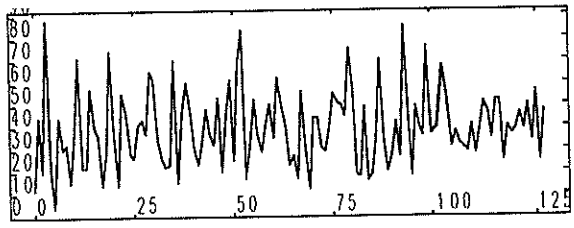


Figure 2 : Number of non null output transitions per state, K = 128.

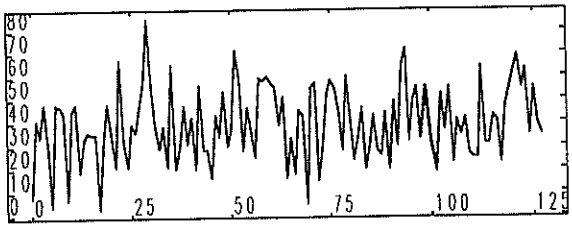


Figure 3 : Number of non null input transitions per state, K = 128.

4. ERGODIC SPEECH SYNTHESIZER

Using a similar EHMM model it is possible to generate output signals which shows some of the characteristics of natural speech. Signals produced from EHMM are very close to speech signals but do not possess a linguistic structure, and are generated in order to verify model's ability to retain information about possible spectral transitions. We use the following algorithm to generate a sequence of T random vectors to be used as excitation parameters on a classical LPC synthesizer :

- 1. Set the initial state index i generating a uniform ramdom number on $[1,K]$
- 2. Decompose R_i in the form $\Theta\chi\Theta'$, where Θ is the matrix of the eigenvectors of R_i and χ is the eigenvalues vector.
- 3. Generate a M dimensional normal deviate λ of zero mean and covariance χ
- 4. Set the output vector as $r_o = \Theta\lambda + r_i^m$, where r_i^m is the mean spectral vector of state i as obtained from the model estimation.
- 5. Generate a uniform deviate μ on $[0,1]$ and choose next state j as $j_{max} \ni \sum_{1}^{j} a_{i,j} < \mu$
- 6. Increment n_v (number of output vectors).
- 7. If $n_v < T$ then go to 2, else stop.

The sequence of vectors produced by the EHMM is transformed to a sequence of reflection coefficients and V/UV flag representation and sent to a classical LPC synthesizer. During the synthesis process an energy match module is active. This module assures the matching of speech output frame energy to the prediction gain of the EHMM model's state which generated that vector. The overall structure of the system is depicted in figure 4.

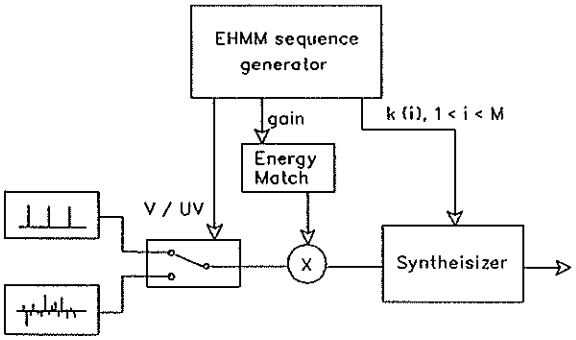


Figure 4 : Structure of the EHMM speech synthesizer.

A sample of the output produced by the EHMM is shown in figure 5, obtained with a model state cardinality K = 64.
Listening tests indicate that many of the acoustic events produced by the model can be perceived as short sequences of phonetic events of Italian language. The most frequently occurring acoustic events are related to steady state sounds as long fricatives and sonorant sounds. This is caused by the high values of transition probability diagonal matrix values corresponding to those tipical steady states sounds. Spectral transitions among speech sounds are clearly represented in the output signal; for instance we observed very good approximations of natural speech sound transitions among fricative, vowels and nasals.

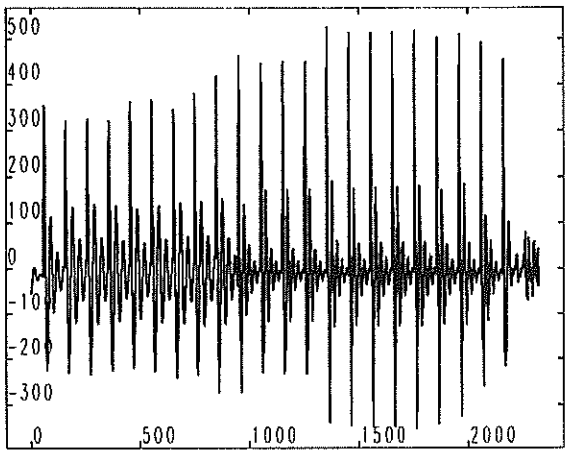


Figure 5 : Output speech from the EHMM

5. APPLICATIONS IN SPEECH SYNTHESIS

A first application relies on the afore mentioned features of the model as a spectral tracking tool which accounts for phonotactical constraints; this makes the model suitable as a steady states localization tool for speech synthesis diphone-like units extraction. A semi-automatic tool based on EHMM spectral tracking is currently under assessment. A tipical output of this system is shown in figure 6, where a natural speech utterance is shown together with a string of state indexes obtained by Viterbi alignement of the utterance with an EHMM whose state cardinality is K = 256. On the top of the figure the tool proposes steady states pointers (here represented by numbers from 1 to 3) resulting from the analysis of the state sequence emitted by the model.

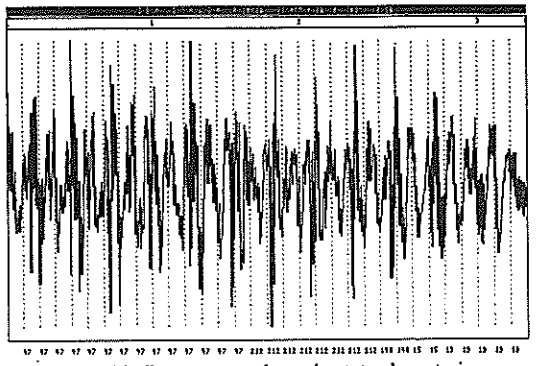


Figure 6. : Viterbi alignement and steady states hypotesis.

A similar approach can be used to provide an automatic segmentation of natural speech in sub-phonemic units to be used as basic units for applications which extends from speech synthesis to speech recognition. The EHMM framework can constitute an alternative technique to build an acoustic lexicon based on non allowed spectral sequences in natural speech, whith respect to Matrix Quantization techniques proposed in [14]
Finally, another application is depicted in [10]. The EHMM is used as the acoustic layer of statistical speech unit representations. In this work phone structure models are built

for the italian language. The structure model is represented in terms of states of an EHMM with state cardinality K = 256 whose connection is inferred from acoustic data. Several repetitions of the phone in different phonetic context are tracked by the EHMM, and a Left to Right HMM model is obtained with a technique similar to the one proposed in [15], although the use of the EHMM model allows a reduction of the complexity of the structure inference algorithm and the share of spectral state representations among phone models.

6. CONCLUSIONS

An EHMM model for speech signal phonotactical constraints organization representation has been presented. An ergodic speech synthesizer has been built to verify model's ability to represent such kind of information. Applications of the model in speech synthesis has been illustrated. Future work will extend the EHMM model emission probability density in order to take into account successive feature vectors, as proposed in [16], and investigate about the extension of EHMM in the direction of 2_{nd} order Markov processes.

REFERENCES

[1] A.B.Poritz, **Linear Predictive HMM and the Speech Signal,** *IEEE Proceedings of ICASSP, Paris,* 1982

[2] E.P.Farges, M.A.Clements, **HMM applied to very low bit rate speech coding,** *IEEE Proceedings of ICASSP, Tokyo,* 1986

[3] A.Falaschi,M.Giustiniani,P.Pierucci, **A Finite States Markov Quantizer,** *IEEE Transactions of ICASSP, Albuquerque, USA,* April 1990

[4] M.Verola, **Un Modello di Markov per la stima delle proprieta' spettrali** *Tesi di Laurea in Ing. Elettronica, Univ. Roma I,* 1989

[5] A.Falaschi,M.Giustiniani,M.Verola, **A Hidden Markov Model Approach to Speech Synthesis,** *Proceedings of EUROSPEECH, Paris,* September 1989

[6] L.E.Baum, **An inequality and associatel maximization technique** *Inequalities, vol III, pp.1-8,* 1972

[7] A.Falaschi, **Automatic selection of phonologically compact phrases,** *Proc. of First SFA Congress, Lyon, France,* April 1990

[8] B.H.Juang, **Design and Performance of Trellis Vector Quantizers** *IEEE Transactions on ASSP-36, N.9,* September 1989

[9] A.Buzo, A.H.Gray,Jr., R.M.Gray, J.D.Markel, **Speech coding based upon Vector Quanization,** *IEEE Transactions on ASSP-28, pp.562-574,* October 1980

[10] A.Falaschi,P.Pierucci **Some experiments on HMM structure inference,** *this volume.*

[11] M.Giustiniani **A new algorithm for pitch detection,** *Proceedings of EUROSPEECH, Paris, France,* September 1989

[12] B.H.Juang, **On HMM and Dynamic Time Warping for speech recognition** *AT&T BLTJ, Vol 63, N.7* September 1984

[13] P.F.Brown, **The Acoustic-Modeling problem in ASR,** *Phd Thesis, Carnegie Mellon University, USA* April 1987

[14] C.H.Lee,F.K.Soong,B.H.Juang, **A segment model based approach to speech recognition,** *IEEE Proceedings of ICASSP, New York, USA,* April 1988

[15] H.Rulot,N.Prieto,E.Vidal, **Learning accurate finite-state structural models of words** *IEEE Proceedings of ICASSP, Glasgow, UK,* April 1989

[16] C.J.Wellekens, **Explicit time correlation in HMM models for speech recognition,** *IEEE Proceedings of ICASSP, Dallas, USA,* April 1987

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

An Algorithm for Automatic Formant Extraction in Continuous Speech

O. Schmidbauer
Siemens AG, ZFE IS KOM31

Otto-Hahn-Ring 6, D-8000 München 83

This paper describes an algorithm for formant extraction in continuous speech which works under 'online' conditions in a speech recognition system. This implies the necessity that the algorithm is robust and avoids crucial descision errors. The algorithm initially preclassifies the speech utterance into 7 broad phonetic categories which correspond to manner of articulation. This allows to use two different algorithms for formant estimation which are specially tailored to the temporal and spectral properties of formants in vowel-like and consonant-like regions of speech. Speech recognition experiments show that formant-based parameters are a powerful feature set for speech recognition. The results show that the formant-based parameters can compete with other feature vectors for speech recognition.

1 Introduction

Formants are defined as the characteristic resonance frequencies of the vocaltract ordered by frequency. They appear as prominent peaks in the short-time spectra of speech. During speech production formants change their frequency values according to different vocaltract shapes, i.e. formants are direct acoustic correlates to the movement of the articulators and therefore may be used as active information in acoustic-phonetic decoding. Thus formants have become a standard in phonetics for describing complex acoustic-articulatory or acoustic-phonetic relations.

For automatic speech recognition formants also seem to be ideal parameters, but so far they have not become a standard feature set in automatic speech recognition. The reason is, that automatic formant extraction is not a trivial problem. Already existing algorithms for automatic formant extraction, e.g. /Laf80/, /McC74/ show the evidence that formant extraction without any errors is impossible. The importance of formants for the phonetic characteristics of speech becomes evident by the fact that errors in formant extraction cause severe recognition errors.

The paper is organized in following order. The next chapter briefly introduces into the problem of automatic formant extraction and will specify the problem. Then the different parts of the algorithm for automatic formant extraction are described. The last section reports about some recognition results achieved with formant parameters compared to results achieved with other feature sets.

1.1 Specification of the problem

Formants basicly differ from commonly used feature sets in speech recognition. They are defined by articulatory phonetics as resonance frequencies of the vocaltract, ordered by frequency, but they are not defined by a mathematical method, which allows to calculate them directly from the speech wave. Formants only may be calculated indirectly via peaks or roots of the power spectrum. Specifying automatic formant extraction in terms of estimation theory, we can formulate the following problem (see figure 1):

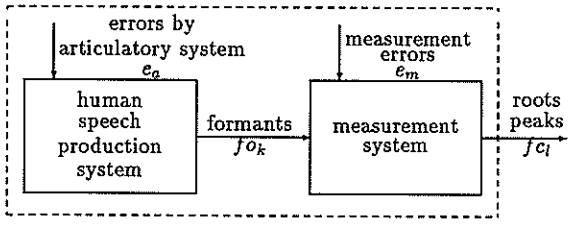


Figure 1: Formant estimation problem

Suppose the peaks and roots p_l of the power spectrum are the only data which can be measured and which give us some information about the unknown quantity 'formants' fo_k inside the system. So, depending on p_l we have to make an estimate for the formants $\hat{f}_k(p_l)$ that the estimation error $E = \hat{f}o_k(fc_l) - fo_k$ is small. 'Small' may be defined in several ways, leading to different estimation methods. Unfortunately this estimation process is heavily influenced by two different noise sources e_a and e_m:

- the errors caused by e_a have their origin in the articulatory system. It confuses the formant order by introducing zeros into the vocaltract transferfunction, depending on the geometry of the vocaltract shape. Thus some formants may highly be damped and therefore may not be detectable.
- noise source e_m causes errors within the measurement system; e.g., as the fundamental frequency is superseeded to the short-time spectra, prominent pitch peaks may be confused with formant candidates.

Existing methods for automatic formant extraction normally try to map the found peaks or roots of short-time spectra to formants by temporal smoothness criteria, e.g.

/Laf80/, /McC74/ and /Sch84/. The background for these procedures is the assumption that, due to the inertia of the articulators, the temporal behaviour of real vocaltract resonances is indicated by continuity.

But exclusively using smoothness criteria is an oversimplification and leads to crucial errors in formant extraction, as in pinciple formant frequency jumps are equally significant as continuous formant tracks in an articulatory-phonetic sense, e.g. in vowel-nasal transitions.

2 The algorithm for formant extraction

The algorithm for formant extraction presented in this paper is closely tailored to the spectral and temporal nature of formants. The method combines acoustic-articulatory knowledge with statistic pattern recognition methods. The algorithm works in the spectral domain with 128-point FFT spectra, bandwidth *8kHz*, calculated every *10ms*. The spectra are calculated via a 16-th order LPC analysis with a *20ms* Hamming window. The formant candidates are determined by peak-picking and root solving.

The algorithm is to be divided into four steps: 1) spectral analysis and preclassification into broad categories of manner of articulation, 2) formant identification *(FID)* in vowel-like segments of speech without smoothness criteria, 3) formant tracking *(FTR)* in VC- and CV-segments with smoothness criteria and 4) preparation of formant parameters for spech recognition.

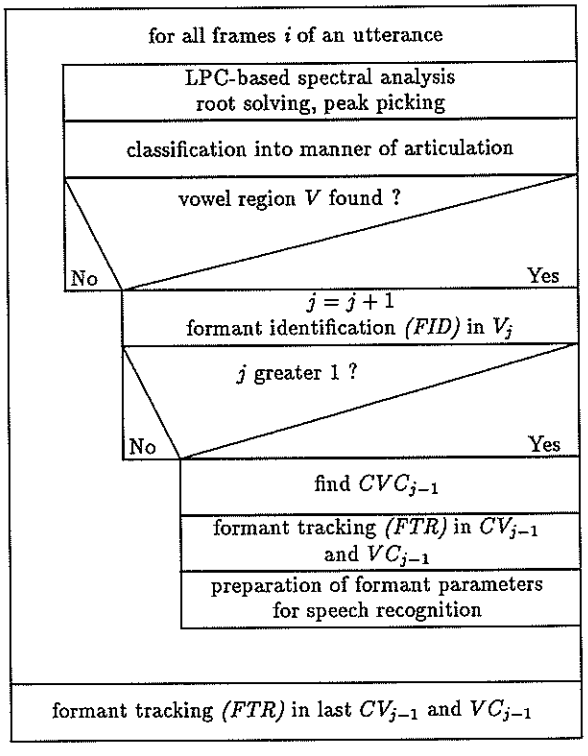


Figure 2: Flow graph of the formant-extraction algorithm

The temporal organization of the algorithm is illustrated in figure 2. The algorithm consists of three nested loops: The outermost loop calculates the spectrum and the candidates for formants (roots, peaks) and classifies the signal into 7 broad phonetic categories. Then the formants of vowel-like speech segments are estimated. The inner loop continues the formant trajectories from the vowel-like segments to the consonant-like regions. Finally formant based parameters which are used for recognition experiments are derived (e.g. the logarithmic energy contained in the formants).

2.1 Preclassification into 7 broad phonetic classes

The speech signal is preclassified into 7 broad phonetic categories *(silence, weak fricative, strong fricative, voiced plosive, nasal, sonorant, vowel)* which correspond to manner of articulation. This is done due to the assumption, that there is no overlapp of formant frequencies in segments with constant manner of articulation. This makes the following steps of the presented procedure, especially step 2 formant identification, more easily. Classification into categories of manner of articulation is performed by mixture density Hidden-Markov-Models (HMM) similar to /Ney88/, using very simple acoustic features like energy contour, zero-crossings rate, low frequency energy (up to 1000 Hz) and the ratio of high to low frequency energy.

2.2 Formant identification in vowel-like segments of speech

Initially, formants are extracted in vowel-like (V-like) segments, as formants are more prominent in vowels than in consonants. The main task of this step is to allocate formant candidates to formants, taking into account that formants may be missing over the whole part of a V-like segment. Within the V-like portions of speech a maximum number of M_{fc} formant candidates is calculated. M_{fc} is usually set to half the number of LPC roots minus one.

Formant identification is based on finding the dominant formant regions within a segment. To perform this, a histogram of formant candidates over frequency is generated, and then approximated with mean values and variances of M_{fc} clusters. The procedure itself consists of three steps: (1) initialization of the cluster procedure, (2) clustering and (3) classification of the formant candidates into formants by a mean sqare estimator.

(1) Initialization of the cluster procedure: To initialize the segment specific formant clusters, we need the mean value m_{fc_l} and variance σ_{fc_l} of the formant candidate frequencies x_{fc_l} over all N_V frames i of a vowel-like segment:

$$m_{fc_l} = \frac{1}{N_V} \sum_{i=1}^{N_V} x_{fc_l}(i)$$

$$\sigma_{fc_l} = \frac{1}{N_V} \sum_{i=1}^{N} (x_{fc_l}(i) - m_{fc_l})^2$$

Assuming that we already know m_{fo_k}, the speaker specific long term mean frequency value of formant fo_k, the center frequencies m_{c_k} of cluster c_k are initialized to:

$$m_{c_k} = \frac{m_{fc_l} + m_{fo_k}}{2}$$

(2) Cluster procedure: The k-means cluster procedure is used to calculate M_{fc} cluster centers. The resulting clusters c_k, ordered by frequency, characterize the segment specific formant frequency regions. They are defined by the mean and variance of M_{c_k} formant candidate frequencies which belong to the $k-th$ cluster:

$$m_{c_k} = \frac{1}{M_{c_k}} \sum_{x_{fc_l}(i) \in c_k} x_{fc_l}(i)$$

$$\sigma_{c_k} = \frac{1}{M_{c_k}} \sum_{x_{fc_l}(i) \in c_k} (x_{fc_l}(i) - m_{c_k})^2$$

Using a sufficient number of speech frames (about 1 minute), the long term formant mean frequencies m_{fo_k} and variances σ_{fc_l} are calculated by the same cluster procedure.

(3) Classification: As it is assumed that there is no overlapp between formant frequency regions within a V-like segment, the computed formant distributions are used to classify the formant candidates into formants. The classification procedure maximizes the probability $p(x_{fo_k}(i) \mid x_{fc_l}(i))$ over all $l = 1,..,M_{fc}$. I.e. the probability that a measured peak frequency $x_{fc_l}(i)$ at time i belongs to formant fo_k, when it was measured as formant candidate fc_l. The probability may be written as:

$$p(x_{fo_k}(i) \mid x_{fc_l}(i)) = \frac{1}{\sqrt{2\pi\sigma_{c_k}}} exp^{-\frac{1}{2}(x_{fc_l}(i) - m_{c_k})^2}$$

Applying the mean square error criterion /Lew86/ to the estimation of formant frequencies leads to the following equation: The resulting frequency x_{fo_k} of formant fo_k is given by the sum of the segment specific mean frequency value m_{c_k} plus the difference of the nearest formant candidate to m_{c_k}, weighted by the maximized probability $p_{max}(x_{fo_k}(i) \mid x_{fc_l}(i))$:

$$x_{fo_k}(i) = m_{c_k} + p_{max}(x_{fo_k}(i) \mid x_{fc_l}(i))\ (x_{fc_l}(i) - m_{c_k})$$

2.3 Formant tracking in VC- and CV-segments

This part of the algorithm continues the formants of the vowel-like (V) segments into neighboured consonantal (C) speech segments, i.e. formant tracking works on CV- and VC-segments. The CV- and VC-segments are well defined by preclassification. They constist of a $30ms$ vowel-like part and of a directly neighboured C-segment. As the formants of the V-region are already known, this part of the algorithm has the task to correct corrupted formant tracks by smoothness criteria. A non-linear smoothing algorithm based on dynamic programming was choosen for this task. This method is able to keep frequency jumps in some formants by optimizing the overall smoothness of the formant tracks.

The smoothness of the trajectory of formant f_k is measured by a cost function $c_k(l,i \mid h,i')$. The cost function measures the deviation of formant candidates to the trajectory of formant fo_k. Assuming that the formant candidates $fc_l(i)$ in frame i and $fc_h(i')$ in frame i' belong to the trajectory of formant fo_k at time i, the costs are given by:

$$c_k(l,i \mid h,i') = \underbrace{[\overbrace{|x_{fc_l}(i) - x_{fc_h}(i')|}^{1} C_1 + \overbrace{(i-i')C_2}^{2}]}_{3} [p(x_{fo_k}(i) \mid x_{fc_l}(i))\, p(x_{fo_k}(i) \mid x_{fc_h}(i')) - 1]$$

with $i' = i-1,..i-4$; $l,k = 1,..M_{fc}$; C_1 and C_1 being constants.

The cost function consists of three main terms: The first term corresponds to the frequency distance in Hz, the second term measures the temporal distance between the formant candidates and the third one is a weighting term which corresponds to the reverse probability that the formant candidates belong to formant fo_k. The function accepts small values for smooth and large values for corrupted trajectories.

The optimization criterion for the allocation of formant candidates to formants is given by the next formula. The criterion states that the total error E given by the sum of the costs over all frames N_{VC} for a VC-, N_{CV} for a CV-segment respectively, has to be a minimum:

$$E = min = \sum_{k=1}^{M_{fo}} \sum_{i=1}^{N_{VC,CV}} c_k(l,i \mid h,i')$$

over all l,k and i'.

This equation can be elegantely solved by dynamic programming. A solution for this problem is presented in /Sch89/.

2.4 Preparation of formant parameters for spech recognition

The formant parameter set which is used in our speech recognition experiments consists of 7 formant frequencies and of two energy terms for each formant, i.e. totally 21 parameters. The energy terms correspond to the logarithmic power which is contained in the frequency region starting from a formant's center frequency and going to the right mr and left minimum ml in the spectrum. With $s(x)$ being the log. power at frequency x, the energy to the left pl_{fo_k} and right side pr_{fo_k} of a formant center is calculated by:

$$pl_{fo_k} = \int_{x=x_{fo_k}}^{x_{fo_k}\ to\ mr} s(x)$$

$$pr_{fo_k} = \int_{x=x_{fo_k}}^{x_{fo_k}\ to\ ml} s(x)$$

The 21 formant parameters $x_{fo_k}, pl_{fo_k}, pr_{fo_k}$ are finally normalized to the speakers mean values and variances. With $v_k(i)$ now being one the 21 formant parameters at time i, the normalized formant parameters $n_k(i)$ are calculated by:

$$n_k(i) = \frac{v_k(i) - m_{v_k}}{\sigma_{v_k}}$$

Expressed in filter bank terminology: The resulting parameters which are used for speech recognition are filterbank coefficients, where the filter channels have variable center frequencies and bandwidths.

3 Experimental Results

The presented algorithm for automatic formant extraction was tested with speech material of 3 speakers (each with 2 versions of 100 phonetically balanced sentences, i.e. about 10 minutes of continuously spoken speech per speaker). The extracted formant parameters were used for classifying the speech signal into 14 categories of *place of articulation* as they are defined in table 1.

abbreviation	articulatory category	related phones
SI	silence	si
GL	glottal	? h
VE	velar	ng x g k
PA	palatal	j ch sch
AL	alveolar	l r n el en
DA	dental-alveolar	s z d t
LD	labio-dental	v,f
LA	bilabial	b p m em
V1	–	U u O o
V2	–	A a er
V3	–	e ae
V4	–	Y y OE oe
V5	–	eh I i

Table 1: Categories of place of articulation and allocation to phones

For each articulatory category we built continuous mixture density Hidden Markov models as they are described in /Ney88/ and /Sch89/. One version of 100 sentences was used for training, the other version was used for testing. The recognition results on *10ms frame level* are shown in Table 2. The pairs of numbers show the class specific mean recognition rates (left) and the overall frame recognition rates. The formant parameters were compared to a 16-component cepstral vector and to a 64-component feature vector as it is used in /Pae88/. It consists of 30 mel-coefficients and of difference and curvature coefficients, taking into account $\pm 40ms$ of context. The overall mean recognition rate over three speakers for 21 formant parameters is 74.9 percent, for the cepstrum 67.4 percent and for the 64-component mel difference vector 78.5 percent. The results show that the formant vector outperforms the cepstral vector (about 7 percent higher). The recognition performance compared to the the 64-component vector is about 4 percent lower, but it has to be taken into account that the dimensionality of the formant vector is three times lower than for the 64-component vector and that no temporal context was considered for calculation.

speaker	21 formant parameters	16 cepstral coefficients	64 mel difference coefficients
SM	74.7 / 84.9	66.8 / 80.3	78.4 / 86.7
TM	74.2 / 84.1	67.3 / 79.5	78.0 / 86.7
BR	75.9 / 86.0	68.1 / 81.1	79.2 / 87.9

Table 2: Recognition rates in percent for three different speakers and three different feature sets

occur.		SI	GL	VE	PA	AL	DA	LD	LA
7693	SI:	97.1	0.0	0.6	0.1	0.0	1.7	0.2	0.0
821	GL:	1.0	48.1	7.3	0.5	2.1	17.2	2.4	0.1
1673	VE:	2.3	0.0	71.8	0.4	1.5	14.2	2.4	1.9
937	PA:	0.0	0.5	0.5	92.5	0.6	2.3	0.9	0.0
3002	AL:	0.1	0.8	0.7	0.0	77.8	2.7	0.1	5.6
4052	DA:	2.0	0.4	1.4	1.6	0.4	90.6	1.6	0.5
1178	LD:	0.9	0.4	0.7	0.0	0.4	4.3	82.9	2.9
1085	LA:	0.4	0.6	1.7	0.2	13.3	17.0	1.9	56.5

occur.		V1	V2	V3	V4	V5	V6
1052	V1:	88.9	0.2	3.4	1.8	0.0	0.2
3063	V2:	6.5	80.9	5.1	0.5	2.5	0.5
1538	V3:	4.9	10.5	59.0	3.8	4.0	4.6
383	V4:	2.1	1.6	6.8	82.2	1.6	0.8
1031	V5:	0.1	2.1	8.1	0.0	71.3	11.7
1218	V6:	0.0	0.0	6.8	0.4	5.1	79.8

Table 3: Confusion matrices for place of articulation categories for speaker BR. Confusions between vowel- and consonant categories are not shown. Therefore the row sum is not 100.

Table 3 shows typical confusion matrices for the formant vector. It can be seen that the recognition performance of the vowel categories is about 80 percent (besides the /schwa/-categorie V3). The consonantal categories GL and LA are badly recognized. This may be due two reasons: 1. The fewer occurances in the training material (especially for GL) compared to other categories, and 2. The context-dependency of the categories (for GL and LA). Due to this observations we will introduce context-dependency on the level of formant-parameters and on the level of articulatory modelling with Hidden-Markov models.

4 References

/Laf80/ Laface P., "A Formant Tracking System toward Automatic Recognition of Speech", Signal Processing 2, S. 113-129, North-Holland Publishing Company, 1980

/Lew86/ Lewis F.L., "Optimal Esimation", John Wiley and Sons, New York, 1986

/McC74/ McCandless S.S., "An Algorithm for Automatic Formant-Extraction using Linear Prediction Spectra", IEEE Trans. ASSP, Vol.22, S.135-141, 1974

/Ney88/ Ney H., Noll A., "Phoneme Modelling Using Continuous Mixture densities", IEEE Proc. ICASSP, pp. 437-440, New York 1988

/Pae88/ Paeseler A., Ney H., "Phoneme-Based Continuous speech recognition results for different Language Models in the 1000-word SPICOS system", Speech Communication 7, North Holland, pp. 367-374, 1988

/Rig88/ Rigoll G., "Formant-Tracking with Quasilinearization", IEEE Proc. ICASSP, pp. 437-440, New York 1987

/Sch84/ Schmidbauer O., "Segment-orientiertes Formant-Tracking mit dynamischer Programmierung", Tagungsband DAGA: Fortschritte der Akustik, S. 797-800, Darmstadt, 1984

/Sch89/ Schmidbauer O., "Robust Statistic Modelling of Systematic Variabilities in Continuous Speech Incorporating Acoustic- Articulatory Relations", IEEE Proc. ICASSP, pp. 616-619, Edinburgh, 1989

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Formant and Anti-Formant Tracker Using Time Weighted ARMA Method

Nobuhiro MIKI, and Nobuo NAGAI

Hokkaido University, Sapporo 060

An ARMA parameter estimator with weighting functions, a simple formant (anti-formant) tracking algorithm using the estimated ARMA parameter sequence, and a software environment on the Apollo's work station are proposed. The weighting filter with a second ARMA type is used for frequency weight, and in the recursive parameter estimation the parameters are obtained with the pitch synchronized weighting. The experimental results are shown for the real speech.

1. Introduction

Power spectrum estimation in the short-time is importance in the speech information processing such as the speech recognition from feature-extraction of the spectral domain, the speech synthesis using Linear Prediction Coding (LPC) or other parametric cordings, and also research of speech science. The formant is the resonant frequency of the vocal tract, and is not the frequency of spectral peak, thus it is important to estimate the Vocal Tract Transfer Characteristics (VTTC). If we can get VTTC as an ARMA parameter representation, the formants are estimated from the root of the AR polynomial, and the anti-formants are estimated from the root of the MA polynomial. There are many algorithms for spectrum estimation and formant tracking method, but most of the algorithms are based on the frame-based LPC analysis which has disadvantage for estimation accuracy. There are some reasons for inaccuracy that

(1) for nasal sounds the influence of anti-formants has been neglected

(2) LPC spectrum is averaged spectrum in the frame

(3) the estimated spectrum is affected from the periodic excitation of vocal cords, especially for high pitch vowels

(4) the estimated first-formant is affected from the preemphasis.

In the above problems, the signal model is the AR model, the frame-based model, and no excitation model. Since a time-varying ARMA model is employed in our model, it is possible to estimate not only anti-formants but also time-varying formants without the influence of anti-formants. To resolve the above problem (3), we introduce the excitation model from estimated innovation sequence and a time weighting method for the ARMA estimation, and for the high pitch vowel of female the tracking is possible stably. Since the excitation model is used, the variation of pitch can be traced also in the algorithm. To reduce the preemphasis problem (4), we employ the emphasis filter with non-minimum phase in the algorithm. In the ordinary preemphasis algorithm, the first order MA filter is used, but the first formant in the low frequency region is estimated inaccurately because the characteristics of the filter in the low frequency region is very sharp.

Since it is very difficult to find the formant trace form the AR polynomials, the knowledge of speech production rules has been used. But the knowledge based system for the formant tracker has to use large memory and requires long computation time. Thus, if the formant tracker can be builded up with the simple tracking algorithm for the AR parameters, it is useful to analyze the speech signal for high-speed processing. We propose a simple algorithm for the formant and anti-formant tracking from the ARMA parameters.

Some experimental results are shown for the formant and anti-formant tracking of the real speech, it is seen that the tracking is performed with high accuracy and high speed. The tracker is runing on an Apollo's work station, the results are displayed on the multi-window, and the person-machine interface for speech analysis is realized with user friendly interfaces.

2. ARMA Estimation Algorithm

The speech production model (SPM) is defined by the following ARMA process:

$$x_k = -\sum_{i=1}^{m} a_i(k)x_{k-1} + \sum_{i=1}^{n} b_i(k)u_k$$

where the X_k is a speech signal from the mouth, and u_k is an excitation signal. Since the speech signal x_k is usually observed through an LPF for the A/D conversion, the measurement system is modeled by an ARMA transfer function, and the observed signal y_k is defined by

$$y_k = -\sum_{i=1}^{s} c_i y_{k-i} + \sum_{i=1}^{l} d_i x_{k-i} + x_k .$$

The coefficients c_i and d_i can be previously determined by the measurement of the characteristics of the LPF.

If the weighting matrix $\mathbf{W}$ and the vector $\mathbf{w}$ are determined from the coefficients c_i and d_i , the ARMA parameter estimation algorithm is shown as the following recursive form:

$$\hat{y}_k = \mathbf{H}_k^t(\mathbf{W}\hat{\mathbf{P}}_k + \mathbf{w}) + \hat{u}_k \quad \nu_k = y_k - \hat{y}_k$$

$$E_k = \mathbf{H}_k^t \mathbf{W} \mathbf{F}_{k/k-1} \mathbf{W}^t \mathbf{H}_k + 1 \qquad \mathbf{F}_{k/k-1} = \mathbf{F}_{k-1} + \hat{\sigma}_{\varepsilon_{k-1}}^{-2} I$$

$$\hat{\mathbf{P}}_{k/k} = \hat{\mathbf{P}}_{k/k-1} + \mathbf{F}_{k/k-1} \mathbf{W}^t \mathbf{H}_k E^{-1} \nu_k$$

$$\mathbf{F}_k = \mathbf{F}_{k/k-1} - \mathbf{F}_{k/k-1} \mathbf{W}^t \mathbf{H}_k E^{-1} \mathbf{H}_k^t \mathbf{W} \mathbf{F}_{k/k-1}$$

$$\hat{u}_k = \begin{cases} y_k - \mathbf{H}_k^t(\mathbf{W}\hat{\mathbf{P}}_{k/k-1} + \mathbf{w}) \\ \nu_k / E_k \end{cases} \qquad \hat{\sigma}_k^2 = \overline{\nu_k^2}$$

where $\mathbf{H}_k$ and $\hat{\mathbf{P}}_k$ are

$$\mathbf{H}_k^t = [-y_{k-1} \cdots -y_{k-(n+s)} \hat{u}_{k-1} \cdots \hat{u}_{k-(m+l)}]$$

$$\hat{\mathbf{P}}_k^t = [\hat{a}_1(k) \ldots \hat{a}_n(k) \hat{b}_1(k) \ldots \hat{b}_m(k)]$$

The above algorithm (WMIS) and the related algorithms have been proposed by the author. [1-3]

Time weighted algorithm

In the algorithm WMIS, the error variance of the estimated parameters increases in the duration after the detection $\nu_k^2 > c\hat{\sigma}_{k-1}^2$, because the error covariance matrix $\mathbf{F}_k$ is changed with large variation of ν_k . Thus the variation of $\hat{\mathbf{P}}_k$ is large after the detection of pitch pulse, and the estimated spectrum is shifted from a pattern to a pattern in each time up date. These variation is not adequate for formant (anti-formant) tracking from the estimated parameter sequence, and it is required to reduce the unnecessary variation of the parameters. We propose a time weighted algorithm for the estimated parameters for the reduction of the unnecessary variation as follows.

$$\hat{\mathbf{P}}s_k = \frac{1}{n_o} \sum_{i=0}^{T_s} W_{k-i} \left(\frac{\nu_{k-i}^2}{\hat{\sigma}_{k-i}^2} \right) \hat{\mathbf{P}}_{k-i}$$

where, $W_{k-i}\left(\frac{\nu_{k-i}^2}{\hat{\sigma}_{k-i}^2} \right)$ is a weighting function,

$$W_{k-i}\left(\frac{\nu_{k-i}^2}{\hat{\sigma}_{k-i}^2} \right) \approx 0 \quad \text{if } \nu_{k-i}^2 \geq c\hat{\sigma}_{k-i-1}^2$$

$$W_{k-i}\left(\frac{\nu_{k-i}^2}{\hat{\sigma}_{k-i}^2} \right) \approx 1 \quad \text{if } \nu_{k-i}^2 \ll c\hat{\sigma}_{k-i-1}^2 \quad W_{k-T_s}\left(\frac{\nu_{k-T_s}^2}{\hat{\sigma}_{k-T_s}^2} \right) \approx 0$$

From the above algorithm, the smoothed parameter $\hat{\mathbf{P}}s_k$ is obtained as the parameters with reduced influence of the unnecessary variation.

Selection of the weighting matrix

The weighting matrix $\mathbf{W}$ can be used for compensation of measurement errors from the measurement system. Since the matrix $\mathbf{W}$ in the eq. (3) represents the filter $D[z]/C[z]$, we can use the $\mathbf{W}$ as generalization of weighting in the frequency region. If it is assumed that the system generating y_k contains the component $D[z]/C[z]$, we can compose the estimator as Fig. 1 (a) or (b). In this figure, we can see that $D[z]$ or $C[z]$ are used as a pre-filter or a kind of weighting filters. If $D[z]/C[z]$ is a non-minimum phase filter, this has no inverse filter, and $C[z]/D[z]$ cannot realize as the pre-filter. However, using the configuration of Fig. 1, we can realize any weighting filters. When we use some recording equipments, amplifiers, and pre-filters for speech data

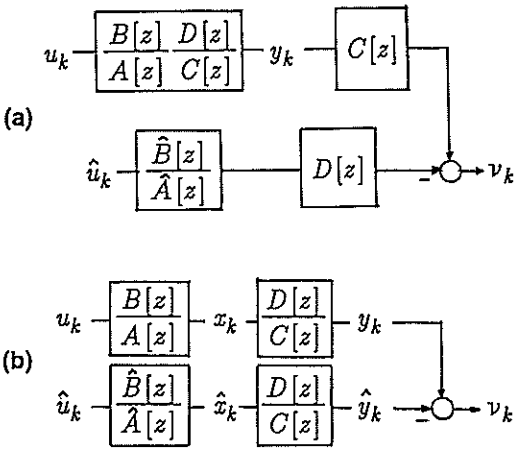


Fig. 1 ARMA parameter estimators

acquisition, since the speech waveform is distorted with the phase characteristics of these systems, a all pass filter of non-minimum phase is frequently required for phase compensation. In the spectra of voiced sound, the power of low frequency component is high, but the one of high frequency is low, thus a pre-emphasis fileter is used for numerical stabilization of the estimator. However since the pre-emphasis filter with one order has the characteristics with sharp cut off in the low frequency such as pitch frequency or first formant of low frequency, the estimated formant trace shows frequently lack of first formant for such Japanese vowel /u/ whose first formant exists in low frequency. In order to improve the above problem, we employ the second order $C[z]$ and $D[z]$ for the weighting whose characteristics is shown in Fig. 2. We see that this characteristics is more close to a straight line than that of first order filters.

3. Simple Formant (Anti-formant) Tracking Algorithm

We propose a formant (anti-formant) tracking algorithm which requires small computation cost. Using the recursive algorithm (WMIS, etc.), we can obtain the ARMA parameters from the real speech signal, thus the spectral peak (null) is computed as the root of the AR (MA) transfer function. But the all peaks or nulls may not always correspond with formants or anti-formants, and only the roots with the reasonable band width are needed.

As the first step, in the analysis frame the all roots are computed from the ARMA parameter, and the roots are sorted in order of the band width from small one to large one. In the second step, all the roots with extremely wide band width (> 450) are abandoned. Using the recursive algorithm with input estimation, we obtain the estimated parameters with some pole-zero cancellations. In the third step, the cancellations are searched, and these unnecessary pairs of formant and anti-formant are eliminated from the parameter. Lastly the remained roots are sorted in order of the frequency, and the formant (anti-formant) frequency and the band width is computed respectively.

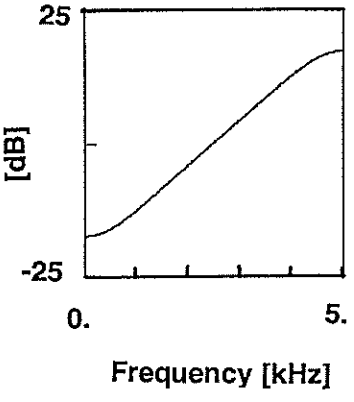


Fig. 2 Frequency response of pre-filter

The computation results are stored in the ASCII formated file, and one can check and see easily in the display window using the pad opening function by the pointing device the mouse. Using the speech signal file, the computation results of formants (anti-formants), and the specification for display functions, we can obtain the graphic representation of formant (anti-formant) traces with the signal waveform in the display window. The formant trace is represented as the sequence of the vertical bar such as Fig. 3. The length of the bar means the band width.

4. Experimental Results

In this experiment the roots of parameter are computed in each 4ms from the parameter sequence. The tracking results of formants (anti-formants) for Japanese vcv sound /onu/ are shown in Fig. 3. In Fig.3, we see that the anti-formant is moving from near the 2nd formant to near the 1st formant, and is crossing the 2nd formant. In the part of the nasal sound /n/, the band width of 1st formant and 3rd formant are varied from vowel like width to extremely narrow width. We can see good tracking of the 1st formant of /u/ , which exists in low frequency, in Fig. 3. From these results, we see that the tracking with the representation of band width is very useful for the feature extraction of nasal sounds in speech analysis. In the experiment of pitch detection, we see that the excitation pulse sequence can be extracted without false pulses .

5. Pitch Detection

The pitch detection as the search algorithm of excitation pulse is very important for speech analysis tools. Since it is impossible to measure the excitation signal as the input signal, the input signal is estimated as $\hat{u}_k$ in our algorithm. In this input

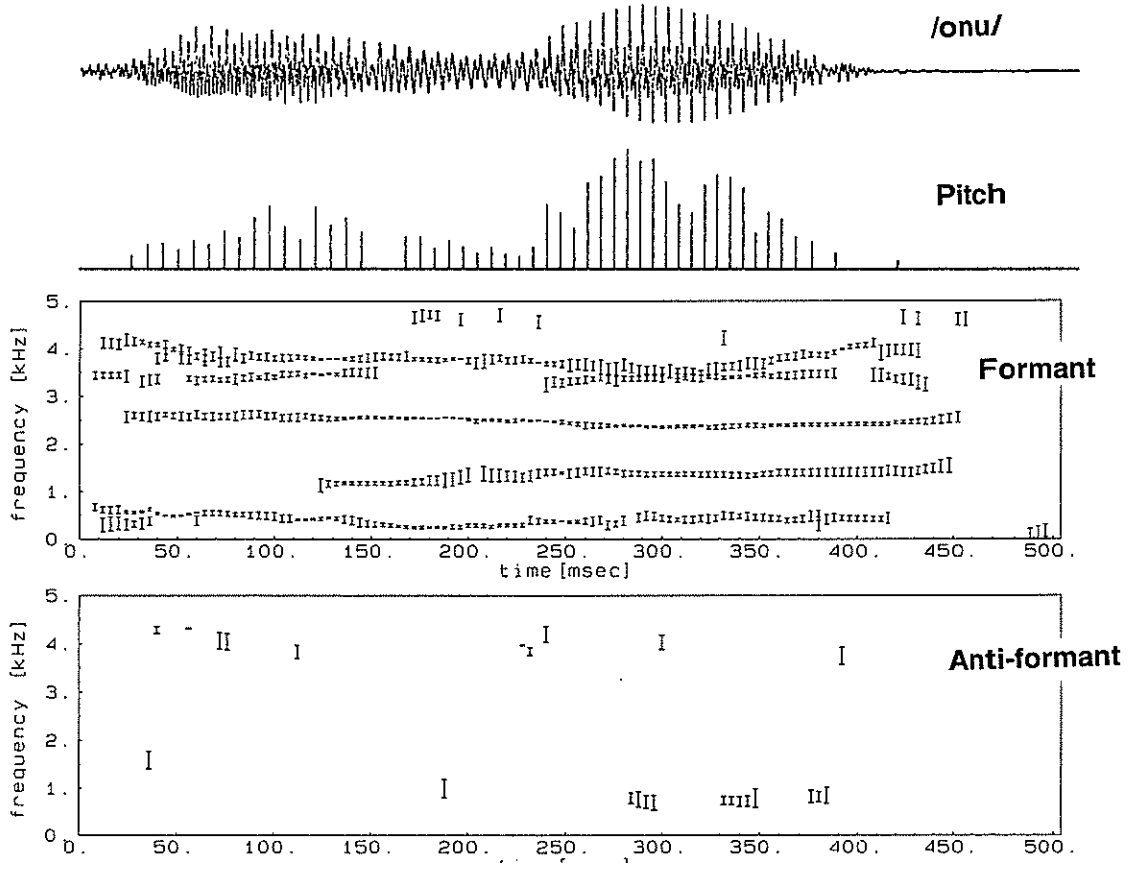


Fig. 3

estimation, we can estimate the excitation pulse from $\hat{u}_k$. In our pitch detector, we employ the estimator of pitch intervals. Using this estimator, we can reduce the detection error for the time of the excitation pulse; if the extracted pulse is regarded as the one within a irregular time interval, then the pulse is neglected, and the estimator searchs the other pulse in the time interval which is determined recursively in time update.

6. References

[1] N. Miki, Y. Miyanaga, S. Saga, and N. Nagai, "Spectrum and pitch estimation of speech using a time-varying ARMA estimation algorithm", IEEE ICASSP 85, Vol. 3, 29.13, pp.1133-1136, (1985).

[2] M. Serizawa, N. Miki, Y. Miyanaga, and N. Nagai, " A study of speech analysis based on a varying ARMA model", IECE, Japan, Rep. CAS87-102, Aug. (1987).

[3] N. Miki, S. Saga, Y. Miyanaga, and N. Nagai, "ARMA spectral estimation using weighted model identification systems", J. Acoust. Soc. Jpn. (E) 7,1 pp.21-28 (1986).

N. Miki, and M. Serizawa, "Spectral Tracking Using ARMA Estimation", The first Symposium on Advanced Man-Machine Interface Through Spoken Language, p113-121, Jan. (1988).

M. Serizawa, N.Miki and N.Nagai,"Recursive estimation of ARMA parameters based on a robust time-varying model for speech analysis", The Second Joint Meeting of ASA and ASJ, G14, Nov. 1988.

[6] Y. Miyanaga, N. Miki, and N. Nagai, "Adaptive identification of a time-varying ARMA speech model," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, pp.423-433, June (1986).

[7] N. Miki, M. Serizawa, and N. Nagai, "Formant and Anti-Formant Tracker Using Time Weighted ARMA Method", Proc. JTC-CSCC '89 pp. 362-366, June 1989.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

PITCH DETECTION BASED ON LOCALIZATION SIGNAL

Jean-Paul LEFEVRE and Gang FENG *

OROS
13 Chemin des Prés, 38240 Meylan, France

An accurate pitch detector is an essential component in a variety of speech processing systems. Though, a lot of algorithms have been published in the literature, none of them is quite satisfactory. In this paper, we demonstrate that the localization signal used in multi-pulse coders is a suitable approach for pitch detection. A detector, based on this localization signal, is presented and the achieved results are discussed.

1. Introduction

The problem of estimating the fundamental period of a speech signal, known as pitch detection, is an essential topic in a variety of speech processing systems. For example, it is well known that the pitch contour of an utterance is useful for recognizing speakers, for speech instruction to the hearing impaired, for noise cancelling systems and is required in almost all speech analysis-synthesis systems like vocoders or adaptive predictive coding systems.

Because of the importance of pitch detection, a wide variety of algorithms devoted to this topic have been proposed in the speech processing literature [1]. The difficulty in reliability when estimating the fundamental period of a speech signal is compounded by deviations from an ideal periodic signal. Although finding the period of a perfectly periodic waveform is straightforward, measuring the period of a speech waveform, which varies both in period and in the detailed structure of the waveform within a period, can be quite difficult. A second difficulty in measuring pitch period is the interaction between the vocal tract and the glottal excitation.

In the past, partially successful attempts were made to use the linear prediction residual signal for pitch extraction. More recently, the principle of pitch extraction from a sequence of selected samples from this residual has been introduced. In particular, it has been stated that the use of the pulses computed in a multi-pulse coder could give good performance for pitch detection [2].

In this paper, we present a pitch detection algorithm based on the localization signal often used for the pulse research in a multi-pulse coder. Section 2 introduces in more details this localization signal. In section 3, a complete description of the pitch detection algorithm, based on this signal, is given. Theoretical advantages of this approach as well as practical considerations, like smoothing procedures, are emphasized. Finally, section 4 provides a discussion of the achieved results.

2. The Localization Signal

The localization signal was firstly introduced by Atal and Remde [3] in order to determine the pulse locations in a multi-pulse excited linear predictive coder. The principle of such a coder consists in determining a few number of pulses (typically one pulse per ms), which will be used as excitation signal $\hat{e}(n)$ in the decoder, by minimizing the following criterion :

$$\sum_{n=0}^{N-1} | [e(n) - \hat{e}(n)] * h_W(n) |^2 \qquad (1)$$

* G. Feng is now with ICP-INPG, 46 Av. F. Viallet, 38031 Grenoble, France

where e(n) is the LPC residual signal within a frame of length N, and $h_w(n)$ is the impulse response of the perceptual filter whose transfer function H(z) is computed from the LPC inverse filter $A(z^{-1})$ by :

$$H(z^{-1}) = 1 / A(\gamma z^{-1}) \quad (2)$$

with $\gamma = 0.85$ (typical value).

From the sequential pulse research procedure, the localization signal T can be defined as the intercorrelation between the LPC residual signal and the autocorrelation of the perceptual filter impulse response $R_w(n)$:

$$T(i) = \sum_{n=i-L_w}^{i+L_w} e(n)\, R_W(n-i) \quad (3)$$

where L_w is the number of significant samples in the impulse response of the perceptual filter. In practice L_w is set equal to the LPC filter order plus a few points.

3. The Pitch Detection Algorithm

We have found that the two following features of the localization signal are particularly propitious to the pitch detection :

- There is few formant structure in the localization signal since it is derived from the LPC residual signal, which has a flatter spectrum than speech signal because of the LPC inverse filtering. Now, it is well known that a good separation of the vocal source excitation from the formant structure is indispensable for pitch detection.
- The localization signal is more smooth than LPC residual signal. In fact, formula (3) can be rewritten as :

$$T(i) = \sum_{n=i-L_w}^{i+L_w} e(n)\, R_W(i-n) \quad (4)$$

since $R_w(n)$ is a symetric function. This relation shows that the localization signal can be seen as the residual signal filtered by a filter whose impulse response is $R_w(n)$. For voiced speech, this filter acts typically as a low-pass one since $h_w(n)$ itself behaves as the impulse response of a low-pass filter.

From this localization signal which offers several quite good features, we have derived a pitch detection algorithm. The main principle of this algorithm lies in performing a peak detection on the autocorrelation of the localization signal. Although, in most cases, the first detected peak corresponds to the pitch value, there are always difficult situations in which the determination is not unique. So it is necessary to introduce a correction algorithm added to the main peak detector. Figure 1 shows the block diagram of the whole algorithm.

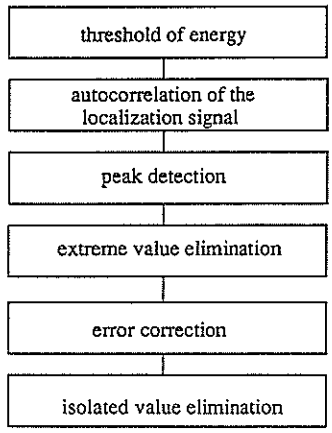


Fig. 1. - Block Diagram of the Algorithm

Firstly, a threshold of energy is used to exclude silences and weak signals which have never to be classified as voiced. The choice of this threshold is not critical. A value of -40 dB below the maximal signal amplitude works very well.

Then, the autocorrelation of the localization signal is computed. It may be useful to mention here that the frame length must be long enough to ensure the detection for large pitch values. Our experience shows that a frame length of 20 ms at least is a requirement.

The peak detection is then performed by comparing the autocorrelation function with a threshold. In the simplest case, the first peak can be used as pitch value. But in our algorithm, the two first peaks are stored in order to allow a subsequent correction procedure. The determination of the detection threshold is relatively critical. It appears that a constant threshold cannot give good results. Therefore, we have selected a variable threshold. At the

beginning, the threshold is set to 0.3. Then it linearly decreases down to a value of 0.15 during 60% of the frame length. After, this value is kept constant until the end of the frame. In practice such a variable threshold improves greatly the performance of detection.

After the peak detection process, peaks whose value is too small or too large are eliminated. In the implemented algorithm, these two limits are respectively fixed to 2.5 ms (400 Hz) and 17.5 ms (57 Hz).

The above described algorithm has been tested on numerous speech signals. Used as it is, without any correction procedure, it gives quite satisfactory results. Though a correction algorithm is always useful, it is very important to obtain at the beginning an efficient detector without correction. Indeed, no algorithm is able to perform a perfect correction, and moreover it may introduce some errors.

Anyhow, besides the above detection, a smoothing procedure is used to correct false values and to smooth the results out. The correction algorithm consists in finding, among the candidate values, their multiples and their halfed values, the value most close to :

- the value retained for the last frame, if it is a voiced one;
- the last value of the previous series of voiced frames, in an unvoiced/voiced transition.

In order to avoid propagation of errors, the correction must be carried out with many care. So, a correction is performed only in a case without any doubt, that is only if the previous value is classified "sure". On the contrary, no correction is performed and the maximal peak of the autocorrelation is used to determine the pitch value.

A value is classified "sure" only if it satisfies one of the three following conditions :

- the value is unique ;
- the distance between the two first peaks of the autocorrelation is large enough ;
- the value is the result of a correction.

The above correction thus ensures the continuity of pitch values. But it does not modify the voiced/unvoiced decision. Sometimes there remains some isolated voiced or unvoiced frames. A further algorithm is introduced to remove these isolated values. For a pitch value, if the previous and the next values are all zero, it will be set to zero. For a zero value, if the previous and the next values are not null, it will be filled by an interpolation.

4. Results and Discussion

Figure 2 gives two examples of the performed computations. All data correspond to one analysis frame. Graph a) depicts the original signal, b) the LPC residual, c) the localization signal and d) the autocorrelation of c). These examples have been chosen to illustrate the performance of the detector during transitions. One can see that the autocorrelation function has a quite good behaviour in these difficult cases, confirming the previous assumptions. From the autocorrelation, it is quite clear that the detection can be done without any difficulty.

Extensive evaluation tests have been carried out with male and female voices from different speakers. Comparisons performed with other pitch detectors (mainly AMDF and other correlation algorithms) have shown significant improvements, more especially during transitions, as discussed with the previous figure.

Figure 3 illustrates typical results achieved with this algorithm. Figure 3b displays the pitch contour obtained before the introduction of any smoothing or correction procedure, while figure 3c displays the final results including all the post processings. As can be seen on this last graph, almost all the classical pitch errors are removed, especially the isolated and doubled values.

Also, it should be mentioned that this algorithm is relatively low cost. All computations are carried out in the time domain. The LPC residual, then the localization signals can be efficiently computed. In fact, only the computation of the autocorrelation takes time with such an algorithm. Finally, this algorithm, using the localization signal as basis for the detection, is perfectly adapted to multi-pulse based coders, in which either a long term predictor or a pitch-based pulse research procedure depends on pitch values [4].

REFERENCES

[1] Hess, W.J., Pitch Determination of Speech Signals - Algorithms and Devices (Springer, Berlin, 1983).

[2] Bakamidis, S.G., Carayannis, G. and Skiadas, N., A New Pitch Detector Based on Preselected Information from the LPC Error Signal, ECST, Edinburgh, pp. 411-414, Sept. 1987.

[3] Atal, B.S. and Remde, J.R., A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates, Proc. Int. Conf. on Acoustics, Speech and Signal Process., Paris, pp. 614-617, May 1982.

[4] Lefèvre, J.P. and Feng, G., A New Algorithm for 4.8 kbit/s Speech Compression, VERBA, Rome, pp. 140-147, Jan. 1990.

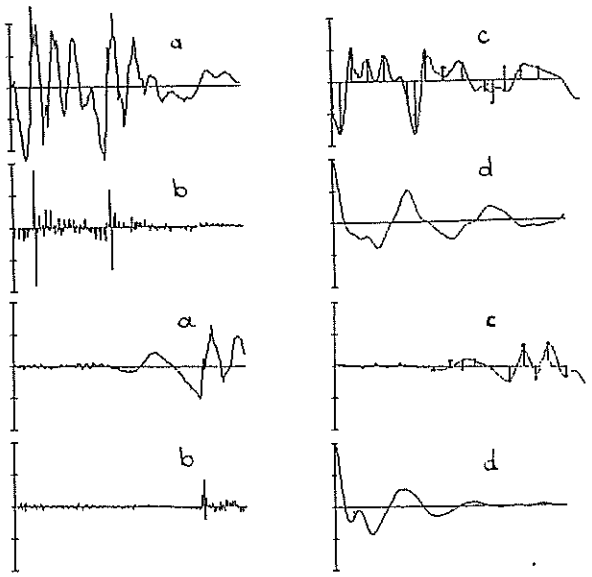


Fig. 2. - Two Examples of Computed Signals

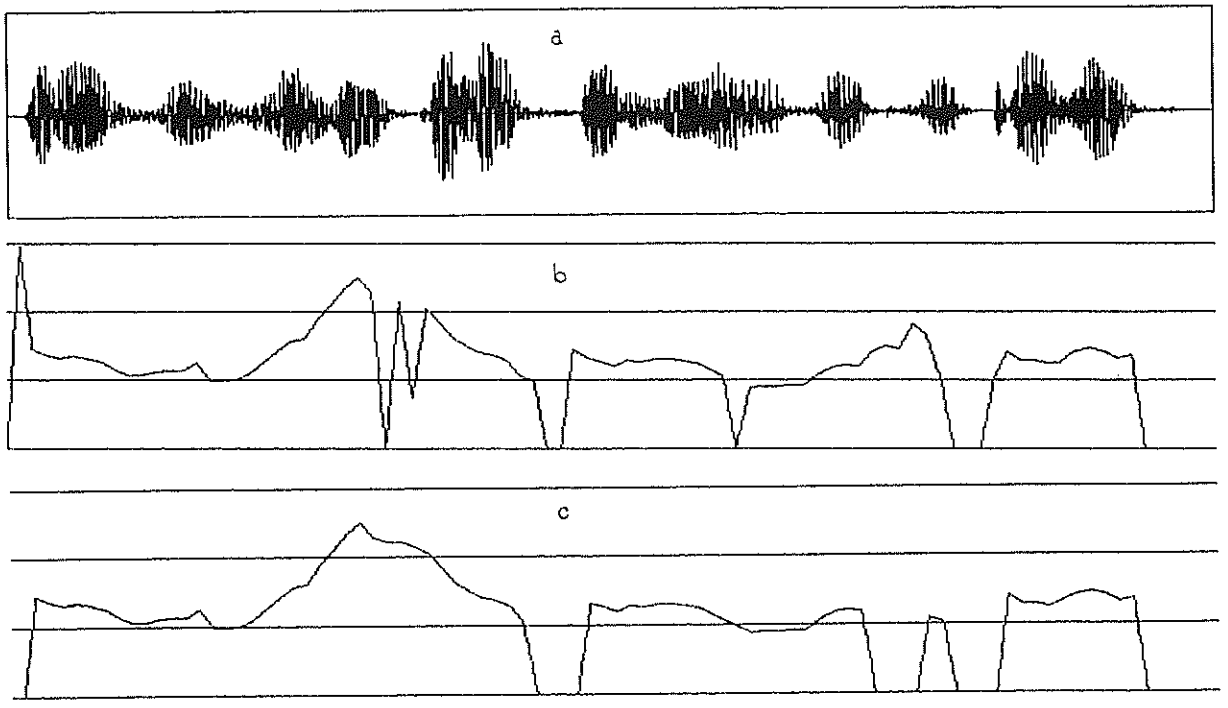


Fig. 3. - Typical Results (a: speech signal, b: pitch contour before smoothing, c: after corrections)

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

PITCH DETECTOR IN SPEECH SIGNALS CORRUPTED BY NOISE

A. Moreno-Bilbao, J. Aracil-Gallardo

Dept. of Signal Theory and Communications. U.P.C.
Barcelona, Spain

Abstract

This paper describes a voiced-unvoiced decisor and a pitch detector robust against noise . A basic pitch extractor, AMDF or Autocorrelation methods have been used and a tracking system is added. Voiced-unvoiced decision is taken based on two enery thresholds. An indirect SNR is measured and applied to change the thresholds. It has been tested under different real noise conditions as car, track, plane, etc. and SNR ratios.

I. INTRODUCTION

Autocorrelation method AC for pitch extraction has been used succesfully in many applications. Preprocessing of the speech signal as low pass filtering, clipping or inverse filtering, [1,2] improves the features of this basic extractor.
AC method can be used for voiced-unvoiced decision. Usually this is done by comparing the first normalized maximum of the function with a fixed threshold.

Pitch detection based on AMDF algorithm [3] is another well known system. Under quiet conditions, AMDF needs a tracking algorithm to avoid jumps to armonics or subarmonics of the pitch frequency. AMDF provides also an internal measure of periodicity by means the quotien between the maximum value of the function and its minimum value (out the origin).

However, in noise environments, two problem arise: Periodic noise produced by some noisy sources can produce peaks (or valleys) in both functions and detect this periodicity in spite of the pitch. Voiced-unvoiced decision is also afected by this kind of noise being biased to voiced.
Non periodic noise afects these methods in the sense that can mask, depending on the SNR, their peaks (or valleys). A fixed threshold for voiced-unvoiced decision is not a good choice because the discrimination also depends on the SNR.

The energy of the signal can help to discriminate between voiced and unvoiced. Usually the system computes the energy in a silence frame and a threshold is calculated from this value. The decision is made by comparing the energy of the speech in each frame with this threshold.

In this paper we present a method to improve the pitch detection and the voiced-unvoiced decision. A pitch tracking is introduced to avoid to detect the periodicity of the noisy source.
An internal SNR measure is computed every frame and applied to adapt the thresholds.

This work was partially supported by PRONTIC grand number 108/88

II. BASIC EXTRACTORS

The autocorrelation of a windowed speech signal is defined as

$$AC(m) = \sum_{k=0}^{M-m} S(k)\ S(k+m)$$

The position P>0 of the first maximum determines the pitch period.

Voiced-unvoiced decision can be made from the normalized autocorrelation

$$R(P) = \frac{AC(P)}{AC(0)} \begin{cases} >THS & \text{voiced} \\ <THS & \text{unvoiced} \end{cases}$$

The AMDF of a windowed speech signal is defined as :

$$AMDF(m) = \frac{1}{L} \sum_{k=1}^{L} |S(k) - S(k-m)|$$

The position P>0 of the first minimum determines the pitch period.
Voiced-unvoiced decision can be made from the normalized function:

$$D(P) = \frac{AMDFmax}{AMDF(P)} \begin{cases} >THS & \text{voiced} \\ <THS & \text{unvoiced} \end{cases}$$

III. VOICED UNVOICED DECISION

Under noisy environments it is necessary to adapt the thresholds for R(P) (or D(P)) [11], and it is convenient to add some aditional information as the energy of the signal.
If the energy in silence frames can be computed, it is possible to implement a threshold for unvoiced-voiced transitions and another for voiced-unvoiced ones. If the SNR of the input signal changes, they have to be adapted. However, even in this case, when SNR is low, this system fails.

For these reasons, we introduce two thresholds but computed from two differents signals: Energy in voiced frames EV and energy in unvoiced frames EUV. If E is the energy of the frame under study:

$EV=0.8EV+0.2E$

if the frame is voiced, and

$EUV=EUV^{0.9}\ E^{0.1}$

if the frame is unvoiced.
EV and EUV give us a smoothed value of the energy in both states. These two values follow the characteristics of the input signal but the variations are slow enough to not be disturbed drastically by an error in the state detection.
Fig1 and Fig2 show the evolution of EV and EUV compared with the energy of the input signal.

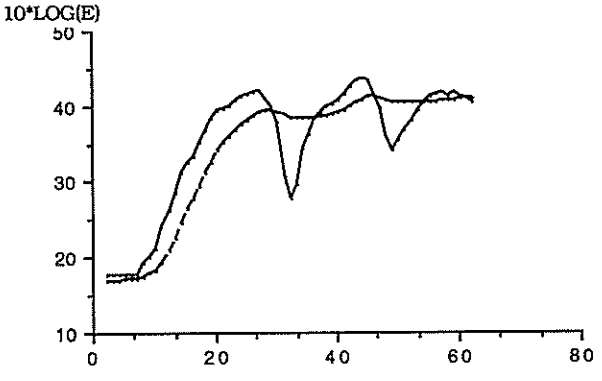


Fig 1. Energy of the signal and smoothed energy in voiced frames.

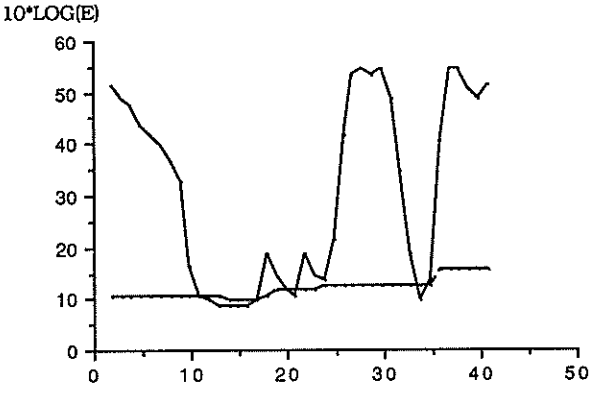


Fig 2. Energy of the signal and smoothed energy in unvoiced frames.

An internal measure of SNR can be obtained from

$$SNR'=10 \log \frac{EV}{EUV} \qquad (1)$$

It is not the SNR of the input signal but can be used as a related function.

From EV and EUV, three thresholds are calculated:

- Voiced threshold:

THV = EV/2

If there are more than four frames unvoiced consecutives THV=EV/4.

- Unvoiced threshold:

THU = EUV f(SNR')

Where f(SNR') is a function calculated empirically from different speakers and noisy environments.

- Sonority threshold:

THS = g(SNR')

THS is a threshold calculated from the signal to noise ratio computed in (1) and it is used to compare with R(P) (or D(P)). THS depends on the kind of basic pitch extractor used in the system and has been empirically founded.

Transitions are determined by:

unvoiced-voiced { R(P)(D(P))>THS and E>THV

voiced-unvoiced { R(P)(D(P))<THS or E<THUV

IV. TRACKING

The tracking algoritm consist in search a pitch in the neighbouring (10 samples) of the previous pitch found.
That means that the algorithm have to validate a pitch and, from it, start the tracking. The validation is done when three contigous frames have E>THV and R(P) (or D(P)) >THS. In each frame, P is the most probable pitch candidate but will be used for the tracking when the third differs less than 10 samples from any of the two previous candidates. In this third frame, usually we get the nucleus of the syllable and for this reason the system is more robust against noisy periodicities.
After a pitch is validated, the system starts the tracking. It calculates the maximum of the autocorrelation R(P) (or minimum D(P)) and the maximum in the tracking window R(T) (or minimum D(T)).

If $R(T)\neq R(P)$, T will be chosen as pitch if R(T)>THS and $R(P)/R(T)<k$, where k has been found empirically.
A similar procedure is applied to AMDF.
Otherwise, a change of pitch occurs and it has to be validated again.

V. TEST SIGNALS

Speech signals are sampled at 8 KHz and windowed in frames of 30 msec delayed 22,5 msec. for applications in a vocoder-LPC. Preprocessing of the speech has been a low pass filter at 900 Hz.

Some noisy environments have been recorded:

1) Diesel motor, 2) plane take-off inside the cabine, 3) plane take-off in the airport and 4) plane in the air inside the cabine accelerating. This noise is extremly difficult because the aceleration produces a wave whose frequency increases in the pitch range of male speakers.

The average spectrum of this noisy signals are shown in Fig. 3 to 6. Motor diesel has a spectrum that don't show strong periodicities but power concentration a frequency close to pitch. Fig. 4 shows mainly an spectrum with two tones near 1300 Hz that are eliminated by the low pass filter. Fig. 5 shows a strong periodicity at pitch frequencies and Fig. 6 shows a spectrum with periodicities in low frequencies. In the results tables we will refer to 1) and 2) non periodic noise and 3) and 4) as periodic noise.

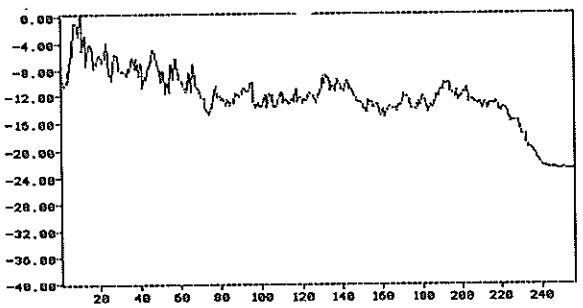


Fig. 3. Spectrum of noise "diesel motor"

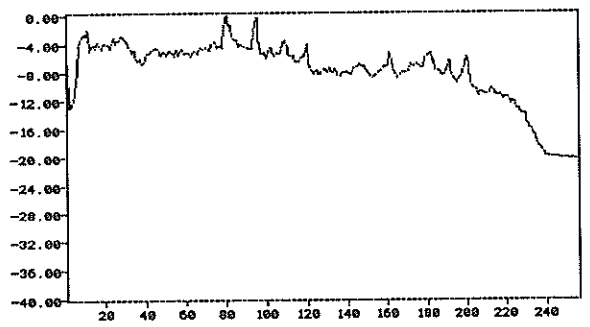


Fig. 4. Spectrum of noise "plane take-off inside the cabine".

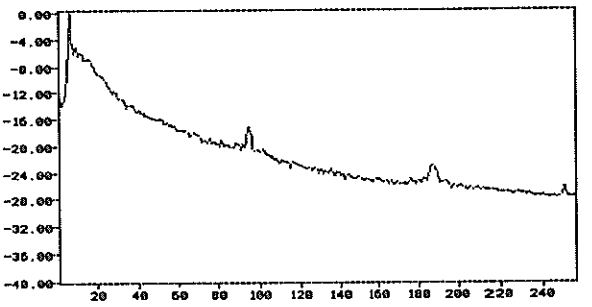


Fig. 5. Spectrum of noise "plane take-off in the airport"

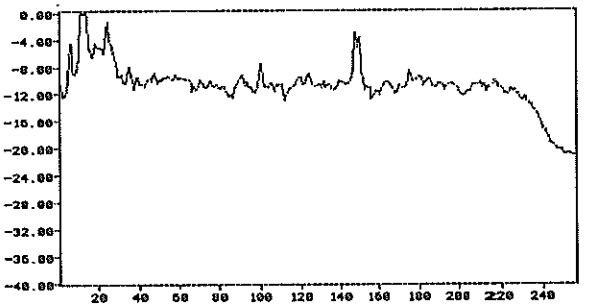


Fig. 6. Spectrum of noise "plane in the air inside the cabine accelerating".

Speech signals from 4 females and 6 males where recorded in a quiet room. 10 sentences of 2 sec each where used to train the system. In average sentences have a 35% of silences and unvoiced frames.

Each sentence was segmented by hand in voiced and unvoiced and AMDF pitch detector was applied to obtain a pitch curve taken without noise.

Noise was added to each sentence to produce the desired SNR computed as:

$$SNR=10 \log \frac{\text{Energy in the all sentence}}{\text{Energy of the noisy signal}}$$

(in the same number of samples).

VI. RESULTS

There have been measured the following errors:

U to V error: the frame was unvoiced but has been detected as voiced. It is computed as the number of errors divided by the total amount of unvoiced frames. Some transitional frames were discarded.

V to U error: the frame was voiced but has been detected as unvoiced. It is computed as the number of errors divided by the number of voiced frames.

Due that we don't have a supervised curve of pitch, we can't calculate fine pitch errors.

Gross pitch error: gross pitch errors (more than 10 samples) are calculated by comparing with the pitch obtained with the AMDF system without noise. They serve as orientation of the system features only.

Tables I and II shows the results for V to U and U to V errors:

SNR		AC	ACT	AMDFT
10		24	0	0
10	P	45	0	0
5		21	14	0
5	P	43	19	3
3		17	21	2
3	P	43	23	7

TABLE I: Error unvoiced to voiced for a) AC autocorrelation, b) autocorrelation with tracking and adaptive thresholds and c) AMDF with tracking and adaptive thresholds.

SNR		AC	ACT	AMDFT
10		3	2,7	20
10	P	3,6	8,4	20
5		8,4	6,8	14
5	P	6,3	6	17
3		10	6,3	16
3	P	6,3	4,7	14

TABLE II: Error unvoiced to voiced in the same cases as Table I.

We obtain the following results:

Unvoiced to voiced: comparing AC and ACT the proposed system decreases the number of errors U to V. Moreover, ACT makes insensible the decision to the kind of noise. In AC system it was very dependent of the periodicity of the noise. AMDFT has a very low error rate and it is slighly dependent of the periodicity.

Voiced to unvoiced errors are maintained very low and almost constant for all the speakers and noise environments with ACT method. The number of errors in AMDF is higher and very dependent of the kind of noise. This system is biased to this kind of errors than can very bad in a vocoder application.

Gross pitch errors: AMDF works correctly with the tracking with non periodic noise. Periodic noise increases gross pitch errors at SNR below 5dB.

Tracking don't improve the autocorrelation method. It eliminates spurious peaks but an error in the tracking can be maintained several frames.

CONCLUSIONS:

It has been presented a system for voiced-unvoiced decision and a tracking algorithm for the pitch. It has been tested under noise environment with a strong energy in the pitch frequencies and periodicities in the range of the pitch. No delay or smoothing has been added to the system to do the decisions.

Voiced-unvoiced decision has been improved. Tracking of the pitch in the autocorrelation method can give problems when a periodic noisy source contaminates the speech signals.

REFERENCES

[1] L.R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection", IEEE Trans. on Acoust., Speech and Signal Processing, vol-ASSP-24, 1976.

[2] M.M. Sondhi, "New Methods of Pitch Extraction", Trans. on Audio-Electroac. Vol AU-16,1968.

[3] M.J.Ross, et al "Average Magnitude Difference Function Pitch Extractor" IEEE Trans. on Acoust., Speech and Signal Processing, vol-ASSP-22, 1976.

[4] H. Kobatake. "Optimization of Voiced/Un-voiced Decisions in Nonstationary Noise Environments" IEEE Trans. on Acoust., Speech and Signal Processing, vol-ASSP-35, January 1987.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A TOOL FOR THE FOCUSING SPEECH SIGNAL ANALYSIS

Goran S. Jovanović

Institute of Applied Mathematics and Electronics
Kneza Miloša 37, 11000 Belgrade, Yugoslavia

The paper investigates a solution for overriding one of the most important problems in the field of the automatic speech recognition: the inter-and-intra subject differences in achieving the acoustic targets in the speech coding process. A general mathematical formulation of the vector local resemblance function for automatic identification of the most important events in the speech signal is presented. The main contribution of the work is an original definition of the function component that is based on the resemblance measurement of the non-neighboring pitch period speech segments.

1. INTRODUCTION

The automatic speech recognition (ASR) systems are usually considered (see for instance Geoff, 1986 [1]) as the pre-processing blocks of the speech understanding systems, although they have significant applications in some other fields, such as the low bit-rate speech transmision, speaker verification and identification, speech to text conversion, language translation, etc. Observing the speech recognition problem in the light of (specific) pattern recognition theory (for instance Fu, 1976 [2]), all ASR systems can be, generally, grouped into three main classes: a) the systems predominantly based on the decision-theoretic or statistical approach (with small amount of the incorporated speech knowledge), b)the systems based on the speech signal structural analysis approach (predominantly relying on the speech knowledge incorporation), and c)the systems representing different combinations of the previous two approaches. In all these approaches the recognition accuracy performance depends (although not at the same extent) on the localization of some significant linguistic events in the speech signal especially for the systems oriented towards the identification of the particular speech elements in the (input) speech signal. The ASR systems from this group are based on the assumption that the speech appears as the result of a specific acoustic coding process that incorporates achieving of the corresponding acoustic targets. Generally, two classes of the acoustic targets can be distinguished: the stationary and dynamic targets. As the other ASR systems, the systems of this group have difficulties coming from the inter-and-intra subject differences in the acoustic realization of the speech elements. Herefrom comes the need for defining the speech signal analizing methods that would reduce, as much as possible, the dissipation of the feature measurement results used for their recognition. From this need came the basic idea for the investigation: the feature measurement should be performed after focusing on the significant events in the speech signal. These events are the parts of the signal having the property of "the best" (quasi) stationarity, as well as the parts corresponding to the locally most abrupt vocal tract changes. So, the basic problem appears to be locating of these significant speech signal events, especially on its voiced portions. Through the exploration having been conducted, a function of the local similarity (FLS) was defined, which gave a good indication of the stationary and dynamic acoustic targets in the speech signal.

2. GENERAL DEFINITION OF THE LOCAL SIMILARITY FUNCTION (FLS)

The logical basis of the definition of the local similarity function is that the systematic vocal tract configuration changes can be followed more appropriately by the non-neighboring segments comparison, because in that way the disturbing influence of the random fluctuations in acheiving the acoustic targets is much more suppressed than in the case of the neighboring segments comparison.

Bearing this in mind, the problem of the focusing analysis can be generally realized starting from the vector function of the local similarity presented in the form

$$FLS = [FLS_1, FLS_2, \ldots, FLS_N] \qquad (1)$$

where each individual component FLS_i, $i=\overline{1,N}$, is defined as

$$FLS_i = FLS_i(j_i,k_i) = d_i\,[f_i(S_{j_i}),\ f_i(S_{j_i+k_i})] \qquad (2)$$

The designations in (2) have the following meaning:

$S_m = s(t)\,|\,t\in T_m$,
t-time, m - index of the analysis frame T_m,
s(t)-the input speech signal

f_i - the corresponding transformation of S_m,
d_i - a suitably chosen distance measure
k_i - skip factor (k_i-1 is the number of the frames between the compared frames)

The rest of the paper concerns the FLS component that enables the focusing speech signal analysis starting from the voiced speech portions.

3.DEFINITION OF THE FLS-COMPONENT FOR VOICED SPEECH SIGNAL SEGMENT ANALYSIS

Let the component FLS_1 in (2) be the local similarity function of the voiced speech portion. In that case appears the problem of negative glottal excitation influence elimination, which generally is not simple. Under the assumption that the locations of individual pitch periods in the speech signal are known, relation (2) can take the form

$$FLS_1 = FLS_1(j_1,k_1) = d_1\,[f_1(S_{j_1}),\ f_1(S_{j_1+k_1})] = d_1\,[f_1(P_{j_1}),\ f_1(P_{j_1+k_1})] \qquad (3)$$

where P_m designates m-th pitch period segment on the voiced speech portion.

The pitch period (segment) identification problem has been efficiently resolved using the FPR method (Jovanović, 1986 [3]), which has great numerical efficiency (nearly two orders of magnitude faster then comparable methods, Miller's DARD method, 1979 [4], for instance). Besides, the FPR method is very reliable and accurate as it represents a specific expert system for the speech signal fundamental peaks identification.

In order to eliminate the disturbances coming from the unvoiced sources of vocal tract excitation, but aiming to preserve the spectral regions of the first two formants, f_1 has been chosen to be low-pass filtration with cutoff frequency of f_g=2 kHz. So, (3) can be represented in the form

$$FLS_1 = d_1\,[\bar{P}_{j_1},\ \bar{P}_{j_1+k_1}] \qquad (4)$$

where $\bar{P}_m$ designates m-th pitch period on the voiced segment in $f_1(\cdot)$.

The pitch period waveforms in (4) can be represented as

$$\bar{P}_{j_1} = f_1(i),\ b_{j_1} \le i < e_{j_1}$$
$$\bar{P}_{j_1+k_1} = f_1(i),\ b_{j_1+k_1} \le i < e_{j_1+k_1} \qquad (5)$$

where b_m and e_m denote the beginning and end of the pitch period. The positions $(i_{o,m})$ of the corresponding maxima $(A_{o,m})$ of the fundamental peaks then satisfy the conditions

$$i_{o,j_1} \subset [b_{j_1},\ e_{j_1}]$$
$$i_{o,j_1+k_1} \subset [b_{j_1+k_1},\ e_{j_1+k_1}]. \qquad (6)$$

For the reason of numerical efficiency preservation d_1 is defined to be minimum Hamming distance between the pitch period segments being compared. After A_{o,j_1+k_1} had been centered to A_{o,j_1}, by the backward time shift of

$$n = i_{o,j_1+k_1} - i_{o,j_1},$$

sample intervals the summation bounderies were determined to insure the complete starting mutual overlaping of the waveforms corresponding to the pitch segments.

4.EXPERIMENTAL RESULTS

The experimental development and verification has been conducted using a speech database B_2 (isolated words-numerals) obtained from 120 speakers in Serbo-Croat language, with one utterance per word. All utterances were digitized at a 10 kHz sampling rate and 12-bit quantization. The database was split into two statistically balanced sets: a training set (RB_2) and a testing set (EB_2).
In figure 1 is illustrated a typical look of the local similarity functions obtained with the skip factor k_1=4. The illustrations correspond to the words of Serbo-Croatian

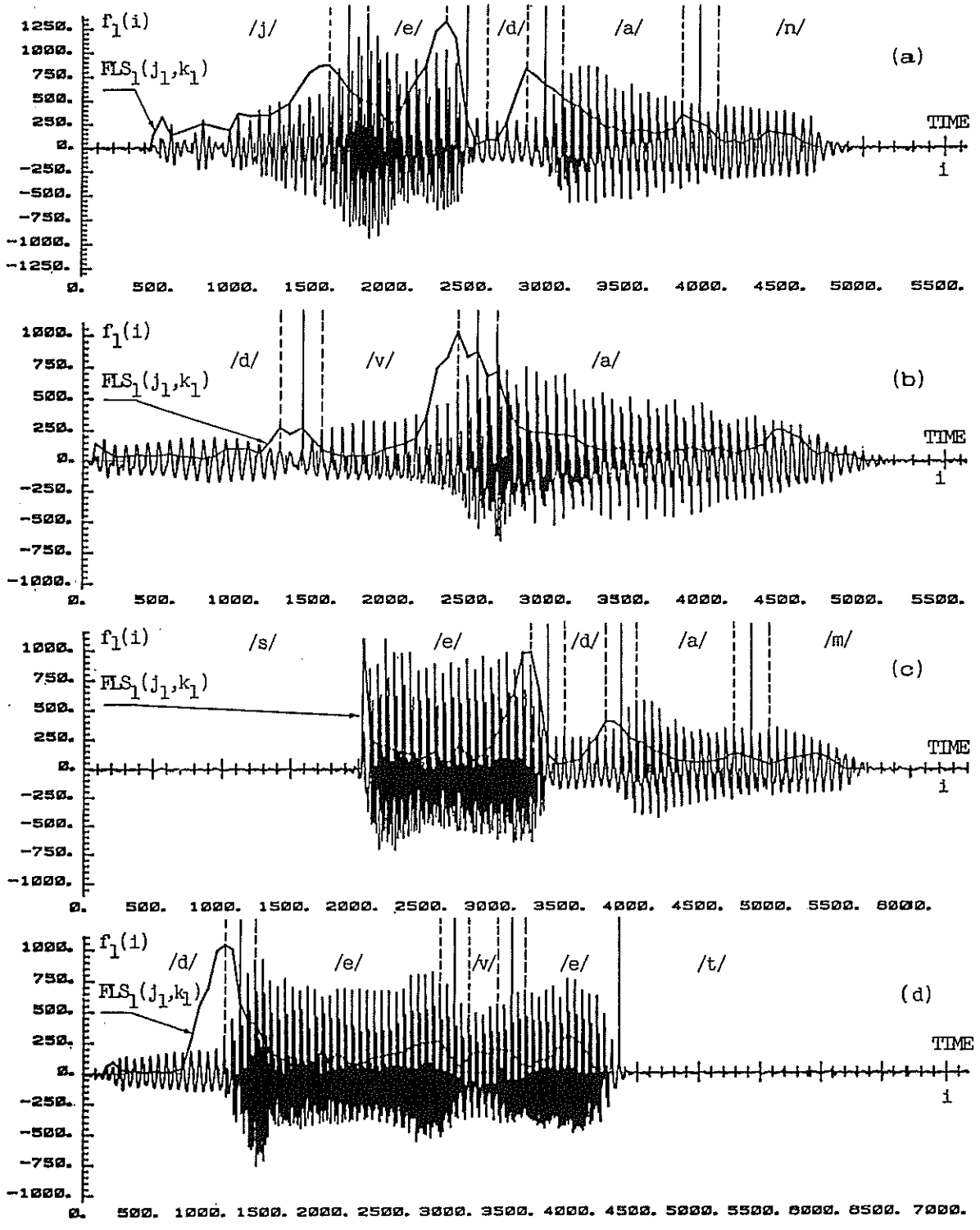


FIG. 1: Illustration of the local similarity function (FLS_1) obtained by the signal $f_1(i)$, with the skip factor k_1=4. The corresponding English words are (a) ONE, (b) TWO, (c) SEVEN, and (d) NINE. The phoneme labeling has been performed manualy using the international phonemic transcription.

language incorporating the abrupt as well as smooth changes of the vocal tract configuration. It can be seen that local maxima of the function indicate very good the inter-phoneme transitions. On the other hand, local minima of the function well indicate regions of the best stationarity. In the illustrations by doted lines are designated the fundamental peaks of the pitch segments being compared. For each transition the left one doted line corresponds to the local maxima of the local similarity function FLS_1.

The performance evaluation of the local similarity function to indicate the inter-phoneme transitions has been made using an expert based system for automatic recognition of different word classes. The system incorporated an algorithm (DPTR algorithm) for inter-phoneme transitions identification, which was developed using the same procedure as in the case of the FPR method, but accomodated for the use with the local similarity function. With a small number of criterions in the classifying algorithm, high recognition accuracy has been achieved with both, the training and control set. Quantitative results concerning the experimental verification are presented in Table-I (for RB_2 set) and Table-II (for KB_2 set). The designations in the tables have the folowing meaning: i-the mnemonic index for the voiced segments of numerals; observe that each of numerals "4" and "8" has two voiced segments when pronounced in Serbo-Croat; N-the number of errors in inter-phoneme FLS peaks detection; Z-the percentage error participation obtained for the given class of voiced segments.

TABLE I - The results for RB_2 data set

i	N_i	Z_i [%]
1	1	3,6
2	1	3,6
3	4	14,3
4_1	1	3,6
5	7	25,0
6	6	21,4
7	0	0,0
8_1	0	0,0
9	1	3,6
0	7	25,0
4_2	0	0,0
8_2	0	0,0

TABLE II - The results for KB_2 data set

i	N_i	Z_i [%]
1	0	0,0
2	1	2,9
3	5	14,7
4_1	0	0,0
5	4	11,8
6	8	23,5
7	1	2,9
8_1	0	0,0
9	3	8,8
0	11	32,4
4_2	1	2,9
8_2	0	0,0

The building up and training procedure of the DPTR algorithm has been performed to achieve the accuracy grater then 95 percent. So, for the training speech sample utterances the total recognition error rate obtained was 3,89 percent. This result as well as the distribution of the errors over the different voiced segment classes were generally confirmed for the test data set (the corresponding error rate was 4,73 percent).

5.CONCLUSION

The investigations having been performed, including experimental verification, indicate the approach to focusing speech signal analysis based on the local similarity function cues of the stationary and dynamic acoustic speech targets is very attractive and worthy of further study. Using the proposed procedure the performance of the speech recognition systems can be significantly improved.

6.REFERENCE

[1] Geoff, B. (ed.), Electronic speech recognition, Collins, London, 1986.

[2] Fu, K.S. (ed.), Digital pattern recognition, Springer-Verlag, Berlin · Heidelberg · New York, 1976.

[3] Jovanović, G.S., "A new algorithm for speech fundamental frequency estimation," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-34, no. 3, pp. 626 - 630, June 1986.

[4] N.J. Miller, "Pitch detection by data reduction," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, pp. 72-79, Feb. 1979.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A Generalized Sample-Selective Linear Prediction Analysis

C. Ma and L.F. Willems
Institute for Perception Research
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract

In this paper, we consider the relation between the covariance linear prediction (CLP) analysis of a frame of a speech signal and the CLP analysis of its subframes . The results of CLP analysis derived from a set of subframes are equivalent to those of a residual-weighted CLP analysis of the complete frame and the solutions of the residual-weighted CLP are the same as those of the generalized weighted average of subframe CLP. A generalized sample- selective CLP analysis is proposed. Those subframes which best reflect the filter model of the speech production can be chosen to improve the accuracy of the estimate of the LPC parameters.

I. Introduction

The process of speech production can be simplified as a source-filter model [1]. The filter can be characterized by an all-pole model represented by the linear prediction equations [2][3]. For voiced sounds, the source is situated in the vibrating vocal folds which modulate the air flow from the lungs. We refer to the vibration frequency of the vocal folds as the fundamental frequency. The unvoiced sound source consists of the turbulent flow formed somewhere in the constricted vocal tract .

In estimating the predictor coefficients, the methods of autocorrelation linear prediction (ALP) and covariance linear prediction (CLP) analysis have become very important. The CLP, in particular, is often used for very short segments of sampled data, for instance in pitch synchronous analysis and closed-glottis-period analysis. When the analysis window is quite wide, for example, covering more than two pitch periods, the performance of CLP is close to that of ALP, but that is not the case for very short segments of sampled data. In order to give a better description of the process of speech production, researchers have paid much attention to the fine structure of formants by means of very short window CLP analysis, or an analysis of only the excitation-free portions, such as the closed glottis portion, to estimate the parameters of the linear prediction model of speech production. But it is not always easy to choose those excitation-free portions, for example, in voiced sounds uttered by females or children, because the pitch period is short. The results are dependent on the data available in the pitch period and are sensitive to the window position [5]. The estimation accuracy of the parameters can be improved by sample selective linear prediction (SSLP) [4], proposed by Yoshiaki Miyoshi et al. In the following sections We shall show that SSLP is a special form of the generalized weighted average CLP analysis.

In practice it is always preferable to obtain an accurate estimate of the LPC parameters so that the source and the filter can be well separated. One example is the glottal inverse filtering technique, which derives the glottal pulse from the speech signal. Improving the estimate of the LPC parameters is one of the main goals in speech processing. The signal of a voiced sound is quasi-periodic. The differences between the successive pitch periods are due to noise or other factors from the glottal source. The pitch synchronous LPC analysis does not benefit from the correlation of the successive pitch periods. However, there always exists some more or less excitation -free portion which best reflects the parameters of the filter model. We cannot use these portions to do pitch synchronous analysis separately, but we can use them in combination to obtain a good estimate of LPC parameters.

In section 2 we shall present the relation between the results of the residual-weighted CLP of a frame and that of its subframe CLP. In section 3 the relation between the frame CLP and the subframe CLP is given. In section 4 a generalized sample-selective CLP method for speech analysis is discussed.

The conclusions are that 1) the results of CLP analysis derived from a set of subframes are equivalent to those of the residual-weighted CLP analysis of the whole frame and 2) the solutions of the residual-weighted CLP are the same as those of the generalized weighted average CLP of the individual subframes. From this we obtain a generalized sample-selective CLP analysis to improve the estimate of LPC parameters.

II. The Residual-weighted CLP of a frame and its subframe CLP

The speech production model can be generally described by the following equations:

$$s_n = \sum_{i=1}^{p} s_{n-i} a_i + e_n \qquad (1)$$

where s_n denotes the nth sample of a speech wave, e_n is the nth sample of an excitation wave, and a_i the ith predictor coefficient. CLP analysis is based on the minimization of the following sum of squared prediction residuals,

$$E = \sum_{n=n1}^{n2} (s_n - \sum_{i=1}^{p} s_{n-i} a_i)^2 \qquad (2)$$

For the sake of simplicity we use a matrix form to represent it. The prediction equations and the error for the CLP are therefore as follows

$$\begin{pmatrix} s_{n1-1} & s_{n1-2} & \cdots & s_{n1-p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n2-1} & s_{n2-2} & \cdots & s_{n2-p} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} s_{n1} \\ s_{n1+1} \\ \vdots \\ s_{n2} \end{pmatrix} \qquad (3)$$

and

$$E = (S\mathbf{a} - \mathbf{s})^T (S\mathbf{a} - \mathbf{s}) \qquad (4)$$

S and s stand for the left-hand matrix and the column matrix of the s_n, respectively, and T represents matrix transpose; $\mathbf{a}^T = (a_1, a_2, \ldots, a_p)$. The Least Square Solution of (3) is

$$\mathbf{a} = (S^T S)^{-1} S^T \mathbf{s} \qquad (5)$$

The above is the normal CLP analysis for the frame of a signal running from $n1 - p$ to n2.

We now choose a window W with a frame length $n2 - n1 + p + 1$ and some subwindows W_k running from $bk - p$ to ek ($bk \geq n1$ and $ek \leq n2$). This is illustrated in Fig. 1. For each subframe W_k, we obtain a set of prediction equations. Putting all subframe equations together, we obtain what we shall call the residual-weighted CLP equations. In this case the total energy of the residual error can be represented by

$$\hat{E} = \sum_{n=n1}^{n2} A_n (s_n - \sum_{i=1}^{p} s_{n-i} \hat{a}_i)^2 \qquad (6)$$

where A_n is the number of times that the prediction equation $s_n = \sum_{i=1}^{p} s_{n-i} a_i$ appears in the set of CLP equations.

In order to relate the predictor $\hat{\mathbf{a}}$ which minimizes (6) to the solution (5), we construct the augmented matrix $(\hat{S}, \hat{s})$ from the augmented matrix (S,s). This matrix is constructed such that the k-th prediction equation is explicitly represented A_n times. It is easy to prove that the Least Squares solution of (3) will be

$$\hat{\mathbf{a}} = (\hat{S}^T \hat{S})^{-1} \hat{S}^T \hat{\mathbf{s}} \qquad (7)$$

or

$$\hat{\mathbf{a}} = (S^T Q^T Q S)^{-1} S^T Q^T Q \mathbf{s} \qquad (8)$$

where $\hat{\mathbf{a}}^T = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p)$, and $Q^T Q$ is a $(n2 - n1 + 1) \times (n2 - n1 + 1)$ diagonal matrix containing the number of times that the predictor appears, which depends on how we choose the subwindows.

From equation (6)-(8), we can see that the results of CLP analysis derived from a set of subframes are equivalent to those obtained by weighting the residual error function e_n of the whole frame with a window in which the amplitude is the element $Q(n,n)$ of matrix Q.

III. The frame CLP and the subframe CLP

The following equations are derived from the augmented matrix which is partitioned according to row.

$$\begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_M \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{pmatrix} = \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_M \end{pmatrix} \qquad (9)$$

In the above equation

$$S_k = \begin{pmatrix} s_{bk-1} & s_{bk-2} & \cdots & s_{bk-p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{ek-1} & s_{ek-2} & \cdots & s_{ek-p} \end{pmatrix} \qquad (10)$$

$$\mathbf{s}_k = \begin{pmatrix} s_{bk} \\ s_{bk+1} \\ \vdots \\ s_{ek} \end{pmatrix} \qquad (11)$$

Because every submatrix S_k and s represent a subframe CLP analysis in which the speech samples are from $bk - p$ to ek ($bk \geq n1$ and $ek \leq n2$), we can rewrite equation (7) as follows:

$$\begin{pmatrix} S_1^T & S_2^T & \cdots & S_M^T \end{pmatrix} \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_M \end{pmatrix} \hat{\mathbf{a}} = \begin{pmatrix} S_1^T & S_2^T & \cdots & S_M^T \end{pmatrix} \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_M \end{pmatrix} \qquad (12)$$

From the product of two partitioned matrices we have

$$\sum S_k^T S_k \hat{\mathbf{a}} = \sum S_k^T \mathbf{s}_k \qquad (13)$$

Each individual subframe analysis, on the other hand, has a solution $\mathbf{a}_k$ given by

$$S_k^T S_k \hat{\mathbf{a}}_k = S_k^T \mathbf{s}_k \qquad (14)$$

where $\mathbf{a}_k$ are the prediction coefficients obtained from the analysis of the k-th subframe. Of course, for each different subframe CLP analysis we find different $\mathbf{a}_k$ coefficients. Comparing the above two formulas, we have

$$\sum S_k^T S_k \hat{\mathbf{a}} = \sum S_k^T S_k \mathbf{a}_k \qquad (15)$$

That is

$$\hat{\mathbf{a}} = \left(\sum S_k^T S_k\right)^{-1} \left(\sum S_k^T S_k \mathbf{a}_k\right) \qquad (16)$$

It is obvious that the solutions of the residual-weighted CLP are a generalized weighted average of the solutions to the individual subframe CLP. We call this the generalized weighted average CLP. The covariance matrix $\sum S_k^T S_k$ is the weighting factor. The error can be calculated by

$$\hat{E} = \sum \mathbf{s}_k^T \mathbf{s}_k - \sum \mathbf{s}_k^T S_k \hat{\mathbf{a}} \qquad (17)$$

Note that the $\hat{\mathbf{a}}$ coefficients are from either the residual-weighted CLP analysis (8) or the windowed signal (16).

IV. The generalized sample- selective CLP

So far we have arrived at the relation (16) between frame CLP and its subframe CLP. The generalized sample-selective CLP will be discussed in this section.

It is useful to analyze a frame of the speech signal with a group of subframes CLP analysis in which only the $\mathbf{a}_k$ coefficients with a small excitation influence are kept. That amounts to an analysis of the speech signal by a generalized sample-selective linear predition. We can see that SSLP is just a special form of this generalized weighted average CLP. The influences of the excitation are included in the analysis frame in SSLP[4]. The generalized weighted average CLP gives us more freedom to choose several subframes to compensate for the scarcity of data and to reduce the noise influence. We can, for example , choose those subframes which do not include any excitation influence. This subframe scheme was also used by P. D. Welch to estimate power spectra in the nonstationary case [7]. The choise of subframes is related to the model of speech production. We will discuss this in the following part.

It is noted that there is a strong correlation from pitch period to pitch period in voiced sounds (exceptions are the voice onset and offset portions). To take advantage of this correlation we can choose subwindows so that they just cover the region where the excitations are relatively small. The information for determining the formants has to be taken from these subwindows, and the result can be optimally obtained from the generalized average of these subframe CLPs.

From figure 2, we can see how formants change according to subframe position. The electroglottogram indicates the status of the vocal folds; the spectral curves are numbered from 0 (top) to 11 (bottom). When the window contains the main excitation the results are unacceptable, for instance, those indicated by curves 3 and 8. When the window is located in the closed glottis portion, good formant estimates are obtained, as illustrated by curves 6 and 11. The analysis conditions for this experiment were as follows. The speech signal was sampled at 10kHz. The pre-emphasis parameter was -0.9 and the window length was 40 samples. The window was moved forward 20 samples every time. Formant frequencies in the open glottis portions deviate from those in the closed glottis regions. Dividing every pitch period approximately into an open glottis portion and a closed glottis portion, we just analyze the data in the closed glottis portion and calculate the average according to equation (16). Due to the correlation between pitch periods the averaging process can also reduce some noise influence. To take an example, curve 0 in figure 2 shows the results from the average for the windows corresponding to spectral curves 3 and 10. As can be seen, an estimation of the formant parameters which fits the speech production model better is obtained.

From the above discussion, we know that we have more freedom in the choice of subframe data than with the SSLP, in addition, the relationship between long-frame CLP , short-frame CLP and the generalized weighted-average CLP is now established.

The conclusions are 1) the results of CLP analysis derived from a set of subframes are equivalent to those of the residual-weighted CLP analysis of the whole frame and 2) the solutions of the residual-weighted CLP are the same as those of the generalized weighted-average CLP of the individual subframe. From this we obtained a generalized sample-selective CLP analysis to improve the estimate of LPC parameters.

Acknowlegement

The constructive criticism and the helpful comments of W. Verhelst and R. Veldhuis are gratefully acknowledged. The authors thank Berry Eggen for providing the speech material and for helpful discussions.

References

[1] J.L. Flanagan,Speech Analysis,Synthesis and Perception. 2nd ed. New York: Springer-Verlag,1972.

[2] J.D. Markel and A.H. Gray, Linear Prediction. 2nd ed. The Hague, The Netherlands:Mouton, 1970.

[3] J. Makhoul, Linear Prediction: A Tutorial Review. Proc. IEEE, Vol. 63, No.4, April 1975, 561-580.

[4] Yoshiaki Miyoshi, et.al., Analysis of speech signals of short pitch period by a sample-selective linear prediction. *Trans. IEEE ASSP-35, No.9, Sept. 1987* , pp.1233-1239.

[5] J.N. Larnar,Y.A. Alsaka, and D.G. Childers, Variability in closed phase analysis of speech. *Proc. ICASSP, pp.29.2.1-29.2.4,1985*

[6] J.N. Holmes, Requirements for speech synthesis in the frequency range 3-4 kHz. F.A.S.E. Symposium on acoustics and speech, Venice, Vol.1, pp.169-172,1981.

[7] P.D. Welch, The Use of Fast Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. In:Modern Spectrum Analysis Ed. by D.G. Childers, IEEE Press, pp.17-20, 1978.

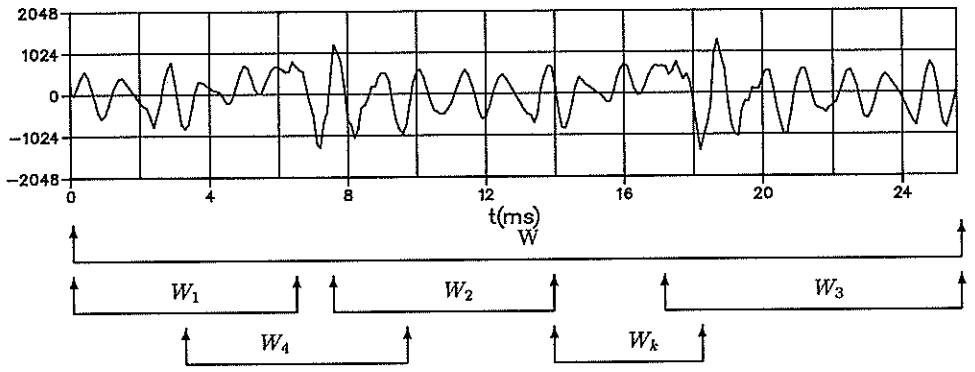


Fig.1. Top: a speech signal. Bottom: an illustration of how the subframes are chosen.

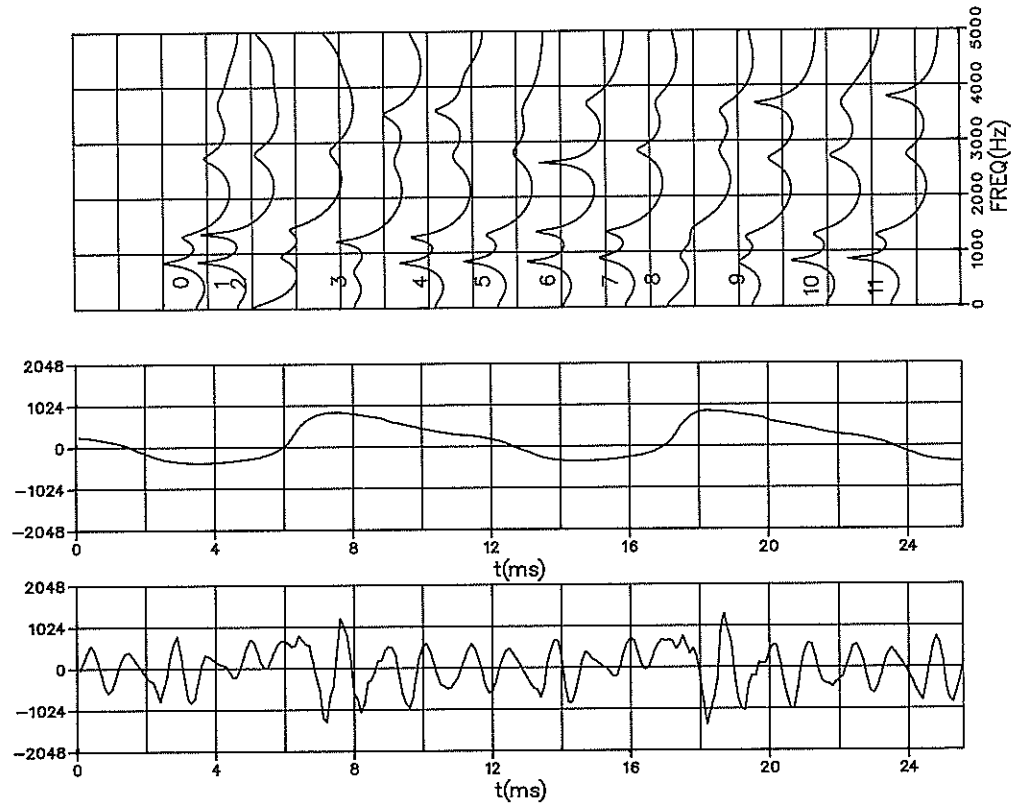


Fig.2. Top: spectrum of the CLP of a 40 samples subframe. Middle: electroglottogram. Bottom: speech signal. The electroglottogram and speech signal are plotted on the same time scale. In the top panel curve 1 corresponds to the result of the CLP of data in the window from 0 ms to 4 ms and curve 2 to the result in the window from 2 ms to 6 ms, and so on. Curve 0 is the generalized average of curves 6 and 11.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

A PC CARD FOR THE REHABILITATION OF DEFICIENT AUDITIVE PEOPLE

MATEOS, J.F.; MACARRON, A.; AGUILERA, S.

Dpto. Ing. Electrónica, E.T.S.I.Telecomunicación (U.P.M.)
Ciudad Universitaria, 28040Madrid - Spain.

ABSTRACT

We introduce a system to help training deaf people in three important features of speech production: Intensity, voicing and pitch. The system is made up of a PC board with an AT&T DSP32 digital signal processor as staple component, and the appropriate software for both the PC and the DSP. Our system computes in real time the values of these three parameters (using frames of 16 msec of speech, digitized at 10 KHZ) and delivers them to the PC each 16 milliseconds. Once the plain results are into the PC, it can store them, post-process them for error correction and graphic display, or perform any other kind of task with them for any application. Our goal is to display them in different ways, either as a pattern the student must imitate with his/her voice, or as a video-game manner, controlled by voice. We want to provide a visual feedback for each of these three parameters of the voice to the speaker.

See pictures 1 and 2. Deaf people (mostly children) can be trained either by playing or by working with their own teacher.

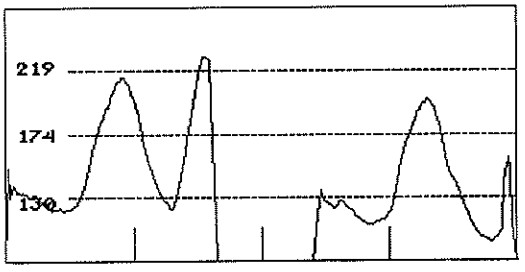


FIGURE1: PC screen using teacher speech as pattern

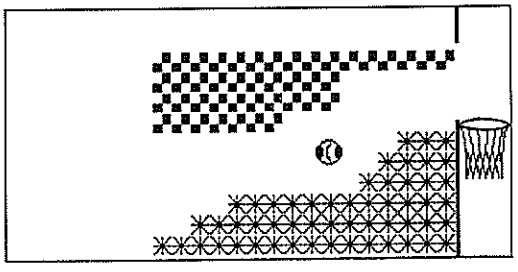


FIGURE2: PC screen controlling video-games by speech

1. PREVIOUS WORKS

The "ancestor" of our ISOTON DIGITAL is other system, named ISOTON (an abbreviation in Spanish for Intensidad, SOnoridad and TONo fundamental) made up by an analog box connected to a first generation microcomputer. The pitch detection was performed by shifting the cutoff frequency of a low pass filter to match the pitch and produce a "clean" sinusoidal waveform. The period of the signal (in voiced intervals) was obtained as the lag between two consecutive peaks of the filtered signal. The silence/speech decision was taken with an energy threshold criterion. The voiced/unvoiced discrimination was performed on the basis of a ratio between the energy of the signal in high and low frequencies. The box delivered the microcomputer the three parameters digitized by an 8 bit A/D converter.

The preceding ISOTON system has been used in deaf people schools for some years, with very good results. It is described in [1], [2] and [3]. The system we introduce here, ISOTON DIGITAL, was designed as a new system, but an upgrade of ISOTON, after considering the following advantages of digital technologies:

-They yield much more reliable hardware and performance than analogies ones, with no adjustment of the components required, what can be an important issue if many units of the systems are to be manufactured.

-Programs can be modified if necessary.

-The same board can be used for other applications on speech processing, only software changes are required.

-PC is a standard. Graphics and development programs can be easily found. Besides, user can utilize it for other tasks.

2. ALGORITHMS

One of our goals was the development of robust and reliable algorithms for the extraction of energy, pitch and voiced/unvoiced character of speech. However we had to implement this in a practical way to achieve real time performance, and an executable code not too long. So, we were limited in computation time and memory. In the following sections, the algorithms used, designed with such constraints, are described.

2.1. ENERGY

We first considered three ways to compute it, using a rectangular window of 16 msec : the plain sum of the squares of the samples, the square root of this value and the log of it. We just compute the first of these, to save time and memory, and then we normalize it with respect to the average noise, computed in silent conditions. The energy value is delivered to the PC, where it is conveniently smoothed for graphic display.

2.2. SILENCE/VOICED/UNVOICED SPEECH DETECTION

Most of the approaches used for this task are based on decisions taken over the value of one or more physical or computational parameters, comparing them with a set of thresholds. We refer to this procedures as "explicit" algorithms. The "implicit" ones try to find a value for the pitch: if no pitch value can be found, the frame is unvoiced. Otherwise, it is considered voiced, but both types of information can be used: First you can try to decide if the frame is voiced or not, using simple threshold criteria, and then use the pitch detection only as a complementary estimator. This is the way our algorithm works, and we like to call it "mixed". In fact, it tries to search for pitch if there is a high probability of the segment being voiced. We considered to use a bunch of estimators of voicing (energy of the signal, pitch, zero crossing rate, first correlation coefficient, comparison between energy in high and low frequencies, etc.), but we realized that many of these estimators were highly correlated, and many of them were a rough measure of the others (e. g., if the signal has many zero crossings in a given interval, it is expected a high content in higher frequency components).

As a result of this consideration, we tried to use estimators with different physical meaning and/or easy computability: energy, zero crossings, first normalized short-term autocorrelation coefficient (R(1)/R(0)) and pitch detection.

One can try to define a threshold for any of these parameters (energy, zero crossings, etc.) so that it divides the possible signals into voiced and unvoiced, but we chose to define two thresholds for them, each of these thresholds set for clearly voiced or unvoiced segments of speech, and an undefinition space between the two thresholds, for segments of speech where a simple estimator is not sufficient to determine the character of voiced or unvoiced. Then, since no single estimator can be relied upon for the voicing decision, we chose to use a combination of criteria, hierarchically arranged, in order to minimize the errors in the decision about voicing decision for segments of speech that are in an undefined region for the thresholds associated with those criteria. This voicing algorithm, with the thresholds chosen after a study of the characteristics of a great deal of speakers, is shown in fig. 3.

2.3. PITCH

We tried some different approaches to this task, most of them well known and thoroughly described in abundant bibliography [4]. Finally, we developed a version of our own of the short-term autocorrelation pitch detector, since we found it reliable and with not very high computational load.

The core of our algorithm, as in any similar method for pitch estimation based on short-term autocorrelation, is the detection of the first significant positive maximum after R(0). The lag corresponding to this peak is then considered as the period of the signal. To remove or attenuate peaks corresponding to high frequency harmonics and noise, the signal is low pass filtered, using a smoothing 9 coefficients FIR filter. These peaks, located near the maximum we are looking for, can carry the algorithm to low precision rates (see Figure 4).

The main features of our algorithm are:

1. Autocorrelation coefficients are only computed for lags corresponding to frequencies between 50 and 600 Hz.

2. Given R(m), a positive peak, and N = number of samples in the window of analysis, m is the pitch

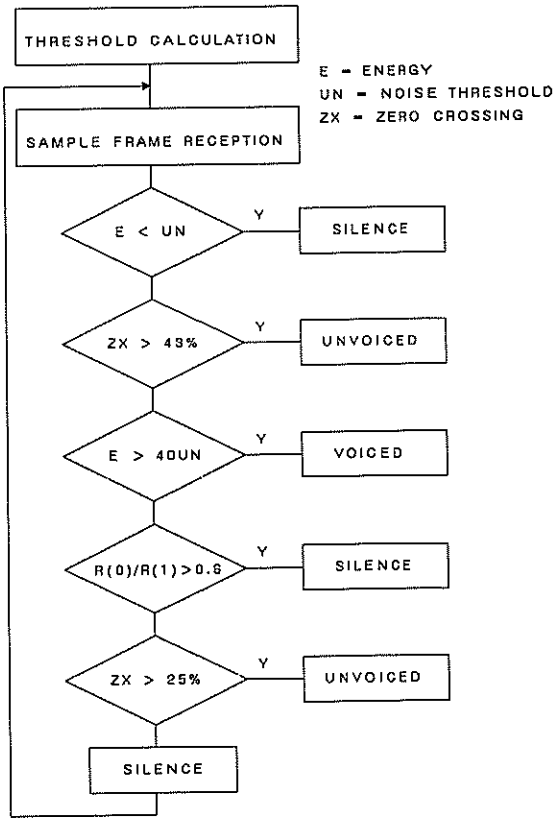


FIG 3: SILENCE / VOICED / UNVOICED DETECTION ALGORITHM

period (in number of samples) if R(m) is at least 70% of R(0), corrected by a factor that takes into account the finite length of the analysis window, N. This factor is just (N-m)/N. Therefore, m has to satisfy that:

$$R(m) > 0.7 \times R(0) \times (N-m)/N \qquad (1)$$

If DSP processor don't found a R(m) value that fulfill (1) condition, in voiced intervals, it puts an alarm flag to be post-process by the PC.

3. After an unvoiced or silence frame, the algorithm is "reset", and looks for it within the range of expected values (50-600 Hz), until a value of m fulfilling (1) is found. The signal is decimated by 2. This reduces the computational load by a factor of 4.

4. If the previous frame was voiced, pitch is expected to be in an interval of +/- 30% of the previous value, for being consistent with this value.

5. If the previous frame was voiced, a variable decimation of the signal is used for computing the autocorrelation. If the pitch was very high, no decimation is applied, since the pitch period is short, and therefore decimation would imply a very low resolution, besides being unnecessary for computational speed. The opposite applies in the case of low frequency, i. e., long period.

6. If the previous frame was voiced, the window length (the default is 32 msec, 320 samples at 10 KHz of sampling frequency) is adjusted to 2 or 3 pitch periods. This is extremely convenient, to reduce the computations involved, but it is very good for capturing the variations of pitch, since the autocorrelation method is a reasonable estimator of an "average" pitch only if this doesn't vary too much within the analysis window. This is also a good "therapy" against having our detector "trapped" to a second or third harmonic.

7. As a minor heuristic "trick" to avoid performing useless computations of R(m), m is increased in more than one unit if the previous value of R(m) was negative or positive but very low. This takes advantage of our search just for big positive peaks, which cannot occur immediately after a very low or negative value of R(m).

In addition to this, some post-processing is performed into the P.C., to correct isolated errors, and to smooth contours in transitions.

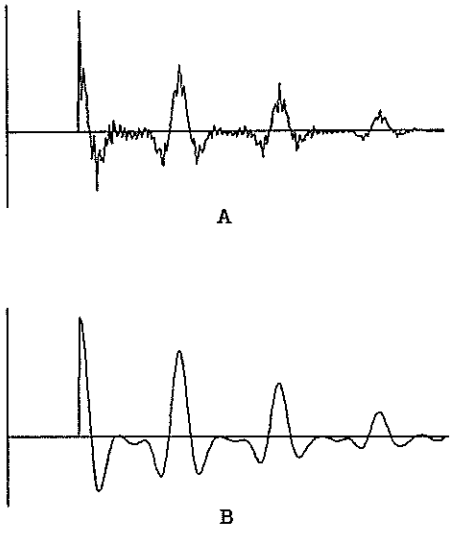


Figure 4: Autocorrelation function of a voiced frame
A) Without smoothing
B) With smoothing

3. HARDWARE

It consists of a single PC board with an external connection to a dynamic microphone. It does not need any external power supply or cables. The board is built with very few simple elements. Thus, the design is very cheap and reliable.

The core of the system is an AT&T-DSP32 digital signal processor. We planned to work with AT&T floating point DSP's, and as we tried to make our algorithms with the lowest cost in computational load and memory, we could implement our system using the simplest member of the family, the DSP32, 40 DIP at 16 MHz. Only the internal RAM (4Kb) is used for both programs and data, and the programs are downloaded by the PC, at the beginning of the operation of our system.

Another important part is the analog to digital converter, used at a sampling frequency of 10 KHz. We have used for this task an 8-bit A-law codec for economical reasons. Since we can tolerate for our application the little distortion due to compression, and the DSP32 performs in a single instruction A-law to float and vice versa conversion, we chose this type of A/D with the only consideration of using a much cheaper A/D than a 12 bit linear one, performing similar precision.

The only other elements in our board are an anti-aliasing and amplifying filter, and the clock circuitry and buffers needed to communicate with the PC.

Communications between codec and DSP and between DSP and PC are made via DMA for maximum throughput.

4. SOFTWARE

The algorithms were developed in 'C' language over a VAX-VMS machine aided by graphical tools. In that stage, we tried to find the optimum version in terms of accuracy and speed, and to achieve the best values of the thresholds and coefficients. Then the whole program was built in the DSP32

assembly language for maximum optimization of the executable code, using the support tools and the emulator over a PC-MSDOS. The final version of the software uses 3712 bytes for data and code, so it fits in the internal 4KB RAM of DSP32, and is fast enough to deliver a group of speech parameters in 0.5 real-time in the worst case, for frames involving the peak computational load.

The program is downloaded into the RAM by the PC when the board is initialized, and it starts collecting 58 ms. data, that is supposed to be silent. Then it computes the statistics of background noise and provides the energy thresholds the rest of the program is going to use. Once it has these values, it restarts all the counters, resets the data tables and goes into an endless loop extracting the three parameters each 16 ms. The tasks performed by this loop are:

1. Reallocation of data in the tables.

2. Computation of energy, zero crossings and R(1)/R(0).

3. Low pass filtering. It is always needed for the next frame, because each analysis window use two frames of 16 msec.

4. First decision about the character of the frame: silence or speech. If silence, it goes to item 8.

5. If the frame is unvoiced, it goes also to item 8

6. Computation of autocorrelations.

7. Extraction of the pitch.

8. Delivery of the values to be represented by the PC.

9. Wait state, controlled internally within the DSP, for the next frame (160 samples).

5. CONCLUSIONS

We have got a system for the education of deaf people, based on a very cheap and robust board which extracts with high accuracy three important parameters from the human voice.

The board performance can be expanded in two ways:

1. Other different computation intensive voice applications (spectrograms, formant tracking, etc.) can be downloaded into the same board instead of the ISOTON DIGITAL program, without changing hardware, since the program and data only use the internal RAM. We can also think of voice synthesis, since the codec includes a digital-to-analog converter.

2. With the same basic design and algorithms, a higher performance version of the DSP32 (25 MHz and/or 100 pin PGA with external memory) or even the DSP32C could replace the DSP used in our board, with highly increased capabilities, using the same software of this application, but with the possibility of adding new features.

Although the concrete implementation is focused on deaf children rehabilitation, the range of application is quite widespread: Phonetics, linguistics, language acquisition, singing learning, and so on.

ISOTON DIGITAL is already been used at educational centers for audition impaired people and deaf children schools, with fairy good results.

ACKNOWLEDGEMENTS

We are very grateful to Mr. Carlos Santamaría and Mr. Jesús Macías, who built the hardware of ISOTON DIGITAL. We are debtful with AT&T Microelectrónica España for giving us the software and hardware tools related with the DSP. We also appreciate the support of the people at DIE-UPM, and our colleagues in the speech group of TELEFONICA I+D, where the two first authors are actually working.

REFERENCES

[1]. AGUILERA,S. "Estudio de parámetros articulatorios del habla, su extracción en tiempo real y aplicaciones en la enseñanza de niños sordos."Ph D. Thesis, Dpto. Ing. Electrónica, U.P.Madrid, 1986.

[2]. AGUILERA,S. BORRAJO,A. et al. "Speech analysis based devices for diagnosis and education of speech and hearing impaired people". Proc. ICASSP86, Tokyo, 1986.

[3]. AGUILERA,S. BORRAJO,A. et al. "Obtención y visualización de algunos parámetros del habla". Mundo Electrónico, N.144, 1984.

[4]. HESS,W. "Pitch determination of speech signals". New York, Springer-Verlag, 1983.

[5]. MACARRON, A. "Software del detector de frecuencia fundamental de la señal de voz, del sistema Isoton Digital". Dpto. Ing. Electrónica, U.P.Madrid, 1989.

[6]. MATEOS,J.F. "Isoton Digital: Filtrado de la señal y extracción de la energía y el caracter sonoro/sordo". Dpto. Ing. Electrónica, U.P.Madrid, 1989.

[7]. AT&T. "DSP32 and 32C, technical information". AT&T.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A Communication Aid for the Hearing Impaired Based on an Automatic Speech Recognizer

D.Kanevsky, C.M.Danis, G.Daggett, P.S.Gopalakrishan, R.Hodgson, D.Jameson, D.Nahamoo
IBM Research Division
T.J. Watson Research Center, P.O.Box 704, Yorktown Heights N.Y., 10598

Abstract

There is evidence that a person's ability to understand two imperfect sources of information is sometimes better that his ability to evaluate each of these sources of information separately. We evaluate this position with respect to using speech recognition technology as an aid to hearing impaired individuals. Our experimental method combined speech presented in noise with a printed version of the same message which contained errors typical of an ASR system. Performance under the dual input condition was compared to performance with each of the two sources of imperfect information (auditory or visual) presented alone. Performance was better under the dual input condition (25% error) than under either the auditory condition alone (43.9%) or under the visual condition alone (44.0%). Some other issues about implementation of the ASR technology in an application for the hearing impaired are discussed.

1. INTRODUCTION

It is known that a person's ability to understand two imperfect sources of information is sometimes better than his ability to evaluate each of these sources of information separately. One example is face-to-face communication in which a hearing impaired subject both sees and hears the speaker. Another example is the use of supplementary tactile information to enhance lipreading performance (Kanevsky & Skurkovich, (1)). In the last two years there has been growing interest in the Speech Recognition Group at IBM in using decoded output from an Automatic Speech Recognizer (ASR) as a second source of information for the hearing impaired. The ability to understand decoded output has been studied under various conditions. In one set of experiments, profoundly hearing impaired subjects could observe the lips of the speaker while simultaneously viewing the decoded output from the ASR. In another set of experiments hearing impaired subjects received printed decoded output from the ASR together with limited acoustic information from hearing the speaker dictating the sentences.

All these experiments led to the strong impression that even though the output from ASR is imperfect, using it significantly increases comprehension of speech for the hearing impaired. We present in this paper the first in a series of experiments which evaluate the usefulness of imperfect ASR output to supplement degraded auditory information. One possible application would be in telephone communication for the hearing impaired. The speech received on the telephone could be decoded by the ASR. Some auditory information from the telephone coupled with the decoded output of the ASR might greatly enhance comprehension.

Since the hearing ability of the hearing impaired varies considerably, it is difficult to estimate the effects of different factors experimentally. We accordingly performed experiments with hearing individuals who viewed a decoded output from an ASR while listening to recorded speech in heavy noise. These experiments are described in Section 3. Discussion of informal experiments is presented in Section 2.

2. INFORMAL EXPERIMENTS

Our experiments were performed using the decoded output from the isolated word speech recognition system developed at IBM research. This system is speaker dependent, with discrete speech input and with 20,000 words in its vocabulary. It can achieve about 95% accuracy for native American speakers (Danis, (3); Brown & Vosburgh, (4)).

Two different methods of using this system were first investigated.

2.1. Fast Match decoding

The current ASR at IBM has two decoding modules - Fast Match and Detailed Match. The Fast Match module produces a list of candidate words for each utterance (Bahl et al.(2)) and the Detailed Match chooses the best words using the candidates proposed by the Fast Match. The output of the Detailed Match module is displayed with about 3 words delay. We performed experiments in which 90 sentences were spoken and decoded using only a Fast Match technique: for each spoken utterance the only word displayed was the one with the highest score on the basis of Fast Match computations. Average decoding accuracy in such experiments was about 65% but decoding was significantly faster. This decoded output for 90 sentences was then used in an experiment in which subjects observed a TV terminal on which sentences were displayed, word by word under one of the following three conditions:

a. Only decoded output from the ASR was displayed. A sentence was displayed on the terminal word by word until the whole sentence was generated.

b. Nothing was displayed on the terminal except the phrase "Hear only". If the subject was a hearing individual he could hear spoken sentences embedded in heavy noise. If the subject was hearing impaired then the sentences were played back without noise.

c. Printed words appeared on a TV screen and were simultaneously and in synchrony played back by audio recorder. Again, for normal hearers acoustic information was presented in heavy noise and for the hearing impaired the noise was omitted.

These 3 conditions were equally and randomly distributed among 90 sentences. Subjects were asked to indicate what they had understood after each sentence was displayed (or played back) - either repeating some words, or paraphrasing what they had understood. The responses were recorded and later analyzed. Responses were compared with reference sentences and qualitative estimates of how good they were on five point scale were produced. Five normal hearing and one hearing impaired person were tested.

In these experiments each tested person reported that he felt that he understood much better in condition (c), in which two sources of information - auditory and decoded output - were used. But it was difficult to determine quantitatively under which conditions comprehension was better. Analysis of problems in this informal experiment resulted in the design of subsequent experiments in which sentences were short, had a fixed length and fixed number of errors was introduced.

2.2 Telephone experiments

In these experiments more than 10 hearing and profoundly hearing impaired subjects used the ASR to help them speak by telephone with members of the Speech Recognition Group at Yorktown. The speech of normal hearers was decoded by the ASR and sent to hearing impaired subjects on terminals at their location. If the hearing impaired had intelligible speech they responded over telephone by voice; otherwise they typed their responses, which were displayed on terminals of the normal hearers participating in the conversations. These conversations were conducted with hearing impaired subjects located at various remote sites (Chicago, White Plains, etc.). Decoding accuracy for the ASR over the telephone varied between 75% and 98% for different speakers. It became quite clear from these experiments that the hearing impaired are able to achieve very good comprehension when the topic of conversation was more or less clear. Normal hearers reported the impression of an "almost normal phone conversations."

In all these experiments the ASR was speaker dependent. The basic question arose whether decoder output which is less accurate than in the above experiments would still significantly improve understanding for hearing impaired individuals. Lower levels of accuracy would result if the ASR were used by a person for whom the recognizer had not been trained; in effect, if the ASR were to be used as if it were speaker independent. One experiment which investigates this question is described in the next section.

3. FORMAL EXPERIMENTS

3.1 Subjects

Six members of the technical staff at IBM's T. J. Watson Research Center participated in three 20 minute sessions each.

3.2 Materials

Sixty sentences of the following form were generated: **The <color> <object/shape> <preposition> the <color> <object/shape>.** Individual sentences were produced by randomly selecting, with replacement, an instance of each category from among eight

alternatives. An example of a stimulus sentence produced in this way is: **The grey square beside the red table.**

The same set of 60 sentences was used in each of the three experimental conditions. However, the sentences were transformed in a different way for each condition. In the Hearing Only (HO) condition the stimulus sentences were recorded digitally (at a bandwidth of 8KHz, with a high-pass filter set at 300Hz and a sampling rate of 20KHz) and presented to the subject concurrently with a recording of white noise. Levels for the two sound sources were determined informally so as to produce approximately 60% correct recognition for the content words. The same sound levels were used for all subjects.

For the Decoder Only (DO) condition the stimulus sentences were presented with 40% of the content words replaced by words which correspond to typical recognition errors produced by IBM's experimental, large-vocabulary, discrete word input speech recognition system. A pool of words for substitution was obtained by processing the sentences dictated for presentation to subjects in the HO condition through the recognizer and noting recognition errors. Additional errors were obtained by having another speaker dictate the same sentences to the recognizer. Errors were introduced into the stimulus sentences by selecting randomly from among four possible errors for each of the content words. Two words were substituted in each sentence to produce an error rate of 40%. An example of a DO sentence is: **The grey rare beside the could table.** Less than half of the substituted errors bore a phonological similarity (e.g., "rare" for "square") to the original word. The remaining words (e.g., "could" for "red") were unrelated to the replaced word.

In the Hearing and Decoder (H&D) condition both the speech combined with noise and the altered sentences in visual form were presented.

3.3 Procedure

All six people completed all three conditions. Presentation order was counterbalanced across participants by assigning a different one of the six possible orders to each subject (conditions H&D and DO were inadvertently switched for one user, such that H&D was presented before DO). Sessions were typically separated by several hours, although the lags ranged from zero (one case) to 24 hours (3 cases).

At the beginning of the experiment subjects were shown the sentence frames and the vocabulary list which was used to generate the original list of stimulus sentences. Their task was to reproduce the original stimulus sentence based on the knowledge of the frame and the vocabulary, as well as the information contained in the transformed sentences.

In the HO condition, subjects listened to the sentence plus noise and wrote as many of the five content words as they could discern. Subjects controlled presentation rate of the sentences. In the DO condition subjects were presented with the list of 60 sentences and asked to correct the errors by selecting the most probable word from the vocabulary based on whatever cue the decoder error seemed to provide. All subjects were aware that recognition errors frequently sound like the spoken word. In the H&D condition users were given the same (visual) list of 60 sentences as in the DO condition and also listened to the sentences presented in noise. The pace of the auditory sentences was controlled by subjects. Subjects made corrections on the sheet with the typed decoder output.

Results and Conclusion

The number of errors were counted for each subject under each condition and divided by the maximum number of errors in that condition to determine the percent error rate. The maximum number of errors was 300 in the HO condition (60 sentences x 5 content words) and 120 in the DO and H&D conditions (60 sentences x 2 decoder errors). The data is plotted in Figure 1. For each subject (arrayed on the x-axis) three points are plotted corresponding to the three conditions. The average percent error for the HO and the DO conditions were very similar, 43.9% and 44.0%, respectively. When the two sources of input were combined the average error rate improved to 25.4%. Inspection of Figure 1 reveals that five of six users performed best when the two sources of information were combined.

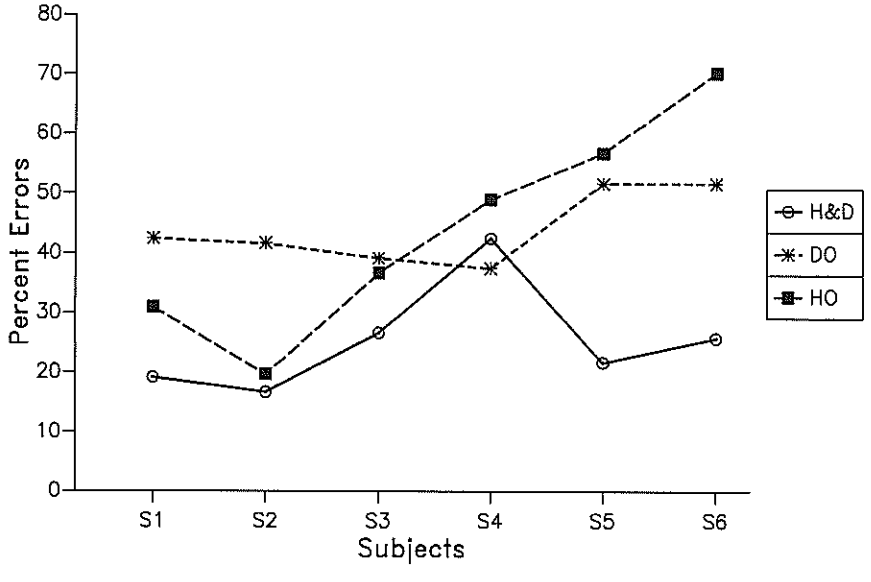
We conclude from this experiment that providing a second source of input with an accuracy of only 60% still provides sufficient information to improve perception of speech presented in noise. This suggests that using a speaker dependent ASR as if it were speaker independent, generating a much lower rate of accuracy than would occur if the system were trained to the user's voice, would still be useful as a telephone communication aid for the hearing impaired. Speaker independent use of an ASR would be more realistic both from logistical and cost perspectives.

These experiments also left some important questions unanswered: What is the optimal time interval between displaying decoded output and producing auditory information (from a real speaker or a recorder)? How well do experiments with normal hearers model the real situation with hearing impaired persons? Later studies will be aimed at answering some of these questions.

References

1. Kanevsky, D. and Skurkovich, G. (1982) Vibrotactile articulator. in *Use of Computers in Aiding Disabled*, J.Raviv. (Ed.). North-Holland.
2. Bahl, L.R., De Gennaro, S.V., Gopalakrishnan, P.S., Mercer, R.L., "A Fast Approximate Acoustic Match for Large Vocabulary Speech Recognition", Proc. EUROSPEECH '89 Conference, Paris, France, Sep. 1989.
3. Danis, C.M., Developing Successful Speakers for an Automatic Speech Recognition System. Proceeding of the HUMAN FACTORS SOCIETY 33d ANNUAL MEETING, pp. 301-304, 1989.
4. Brown, N.R. and Vosburgh, A.M., Evaluating the Accuracy of a Large-Vocabulary Speech Recognition System. Proceeding of the HUMAN FACTORS SOCIETY 33d ANNUAL MEETING, pp. 296-300, 1989.

Figure 1
Percent Errors for Individual Subjects in the three Conditions



SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ROBUST SPEAKER-INDEPENDENT WORD RECOGNITION USING SPECTRAL SMOOTHING AND TEMPORAL DERIVATIVES

Ted H. Applebaum and Brian A. Hanson

Speech Technology Laboratory
Division of Panasonic Technologies, Inc.
3888 State Street, Santa Barbara, CA 93105, U.S.A.

Ambient noise and the changes it induces in a talker's speech production ("Lombard effect") both pose problems for automatic speech recognition. The goal of this work is to find a representation of speech which will support high recognition rates of either normal or noisy-Lombard speech by a recognizer trained on normal speech.

Four features were considered: cepstral coefficients and their first three temporal derivatives. Regression- and difference-based implementation of the features were evaluated. Long regression window lengths were found necessary for high order regression features. Spectral smoothing by reducing the number of cepstral coefficients was examined as a way of maintaining high recognition rate at shorter regression window lengths. Combinations of features which are effective for recognition of the digit vocabulary in either normal or noisy-Lombard input speech conditions were found.

1. INTRODUCTION

The goal of this work is to find a representation of speech which will support high recognition rates in both normal and noisy Lombard speech conditions. The representations considered consist of combinations of temporal derivatives of cepstral coefficients. The *first* temporal derivative of cepstral coefficients has been widely used in speech recognition. The temporal derivative has been implemented as either a first difference (e. g. [1-3]) or first order regression (e. g. [4-10]). The first temporal derivative of cepstral coefficients, in a difference implementation, has been shown to increase recognition rate for clean, noisy, Lombard and noisy-Lombard speech [11].

Furthermore, the *second* temporal derivative has been used in speaker verification [4,5] and in speech recognition [7,11]. Furui [7] used static, first and second order regression features to recognize normal (non-noisy, non-Lombard) speech, but found that the use of second order regression feature gave no significant improvement of recognition rate beyond what was achieved by the combination of static and first order regression features. [11] introduced a difference-based implementation of the second temporal derivative, and found improved recognition of noisy, Lombard and noisy-Lombard speech by a recognizer trained on clean speech. However, as in [7], no gain in recognition rate due to the second temporal derivative feature was found with normal speech input.

The current work continues the research presented in [11], and focuses on the regression implementation of temporal derivative features for regression orders zero through three. The regression implementation is compared to the difference implementation from [11]. Interaction of temporal derivatives with temporal and cepstral smoothing is explored, and an attempt is made to compensate for the long regression window lengths required for high order regression features. Combinations of regression features are found which are effective for recognition of the digit vocabulary in both normal and noisy-Lombard input speech conditions.

2. PRELIMINARIES

2.1. Temporal Derivative Features

The regression implementation of temporal derivative features is the main focus here. Comparisons are also made to the difference implementation studied in [11].

Let be $P_r(X,N)$ be the r-th orthogonal polynomial of length N. The first few orthogonal polynomials are [12]:

$$P_0(X,N) = 1$$

$$P_1(X,N) = X$$

$$P_2(X,N) = X^2 - \frac{1}{12}(N^2-1)$$

$$P_3(X,N) = X^3 - \frac{1}{20}(3N^2-7)X$$

Let $C_k(t)$ be the k-th cepstral coefficient from a frame of speech at time t, and let ΔT be the step size between frames. The regression implementation of the r-th temporal derivative (or "r-th regression feature") is then:

$$R_{rk}(t,N,\Delta T) = \frac{\sum_{X=1}^{N} P_r(X,N)\, C_k\left[t + (X - \frac{N+1}{2})\,\Delta T\right]}{\sum_{X=1}^{N} P_r^2(X,N)}$$

where N (the number of analysis frames in the regression window) is an odd integer.

The notation above will be simplified, where appropriate, by omitting the k and t dependence, assuming a fixed value for ΔT and expressing the window length by the product $(N\,\Delta T)$. For example the first order regression feature with $N = 15$, and $\Delta T = 10$ msec will appear as $R_1(150$ msec). Further, the unit of window length will be assumed to be milliseconds, and the combination of features will be indicated with "+", as in "$R_1(150)+R_2(250)$".

A difference implementation of temporal derivative features was considered in [11], where the zero-th, first and second derivatives of the cepstral coefficients were called the "static", "dynamic", and "acceleration" features. These features are:

$$S_k(t) = R_{0k}(t, 1, \Delta T) = C_k(t) \qquad \text{(Static)}$$

$$D_k(t) = C_k(t+\delta_D) - C_k(t-\delta_D) \qquad \text{(Dynamic)}$$

$$A_k(t) = C_k(t+\Delta_A+\delta_A) - C_k(t+\Delta_A-\delta_A) - C_k(t-\Delta_A+\delta_A) + C_k(t-\Delta_A-\delta_A) \qquad \text{(Acceleration)}$$

where δ_D, δ_A and Δ_A are free parameters with the dimension of time. The *dynamic feature window length* is defined here as $2\,\delta_D + \Delta T$. The *acceleration feature window length* is defined as $2\,(\delta_A + \Delta_A) + \Delta T$.

2.2. Recognition System

The experiments described below used an isolated word recognition system based on twelfth order perceptually-based linear prediction analysis [13], with an analysis step size of ΔT = 10 msec. Cepstral coefficients 1 through 12 were used, except where explicitly noted in section 3.2. Each feature was vector-quantized independently. VQ-codebook design and quantization used the index-weighted cepstral (RPS) distance [14]. A five state, discrete density hidden Markov model was trained for each vocabulary word. Output probabilities due to each feature were combined at the frame level [2]. Decoding was done by the Viterbi algorithm without a grammar. Note that the relative normalizations of the features have no effect on recognition results of this system. More information about this recognition system is available in [11] and [15].

2.3. Databases

The vocabulary consisted of ten English digits and the word "oh". The *training* data consisted of one "normal" (non-noisy, non-Lombard) repetition by 96 talkers. The *testing* data consisted of two "normal" and two Lombard repetitions by 24 talkers (different from the 96 training talkers). While recording the Lombard part of this database the talkers listened through calibrated headphones to 85 dB SPL white Gaussian noise. (The effective signal-to-noise ratio at the time of recording the Lombard speech data was 10 to 15 dB). White Gaussian noise at 18 dB SNR was later added to the Lombard speech data to produce "noisy-Lombard" data. These databases are described in more detail in [11].

3. EXPERIMENTS

In the following experiments the recognizer was trained on normal (non-noisy, non-Lombard) speech. Recognition performance of the individual regression features were measured in normal and noisy-Lombard input speech conditions. The regression implementation of the temporal derivative features was compared to a difference implementation. The interaction of temporal smoothing (due to averaging within the regression window) and spectral smoothing (due to truncating the cepstral expansion of the spectrum) was considered. The regression window lengths were then adjusted to optimize recognition performance of sets of combined features for noisy-Lombard test data. Finally, the performance of the optimized feature set was checked for normal speech test data.

3.1. Recognition based on one feature

The recognition performance of the individual regression features were considered in normal and noisy-Lombard input speech conditions. Figure 1 shows the recognition rate vs. window length ($N\,\Delta T$) for regression features R_0 through R_3. The left graph shows results for normal test data and the right graph shows results for noisy-Lombard test data.

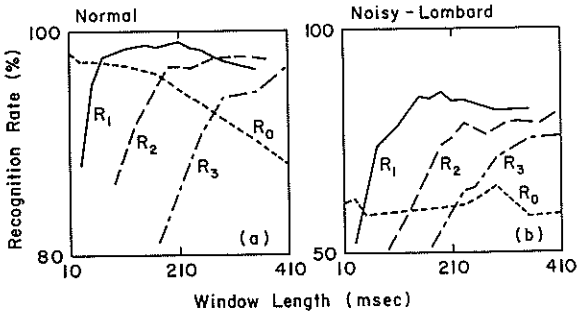


Fig. 1 Single feature recognition rates. Recognition rate vs. window length: **(a)** normal (non-noisy, non-Lombard) test speech and **(b)** noisy-Lombard test speech. All training is done on normal speech.

3.1.1. Normal test data.

For normal test data each feature achieves a recognition rate of over 96% for some window length. The R_0 feature has a unique peak recognition rate at R_0(10msec), which is equivalent to S, the static feature. For higher regression orders the best recognition rates are achieved at successively longer window lengths. The *optimal* regression window lengths get quite large (over 300 and 400 msec for R_2 and R_3 respectively). These window lengths exceed typical syllable lengths [16]. Averaging over such long regression windows will result in loss of temporal resolution, and will particularly reduce recognition rates for more confusable vocabularies, where fine phonetic distinctions must be made. Therefore, where near-optimal recognition rates are achieved for a range of window lengths, the smallest *effective* window length is preferred.

3.1.2. Noisy-Lombard test data.

The best single-feature recognition rate for noisy-Lombard test data is achieved by the R_1 feature, and recognition rate declines for features R_2 and R_3. The R_0 feature never achieves recognition rates comparable to the other features for noisy-Lombard test data. R_0 exhibits two local optima of recognition rate, i.e. at window lengths of 10 to 50 msec and 250 to 350 msec respectively. For noisy-Lombard speech, as for normal speech, the best recognition rates for the higher regression order features are achieved at successively longer window lengths. For regression orders greater than zero the recognition rate increases rapidly with window length for short windows, and exhibits only small changes when the window length exceeds some critical value. This *smallest effective window length* increases with regression order. In the case of our noisy-Lombard test speech the smallest effective window lengths for R_0 through R_3 were approximately 10, 150, 230 and 290 msec respectively.

3.1.3. Regression vs. difference implementations.

Recognition results based on a difference implementation of the temporal derivative features, and using the same recognition system and database as the present study, were previously reported [11]. Figure 2 shows recognition rate vs. window length for both regression and difference implementations of temporal derivative features. The left graph shows the first derivative features R_1 and D ("dynamic" feature). The right graph of Figure 2 shows the second derivative features R_2 and A ("acceleration" feature). The window lengths of the dynamic and acceleration features are as defined in section 2.1. For the acceleration feature, δ_A is fixed at 70 msec. For either first or

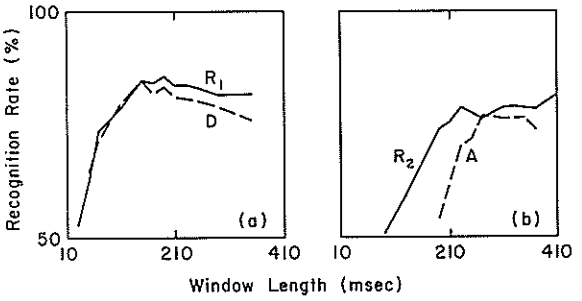


Fig. 2 Comparison of regression and difference implementations of temporal derivative features. Recognition rate for noisy-Lombard test speech vs. window length: (a) regression feature R_1 (solid line) and difference feature D (dashed line), and (b) regression feature R_2 (solid line) and difference feature A (dashed line). The window length for D is $2\,\delta_D + \Delta T$ and the window length for A is $2\,(\delta_A + \Delta_A) + \Delta T$.

second derivative features the regression and difference implementations achieve similar optimal recognition rates, and exhibit nearly the same trends in the dependence of recognition rate on window length. The difference in the smallest effective window lengths for the R_2 and acceleration features may be due to our definition of "acceleration feature window length", and the fact that the δ_A was fixed while Δ_A varied.

3.2. Cepstral smoothing and temporal derivatives

Recognition rate can sometimes be increased by truncating the cepstral expansion of the spectrum. We hypothesized that, with fewer than 12 cepstral coefficents, the effective window length may be reduced. In Figure 3 recognition rate is shown vs. number of cepstral coefficients for two values of the R_2 window length. The longer window (230 msec) is near the smallest effective value for R_2. No gain in recognition rate is seen for cepstral smoothing in this case. The shorter window (130 msec) is less than the smallest effective value for R_2. Recognition rate is seen to increase as the number of cepstral coefficients is reduced from 12 to 5, and drop sharply as the number of cepstral coefficients is further reduced. Although cepstral smoothing improves recognition rate for the shorter regression window, the highest recognition rate obtained for the shorter window length (71.8%) is well below the recognition rate achieved at the near-optimal window length (78.9%).

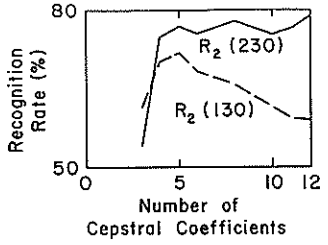


Fig. 3 Cepstral smoothing effects. Recognition rate for noisy-Lombard test speech vs. number of cepstral coefficients: for near-optimal R_2(230 msec) case (solid line) and sub-optimal R_2(130 msec) case (dashed line).

Similar results for R_1 further demonstrated the interaction between temporal smoothing and cepstral smoothing. Again in this case reducing the number of cepstral coefficients from 12 to 5 improved recognition rate for a sub-optimal (i.e. short) 90 msec regression window. However the best recognition rate for the shorter regression window was less than the recognition rate achieved for a near-optimal, 170 msec window with 12 cepstral coefficients.

3.3. Optimization of combined feature recognition rates

The regression window lengths were adjusted to optimize recognition performance of sets of combined features for noisy-Lombard test data. Figure 4 shows recognition rate vs. window length as the window length is varied for one of the regression features, and held constant for each of the others. Effects of varying the window lengths of features R_0 through R_3 are shown in Figs. 4(a) through 4(d) respectively. The best values obtained at each stage of optimization were retained in the next stage.

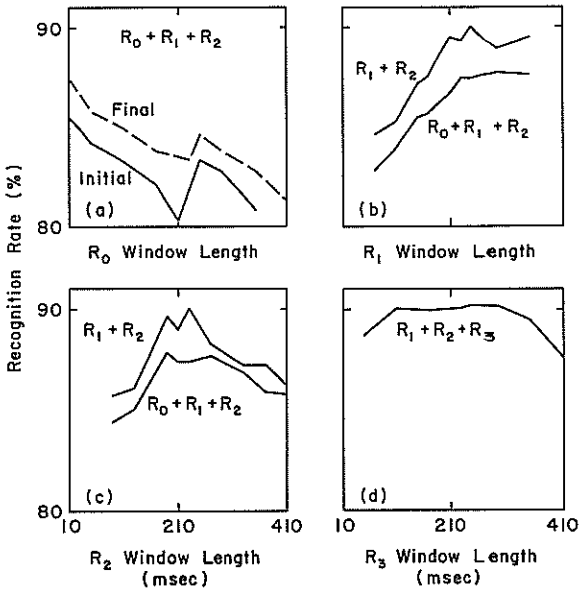


Fig. 4 Combined feature recognition rates. Recognition rate for noisy-Lombard test speech vs. window length: window lengths of features R_0 through R_3 are varied in graphs (a) through (d), respectively. The fixed window length features are: (a) initial R_1(150), final R_1(250) and R_2(230), (b) R_0(10) and R_2(230), (c) R_0(10) and R_1(250), and (d) R_0(10), R_1(250) and R_2(230).

In Fig. 4(a) (*solid line*) the shortest window length for R_0 is found to give the best recognition rate. A second peak of recognition rate is present near 250 msec, but is much weaker than for R_0(10). In Figs. 4(b) and 4(c) the feature set R_1(250)+R_2(230) gives the best recognition rate. Notice that including R_0 with these features causes the error rate to go up by about 20%. In Fig. 4(d) feature R_3 is included with the best combination of features found in the preceding stages. No significant further improvement of recognition rate is found for R_3 at any window length. Finally, returning to Fig. 4(a) (*dashed line*), the original choice of R_0(10) was verified in combination with the best values of the other parameters.

The performances of the optimized feature sets determined above were checked for normal speech test data. The feature sets R_0(10)+R_1(250)+R_2(230) and R_1(250)+R_2(230) achieved recognition rates of 99.6% and 99.8% respectively.

4. DISCUSSION

Although each regression feature produced similarly high recognition rates for normal test speech, the regression features varied greatly in their performance for noisy-Lombard speech. By itself R_0 never achieved recognition rates above 65% for noisy-Lombard speech, while the higher order regression

features each achieved over 76% recognition rate at sufficiently long window lengths. In combination with the other features R_0 *decreased* recognition rates for noisy-Lombard speech.

The combination of R_1(250 msec) and R_2(230 msec) achieved recognition rates as good as, or better than, any of the other feature combinations (90.1% for noisy-Lombard, and 99.8% for normal test speech). No significant further improvement was achieved by the inclusion of R_3 [see Fig. 4(d)].

For noisy-Lombard speech, the best combination of difference-based features found in [11] was D+A, with $\delta_D = 60$ msec, $\delta_A = 70$ msec and $\Delta_A = 60$ msec. The best recognition rate achieved by difference features was 84.4% (compared to the 90.1% achieved by regression features $R1+R2$). The better performance of the regression implementation may in part be due to the fact that high order difference features have more adjustable parameters than do regression features, and are therefore harder to optimize.

The choice of regression window length is influenced by the effect of other smoothing or "sharpening" operations. The length of the regression window controls the amount of smoothing in the time domain. Long regression windows compensate for excessive "sharpening" due to high order temporal derivatives. For single features, higher order derivatives require longer window lengths.

Smoothing, due to reducing the number of cepstral coefficients, interacts with "sharpening" due to high order temporal derivatives. We attempted to use this interaction to maintain high recognition rates when the regression window length is reduced. We found that recognition rate for short (sub-optimal) regression window lengths could be improved by sharply reducing the number of cepstral coefficients. However the best recognition rates were obtained with long regression windows and 12 cepstral coefficients.

The results considered above may be sensitive to the size and confusablity of the vocabulary. Study of regression features should be extended to a more confusable vocabulary, where finer phonetic distinctions must be made.

5. CONCLUSIONS

The following conclusions are drawn for our small, and relatively non-confusable, digit vocabulary.

1. *Individual features:* The smallest effective regression window length increases with regression order for either normal or noisy-Lombard test speech. For noisy-Lombard test speech, the zero-th regression feature R_0 exhibits two locally optimal window lengths (10-50 msec and 250-350 msec). When R_0 is combined with the features R_1 and R_2, the R_0 feature with the shortest window (which is equivalent to the static feature) is preferred.

2. *Comparison of regression and difference implementations:* The regression and difference implementations of the zero-th, first and second temporal derivative features exhibit the same general trends. For single features, recognition rate rises sharply with window length at short window lengths, and changes much more slowly for slightly longer window lengths. In either the regression or difference implementation, the highest recognition rates are achieved by the combination of first and second derivative features. Introducing the zero-th order regression feature reduces the recognition rate for noisy-Lombard speech input. The recognition rate for noisy-Lombard speech with the best combination of regression features is more than 5% higher than that with the best combination of difference features.

3. *Cepstral smoothing:* The choice of regression window length is influenced by the number of cepstral coefficients. Reducing the number of cepstral coefficients can partially compensate for short, sub-optimal, regression windows. However cepstral smoothing did not improve recognition rates for features with near-optimal window lengths.

4. *Optimal recognition with combined features:* A second temporal derivative feature, in either a regression or difference implementation, is useful for recognizing noisy-Lombard speech. A third temporal derivative feature (R_3) yields no further gain in recognition rate. Combining temporal derivative features is a robust way to address recognition of noisy and Lombard speech, as the best combination of features for recognition of noisy-Lombard speech (R_1+R_2) also performs well for normal speech input.

REFERENCES

[1] Shikano, K., **Evaluation of LPC Spectral Matching Measures for Phonetic Unit Recognition**, CMU-CS-86-108, Carnegie Mellon Univ., Pittsburgh PA, 1986.

[2] Gupta, V. N., M. Lennig, and P. Mermelstein, "Integration of Acoustic Information in a Large Vocabulary Word Recognizer," **ICASSP**, p. 17.2, 1987.

[3] Lee, K.-F., **The SPHINX System**, CMU-CS-88-148, PhD. Thesis, Carnegie Mellon Univ., Pittsburgh, PA, 1988.

[4] Furui, S. and A. E. Rosenberg, "Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech," **ICASSP**, pp. 1060-1062, 1980.

[5] Furui, S., "Cepstral Analysis Technique for Automatic Speaker Verification," **ASSP**, vol. 29, pp. 254-272, 1981.

[6] Furui, S., "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," **ASSP**, vol. 34, pp. 52-59, 1986.

[7] Furui, S., "Speaker-Independent Isolated Word Recognition Based on Emphasized Spectral Dynamics," **ICASSP**, pp. 1991-1994, 1986.

[8] Furui, S., "A VQ-Based Preprocessor Using Cepstral Dynamic Features for Speaker Independent Large Vocabulary Word Recognition," **ASSP**, vol. 36, pp. 980-987, 1988.

[9] Soong, F. K. and A. E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," **ICASSP**, pp. 877-880, 1986.

[10] Junqua, J.-C., "Evaluation of ASR Front-end in Speaker Dependent and Speaker Independent Recognition," **JASA**, Supp. 1, vol. 81, p. S93, May 1987.

[11] Hanson, B. A. and T. H. Applebaum, "Robust Speaker-Independent Word Recognition using Static, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech," **ICASSP**, 55S-15b.13, 1990.

[12] Draper, N. R. and H. Smith, **Applied Regression Analysis**, Wiley, New York, 1981.

[13] Hermansky, H., B. A. Hanson, and H. Wakita, "Low-dimensional Representation of Vowels Based on All-Pole Modeling in the Psychophysical Domain," **Speech Communication**, vol. 4, pp. 181-187, 1985.

[14] Hanson, B. A. and H. Wakita, "Spectral Slope Distance Measures with Linear Prediction Analysis for Word Recognition in Noise," **ASSP**, vol. 35, pp. 968-973, 1987.

[15] Applebaum, T. H., B. A. Hanson, and H. Wakita, "Weighted Cepstral Distance Measures in Vector Quantization Based Speech Recognizers," **ICASSP**, p. 27.9, 1987.

[16] Massaro, D. W., "Perceptual Images, Processing Time and Perceptual Units in Auditory Perception," **Psych. Review**, vol. 79, pp. 124-145, 1972.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A New Method to Improve Speech Recognition in a Noisy Environment

H.G. Hirsch, A. Corsten
Technical University of Aachen, FRG
Templergraben 55, D-5100 Aachen

1. INTRODUCTION

In many applications for speech recognition systems the recognition rate decreases in the presence of background noise (e.g. voice dialing in a car or controlling machines in a workshop hall). Therefore, several preprocessing algorithms and methods to improve the robustness of recognition have been proposed. Single microphone /1/,/2/,/3/ and multi-microphone methods /1/,/4/,/5/,/6/ for speech enhancement have been developed.

These methods can be applied to speech recognition too. Improvements, using a single microphone spectral subtraction technique, have been reported in e.g. /7/,/8/. However, the use of this method is restricted to nearly stationary noise situations. Furthermore, severe difficulties arise from the required speech pause detection.

This automatic distinction between noise and noisy speech segments is a very difficult task if the noise signal is not stationary. In standard word recognizers one of the major reasons for the increase of error-rate in noisy environments is the failure of the word boundary detection.

The author has developed earlier a method to improve recognition accuracy of reverberated speech /9/. This approach has been adapted to the problem of noise-suppression. The most interesting aspect is that no speech pause detection is required.

At first this new algorithm will be described. Then the speech recognition system used for this study will be shortly introduced. Finally, the implementation of the noise reduction scheme as part of the analysis section of the speech recognizer will be described. The results will be shown for different noise situations with various S/N-ratios.

2. NOISE REDUCTION BY FILTERING SUBBAND ENERGY SIGNALS

A method has already been developed /9/ to improve speech recognition in the situation of a hands free speech input in a room with a certain reverberation time. It is known that reverberation has a low pass effect on the temporal subband energy contours /10/,/11/.

With this knowledge the idea of an inverse high-pass filtering of subband energy signals was introduced to reduce reverberation and improve recognition rates. In the following it will be shown that this kind of spectral preprocessing can be used for noise suppression too.

The basis is a short-term spectral analysis. If the noise is short-term stationary, the contributions to the subband energies will be almost constant. Stationary noise can be interpreted as DC-components of subband energies after a spectral analysis of the temporal subband energy signals.
Under this assumption the noise can be eliminated by applying high-pass filtering to the subband energy signals. This procedure is related to the spectral subtraction technique /2/,/3/, which also tries to reduce the DC-component. However, the interesting feature of the new method is that no speech pause detection is required. The principal processing steps are shown in figure 1.

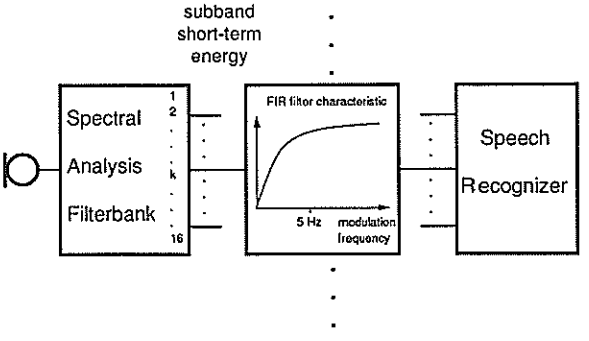


Figure 1: Speech recognition with high-pass filtered subband energy signals

The short-term subband energy is estimated with an analysis filterbank. Then each temporal subband energy signal is filtered with a high-pass FIR-filter.

The terms modulation and modulation frequency are introduced to describe the temporal fluctuation of subband energy. The filtered subband energy signals are used as input for the speech recognizer.

It is known that only modulation frequencies of up to 25 Hz exist in speech signals /10/. The spectral peak of the energy signal is located at about 3 Hz.
A filter characteristic which compresses completely the DC-component while frequencies above a modulation frequency of 3 Hz are not suppressed at all has given the best recognition results. This transfer function implies a total suppression of stationary noise components as well as a suppression of noise components with low modulation frequencies.
The same constant FIR filter is used in each band. Any kind of adaption is not necessary.
Different versions of the spectrum of a speech segment are shown in figure 2.

It can be recognized that noise is reduced and that speech does not deterioate considerably by this kind of preprocessing.
After filtering, negative values for the subband energy may occur. One possibility is to set these negative values to zero.
However, problems may occur for the word end detection when using an isolated word recognizer. Therefore the following method has been introduced. At first negative values are set to zero. The magnitude of a negative value is used after detection of a word beginning. With this the word end detection as well as the recognition rate can be improved.

3. THE SPEECH RECOGNIZER

The basis of this speech recognition system is a "68000" microcomputer. A special integrated circuit (NEC7763) is used for speech analysis. This chip consists of an analog 16 channel filterbank where centre frequencies are spaced nonlinear at a bark scale. An estimation of the short-term subband energy as well as an A/D-conversion of these energy values are also integrated. The estimation of the spectra is done each 16 ms.

To use this system for isolated word recognition word boundary detection is based on energy thresholds.
A certain number of spectra are extracted with a trace segmentation algorithm. The comparison of a test template with a reference template is done by dynamic programming technique. A speaker independent recognition system was used for this study. The speaker independent reference templates in each word class are calculated using a clustering algorithm.

The vocabulary consists of 30 German words which could be used to serve a bank automaton.
Three reference templates were calculated for each word class. With this a speaker independent recognition rate of about 98 % can be reached in a noise free environment using a close by talking microphone.

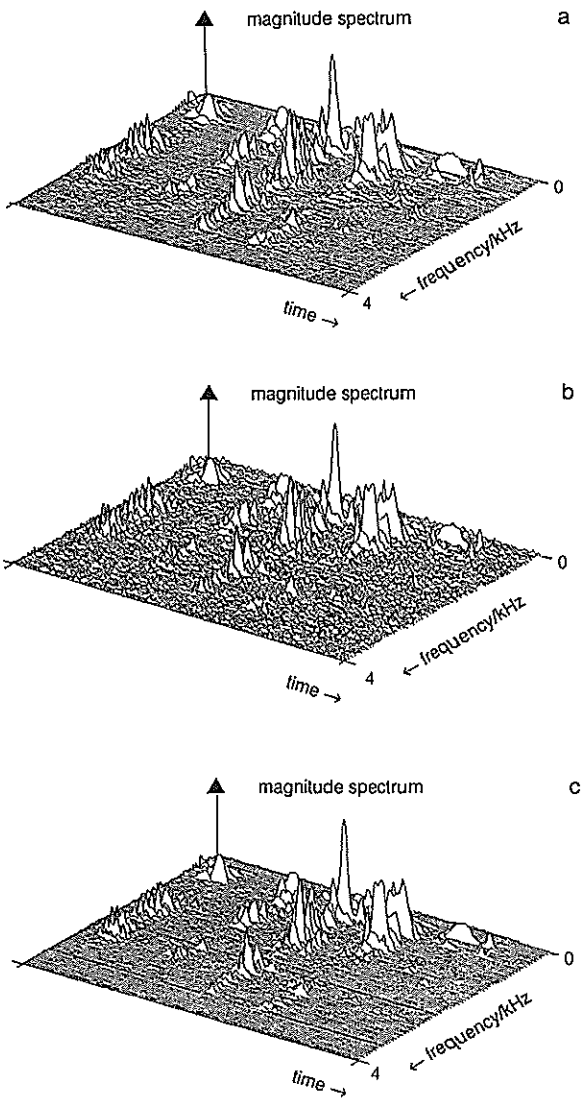


Figure 2: Spectra of a speech segment

a) original
b) disturbed with white Gaussian noise
c) after high-pass filtering of subband energy signals

4. RECOGNITION IN A NOISY ENVIRONMENT

The high-pass filtering of the subband energy signals was integrated into the speech recognition system. The FIR-filtering in the 16 channels was accomplished with the miroprocessor. To achieve real time only a simplified version of the high-pass impulse response is used such that no multiplication and only shift operations are required.
The impulse response and the frequency response are shown in figure 3.

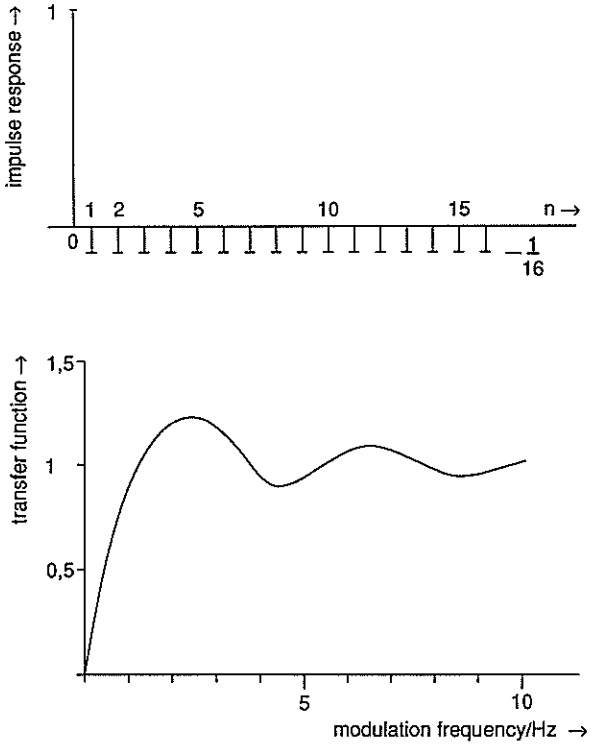


Figure 3: Impulse Response and frequency response of the high-pass filter

To test the recognition system with different noise signals at various S/N-ratios the following set up was used. 300 test words of the 30 word vocabulary from 10 speakers were recorded on a DAT recorder. Noise signals were recorded on a standard audio cassette recorder. The analog signals of DAT and cassette recorder were added with an audio mixer. The output of the audio mixer was connected to the analog input of the speech recognizer. In this way always the same test words can be used and a variation of the S/N-ratio is possible. The results when using white Gaussian noise to disturbe speech can be seen in the following figure.

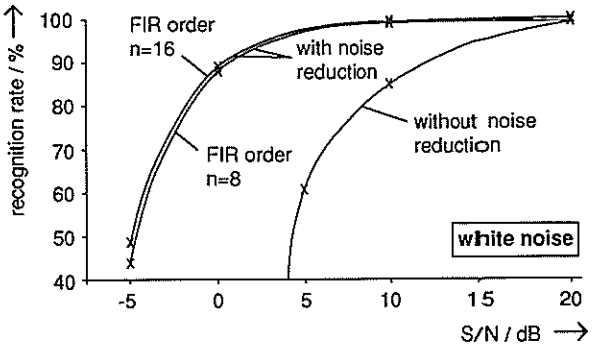


Figure 4: Recognition rates of noisy speech with and without noise reduction

Figure 4 shows a considerable improvement of the recognition rate by using the noise reduction technique. Reference templates have been estimated from undisturbed speech signals in the case without noise suppression. In the other case references were also taken from undisturbed speech but after preprocessing with the high-pass filtering technique. It can be seen in the figure that the results with the filter of 16. order are slightly better in comparsion to the filter of 8. order. A gain of more than 10 dB can be obtained at a recognition rate of 95 %.
The following figure give the results for noise signals recorded in a car.

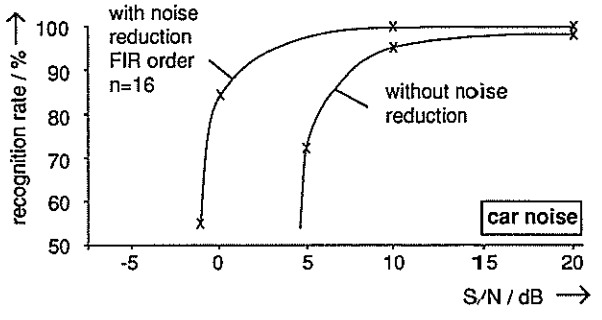


Figure 5: Recognition rates of noisy speech with and without noise reduction

In this application the improvement is not as significant as in the case of using white noise. The reason for this is the fact that the car noise is not stationary due to e.g. changing the gear or using the blink operator.

5. CONCLUSIONS

A noise reduction method has been presented which is based on high-pass filtering of subband energy signals. The main advantage of this method is the fact that no speech pause detection is necessary nor any kind of adaption.
A considerable improvement of recognition rate has been obtained with this method in a speaker independent recognition system.

6. REFERENCES

|/1/ Lim,J.S.:
Speech Enhancement, Prentice-Hall, 1983

/2/ Vary,P.:
Noise Suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits, Signal Processing, 1985, pp. 387-400

/3/ Boll,S.F.:
Suppression of Acoustic Noise in Speech Using Spectral Subtraction, IEEE ASSP, Vol.27, No.2, 1979, pp.113-120

/4/ Widrow et al.:
Adaptive Noise Cancelling: Principles and Applications, Proc. of the IEEE, Vol. 63, No.12, 1975, pp. 1692-1716

/5/ Van Compernolle,D. et al.:
Speech Recognition in Noisy Environments with the Aid of Microphone Arrays, Proc. European Conference on Speech Communication and Technology, 1989, Paris, pp. 657-660

/6/ Zelinski,R.:
A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms, Proc. ICASSP 88, New York, 1988, pp. 2578-2581

/7/ Hirsch,H.G. et Rühl,H.W.:
Automatic Speech Recognition in a Noisy Environment, Proc. European Conference on Speech Communication and Technology, 1989, Paris, pp.652-655

/8/ Rühl,H.W.et al.:
Speech Recognition in the Noisy Car Environment, Proc. European Conference on Speech Communication and Technology, 1989, Paris, pp.262-265

/9/ Hirsch,H.G.:
Automatic Speech Recognition in Rooms, Signal Processing IV: Theories and Applications, Eusipco1988, pp.1177-1180

/10/ Houtgast,T. et al.:
Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function. I. General Room Acoustics, Acustica 46, 1980, pp.60-72

/11/ Houtgast,T. et Steenecken,H.J.M.:
The Modu-lation Transfer Function in Room Acoustics, Technical Review Brüel & Kjaer, No. 3, 1985

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A COMPARATIVE STUDY OF FEATURE EXTRACTION METHODS FOR NOISY SPEECH RECOGNITION

Juan Gómez - Mena, Luis Sánchez - Sandoval* and Ramón García - Gómez.

Escuela Técnica Superior de Ingenieros de Telecomunicación
Dpto. Señales, Sistemas y Radiocomunicaciones. UPM
Ciudad Universitaria s/n, 28040 Madrid.

ABSTRACT: In real life situations, the background noise variability is large and some means of correction has to be provided to overcome unwanted degradation. Several solutions have been proposed, although it is possible to find common features, in this manner, we can classify these methods in two groups: the first one includes those techniques developed to reduce or cancel the interfering signal -based, principally, in estimation theory- which can be called *'probabilistic models'*. And, the second group deals with feature extraction methods with low sensitivity to disturbing signals. The main concern here is to improve the evaluation of parameters and to obtain distance or distortion measures more robust to be appropiate for any kind of situation.

1. INTRODUCTION

Speech recognition systems have now acquired considerable degree of development and almost perfect performance in controlled environments such as noise-free laboratory conditions. But, it is in a more practical situation - for example: an office or a car- where the system deteriorates to an unacceptable level. Noise is considered the main responsable of such degradation. Since recognition systems are based in parametric models like the Hidden Markov Model (HMM), they are very sensitive to disturbances in the parameters and it has been established that the presence of additive noise in the incoming signal modifies the value of these parameters. For example, the norm of the cepstrum vector for noisy speech is smaller than the norm of the cepstrum for clean speech, being smaller with smaller signal-to-noise ratio (SNR) [5].

Furthermore, In real life situations, the background noise variability is large and there could have been existed different background acoustic conditions when the system was trained and when it was tested. This is not very unlikely for most of the places where a recognizer could be used and located. Hence, some mean of correction has to be provided to overcome unwanted degradation.

Several solutions have been proposed, although it is possible to find common features, in this manner, we can classify these methods in two groups: the first one includes those techniques developed to reduce or cancel the interfering signal -based, principally, in estimation theory- which can be called *'probabilistic models'*. And, the second group deals with feature extraction methods with low sensitivity to disturbing signals. The main concern here is to improve the evaluation of parameters and to obtain distance or distortion measures more robust to be appropiate for any kind of situation. In such context, this approach can be named *'robust parameters and distances'*.

Accordingly, this communication is divided in four parts, the first and second are dedicated to review, briefly, the probabilistic models and the robust parameters and distances. In the third part, we describe our own experiments taking, basically,

(*) ENEP-Iztacala, Universidad Nacional Autónoma de México. Proyecto en Neurociencias, Apdo. postal 314, Tlalnepantla, 54030 México.

the second approach and we make a few comparisons with results published in current literature. Finally, in part four, we give our comments and conclusions.

2. PROBABILISTIC MODELS.

Recognition of noisy speech could involve speech enhancement, its objective is to improve one or more aspects of degraded speech to restore its quality. This goal is pursued by devising more reliable methods for estimation of parameters based on probabilistic considerations. Ideally, from statistical theory, estimation of unknown parameters from the observed data requires knowledge of the probability distributions (PD's) of the processes occuring. In the case of speech enhancement, the PD's of the speech and noise process are, usually, not known, in consequence, they are, either, assumed or approximated from the noisy speech, which, although restricted, allows a practical and useful estimation.

The basic procedures for the estimation of the parameter set of the speech model are the *Maximum a-posteriori (MAP) estimator* and the *Maximum Likelihood (ML) estimator* modified and simplified for specific recognition systems. For example, Hansen y Clements [2], simplify the MAP estimation to a minimum square error estimate (MMSE). Detailed description of the algorithms is given in the references and we just highlight here a few results, in table 1, of some of the probabilistic models.

Authors	Algorithm	Speech Model	Noise Model	Best results
Ephraim et al.	ML and Wiener filtering	discrete HMM	AR, white Gaussian	8 - 10 db. SNR improvement
Hansen and Clements.	MAP, constraints on pole locations and Wiener filtering	discrete HMM, LPC based.	white Gaussian and colored.	Improvement in recognition rate from 5% for none enhancement to 35% for constrained enhancement at 10db SNR.
Nadas et al.	MIXMAX	discrete HMM, filter-bank based.	Gaussian density	With an average of 31 db SNR, an error rate of 9 9% for quiet/noisy case and 7% for noisy/quiet case.

TABLE 1.- Highlights of some of the probabilistic models.

3. ROBUST PARAMETERS AND DISTANCES

In a great number of speech recognition systems, word identification is done comparing the representative features of the word to be recognized with the ones kept in the recognizer's data base, known as reference patterns. The quantity of similarity is established by a distance or distortion measure which is inversely related, that is: the smaller the distance, the greater the similitude or, in other terms, more confidence in the identification. Hence, a good distance measure is a critical issue in the recognition task. One of the most desired properties of a distance is insensitivity to disturbances produced by -in almost all cases- noise and several proposals have appeared to reach such situation.

Among the several features that could be extracted from the speech signal, the spectral characteristics are belived to be the most representative. One of them, the cepstrum, is a parameter commonly used in the formulation of distance measures, being the simplest *the Euclidean cepstral distance* given by:

$$d = \sum_{i=1}^{L} \left(c_i - c_i' \right)^2$$

where $\{ c_i \}$ and $\{ c'_i \}$ are the cepstra corresponding to the all-pole models $1/A(z)$ and $1/A'(z)$ for the observed and the reference speech signals, respectively. L is the length of the cepstrum vector.

Although the Euclidean distance works fairly well in clean environments, it was found that, for noisy speech, its performance drops drastically, making the recognizer almost useless. One of the approaches proposed to add some inmunity to disturbances has consisted in weighting the cepstrum vector by some function or window that, in a way, gives some robustness to the distance against modifications introduced by noise [3], [5], [8]. Hence, the weigthed cepstral distance measure is:

$$d = \sum_{i=1}^{L} (w_i(c_i - c_i'))^2$$

Juang et al. [3], propose *'the bandpass liftering'* window defined by:

$$w_k = 1 + \frac{L}{2} sin\left(\frac{\pi k}{L}\right) \quad for\ k = 1\ to\ L$$

When $w_k = k$ the weighted distance is referred to as *'root power sums'* [8]. A recent study [5], has investigated the effects of noise on the norm of cepstral vectors which has led to the following distance:

$$d = |c'|(1 - cos\beta)$$

where $cos\beta$ is defined by:

$$cos\beta = \frac{c \cdot c'}{|c||c'|}$$

and $|\cdot|$ is the norm of the vectors. This distance has been called, *'the cepstral projection distance measure'*. All of this modified distances and some more [4], [9], have shown to improve the recognizer's performance. Although up to know there is not enough comparative studies to decide which is the best weight or modification, nevertheless it seems that there will be a best weight for a specific recognizer and every recognizer should try several before a decision is done.

Another approach has prefered to develop a more robust evaluation of the filter coefficients *A(z)*. If this would be possible, the Euclidean distance measure with minor modifications could be used in a greater range of SNR's. Mansour and Juang [6], took this line and called their development *'The Short-time Modified Coherence (SMC) representation'*. They argue that evaluation of the LPC coefficients from the autocorrelation sequence followed by a spectral shaper is more robust than the standard LPC analysis. Some of their results are given in Table 2. The Euclidean cepstral distance with bandpass liftering was the distance used. For a complete description of the system the reader should see the reference. An additional advantage of the latter approach is that it can also be applied in recognition systems where the identification of the unknown word is done by a probability value instead of a distance. as will be described in the next section.

SNR (db)	SMC Cep	LPC Cep
60	94.8	95.6
20	89.2	73.3
15	84.2	54.3
10	78.2	30.5
5	64.6	12.56
0	45.0	
−5	22.2	

TABLE 2.- Correct recognition percent for alpha-digits.

4. ROBUST SPEECH RECOGNITION.

Since the proposition of a robust estimation of the A(z) coefficients was interesting, We decided to try this method on a recognizer different from the one reported in [6]. The system chosen was built from continuous HMM based on LPC parameters. This system has proven a good recognition rate in clean environments, and the idea was to explore how those implementations worked in another system and, at the same time, to introduce robustness on it.

The recognizer was trained with 35 spanish words spoken by 5 male speakers and one repetition. The vocabulary consisted of numbers [cero(0), uno(1), . . . , nueve(9)], commands, city names and people names. Training conditions were considered clean (SNR ≧ 40 db.). A second repetition was used for testing and adverse conditions were simulated adding filtered white gaussian noise.

Results are given in table 3. On the first row the values obtained for a standard LPC analysis. The features used were cepstrum, regression on cepstrum and regression on energy. On the second row the recognition rate when the SMC method was used before the evaluation of the cepstrum.

SNR	≧40	20	15	10	5
LPC standard	0.6	32	65	71	95
SMC	0.6	1.1	4	12	33

TABLE 3.- Percent of error on recognition rate for a recognizer based on continuous HMM.

One of the first things to note is that the SMC method introduced an improvement of ,approximately, 15 db in the SNR with respect to the standard LPC. This results are in agreement when compared with the ones on [6], which could confirm that the SMC method is versatil and could be included in the feature extraction block of the speech recognition systems. But we belive that more studies should be done.

REFERENCES

[1] Ephraim Y., Malah D. and Juang B. H., "On the application of Hidden Markov Models for enhancing noisy speech", Proc. IEEE-ICASSP, pp. 533-536, 1988.

[2] Hansen J. H. L. and Clements M. A., "Constrained iterative speech enhancement with application to automatic speech recognition", Proc. IEEE-ICASSP, pp.561-564, 1988.

[3] Juang B. H., Rabiner L. R. and Wilpon J. G., "On the use of bandpass liftering in speech recognition", Proc. IEEE-ICASSP, pp. 765-768, 1986.

[4] Li J. and Krishnamurthy A. K. "A modified frequency-weighted Itakura spectral distortion measure", IEEE Trans. Acoust. Speech, Signal Processing, vol. 37, No. 10, pp. 1614-1617, 1989.

[5] Mansour D. and Juang B. H., "A family of distortion measures based upon projection operation for robust speech recognition", Proc. IEEE-ICASSP, pp. 36-39, 1988.

[6] Mansour D. and Juang B. H., "The short+time modified coherence representation and its application for noisy speech recognition", Proc. IEEE-ICASSP, pp. 525-528, 1988.

[7] Nádas A., Nahamoo D. and Picheney M. A. "Speech recognition using noise adaptive prototypes", IEEE Trans. Acoust. Speech, Signal Processing, vol. 37, No. 10, pp. 1495-1503, 1989.

[8] Schroeder M. R. "Direct (nonrecursive) relations between cepstrum and predictor coefficients", Proc. IEEE-ICASSP, pp. 297-301, 1981.

[9] Soong F. K. and Sondhi M. "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise", IEEE Trans. Acoust. Speech, Signal Processing, vol. 36, No. 1, pp. 41-48, 1988.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ACOUSTIC-PHONETIC STUDY OF LOMBARD SPEECH IN THE CASE OF ISOLATED-WORDS

Yolande ANGLADE and Jean-Claude JUNQUA *

C.R.I.N. / I.N.R.I.A LORRAINE BP239 F-54506 Vandoeuvre-les-Nancy cedex and
* SPEECH TECHNOLOGY LABORATORY, Santa Barbara, California.

In this paper, we investigate the acoustic-phonetic changes between speech produced in quiet and speech produced in noise (Lombard effect). In addition to replicating previous studies, we examined more parameters, studied the influence of the context on each phoneme and the incidence of the Lombard effect on male and female speakers. Our results attest the important influence of the Lombard effect on the speech acoustic parameters and the great variability of this influence among speakers.

1 INTRODUCTION

In the presence of noise, speakers modify their vocal effort, by what is called the Lombard effect, in order to preserve their message's intelligibility. Little effort has been made to report the acoustic-phonetic changes due to this effect, but recently [1, 2] it has been shown in the case of speech recognition that the speech production variation caused by noise exposure at the ear is far more degrading than that caused by exposure to ambiant acoustic noise. Several studies found distinctive differences between normal and Lombard speech [3, 4, 5]. These studies analyzed prosodic and spectral features, at a phonetic or word level. Their main purpose was to improve performance of recognizers in the presence of noise by including knowledge about the acoustic changes. However, the results obtained are not yet satisfactory. In order to define the differences, we extracted at a phoneme level, from a database of isolated-words, some representative features used in speech recognition, at a phoneme level. The results were analyzed with statistical tests that indicate significant changes. In this paper, we will give a global synthesis of the results obtained.

2 DATABASE

We conducted acoustic analyses on a database which is a super-set of the alphanumeric data plus control words (49 english words, see Table 1).This vocabulary was produced by 10 speakers, 5 males and 5 females from different states of the U.S.A., twice under normal conditions and twice under noisy conditions (white-Gaussian noise at 85 dB SPL through calibrated headphones TDH49). All words were recorded in isolation (at 10kHz) and manually labeled at the phonetic level. Most of the American phonemes are present in the database. We ran all the statistical tests on 25 of them, and 9 were not numerous enough to be entirely treated. The phonemes were grouped into several classes (see Table 2).

Before conducting the acoustic analyses, the database was manually labeled into 5280 phonemes.

a	b	c	d	e	f	g
h	i	j	k	l	m	n
o	p	q	r	s	t	u
v	w	x	y	z	zero	one
two	three	four	five	six	seven	eight
nine	enter	erase	go	help	no	off
on	repeat	right	rubout	start	stop	yes

Table 1 : the vocabulary studied.

3 STATISTICAL TESTS

We used several statistical tests [6] (with a decisive probability of 0.05) to extract the significant differences between normal and Lombard speech, for each phoneme and each parameter computed on the database:

1. a Student T-test applied to the following:
   - all speakers for the phoneme in all phonetic contexts,
   - each speaker for the phoneme in all phonetic contexts,
   - all speakers for the phoneme in each phonetic context.
2. a two-way analysis of variance (noise and context).
3. some preliminary tests, in order to run the Student T-test and the two-way analysis of variance according to their hypotheses (same variance for the two populations and Normal distribution).

We examined the results taking into account all the speakers, and male and female speakers separately.

vowels and diphtongs	/IY/, /IH/, /EY/, /EH/, /AA/, /AO/, /UW/, /OW/, /AH/, /EL/, /AX/, /UR/, /AY/, /AW/
glides	/Y/, /W/
liquids	/L/, /R/
nasals	/M/, /N/
fricatives	/HH/, /F/, /TH/, /S/, /V/, /Z/
affricates	/CH/, /JH/
plosives	/P/, /T/, /K/, /B/, /D/, /G/

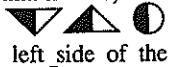
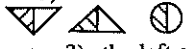
Table 2 : phonetic classes (Arpabet notation)

4 PARAMETERS

The parameters presented in Table 3, computed on a phoneme basis, were selected because of their importance in speech recognition. They have been extracted using Snorri [7] which is a speech analysis tool developed at C.R.I.N. / I.N.R.I.A. LORRAINE. Most of the features were calculated at three different points of each phoneme, such that the influence of its left and right contexts was minimized. Then these three values were filtered, and their average was used for the statistical tests.

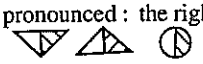
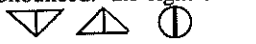
5 RESULTS

Studying the results of the statistical tests, our purpose was to answer the following questions (in the following, the symbols used in the Tables 4 and 5 are given):

1. is there a general evolution tendency for the parameter?
2. is this tendency :
   - a diminution ? : symbol 
   - an augmentation ? : symbol 
   - a stability ? : symbol 
3. is this tendency speaker dependent (indicated as described below by the left side of the symbols) ?
   - not at all, or not much (the number n of speakers that follow the tendency is such that n >= 8): the left side of the symbol is black 
   - pronounced (8 > n >= 5): the left side of the symbol is striped 
   - very pronounced (5 > n >= 3): the left side of the symbol is white 

Duration (phoneme and word)	Formants F1, F2, F3, F4 with their bandwith
Energy between 0-250Hz	Pitch
Energy between 250-500Hz	Zero-crossings
Energy between 500-1000Hz	Low-band spectral tilt 0-2000Hz
Energy between 1000-2000Hz	High-band spectral tilt 2000-5000Hz
Energy between 2000-3000Hz	lowest frequency of the fricative energy
Energy between 3000-4000Hz	Norm of the cepstral coefficients
Energy between 4000-5000Hz	Burst (strength, frequencies of high energy points)
Spectral energy center	

Table 3 : parameters selected.

4. is this tendency dependent on the phoneme context (indicated as described below by the right side of the symbols) ?
   - not at all, or not much: the right side of the symbol is black 
   - pronounced : the right side of the symbol is striped 
   - very pronounced: the right side of the symbol is white 

Some symbols are cut along the vertical ligne in the graphical tables of results (in fact half of the symbol is represented). This corresponds to cases where the tendency is different for female and male speakers. For these cases it was not possible to analyze the influence of the phonetic context because the number of phonemes was not sufficient to run the statistical tests.

Since it is impossible to report all the results in this paper, we will give here :

1. two graphical tables showing some interesting differences in the values of selected parameters,
2. the main results of our work for all the parameters studied.

For all the results, the utterances produced without noise are taken as references.

5.1 Graphical tables

Table 4 and 5 show, for female and male speakers respectively, the changes that occur between normal and Lombard speech for several parameters : the energy in

phon.	Energy 0-250Hz	Energy 250-500Hz	Formant 1	Zero - cross.	Pitch
IY					
IH					
EY					
EH					
AA					
AO					
OW					
UW					
AH					
AY					
Y					
W					
L					
R					
N					
F					
S					
V					
Z					
P					
T					
K					
B					
D					
G					

Table 4 : Variations concerning the energy in the two first frequency bands analyzed, the first formant, the pitch and the zero-crossings parameters in the case of female speakers.

phon.	Energy 0-250Hz	Energy 250-500Hz	Formant 1	Zero - cross.	Pitch
IY					
IH					
EY					
EH					
AA					
AO					
OW					
UW					
AH					
AY					
Y					
W					
L					
R					
N					
F					
S					
V					
Z					
P					
T					
K					
B					
D					
G					

Table 5 : Variations concerning the energy in the two first frequency bands analyzed, the first formant, the pitch and the zero-crossings parameters in the case of male speakers.

the two first frequency bands analyzed, the first formant, the pitch and the zero-crossings. It is interesting to note that the tendencies (described in the next section) are sometimes different in the two tables, especially for the pitch and the zero-crossings parameters.

5.2 Main results of our work

- *duration*: the consonants at the beginning and end of the words tend to be shortened, while the vowels are generally lengthened. This leads to an increase of the word duration;
- *energy*: between 0–500Hz, the energy of the vowels, diphtongs and liquids decreases for all the speakers (from 17% to 37%). The same tendency is observed up to 1000Hz, but in a more moderate way. We observed an increase in energy between 4000–5000Hz only for the females. The context has no real effect on the results. The nasals, fricatives, affricates and plosives are associated with a decrease in energy for all the frequency bands;
- *spectral gravity center*: it increases for all the speakers and all the phonemes (particularly for the vowels and diphtongs);
- *formants*: there is a general increase of the first formant for the vowels and diphtongs. This tendency

does not depend on the speaker for the females and is more variable for males. Furthermore, it is highly independent of the phoneme context. The second formant follows the same pattern but only for the females;

- *pitch*: there is an increase for the vowels and diphtongs which is independent of the phoneme context. It is more pronounced for males. We observed a great variability among speakers for the females;
- *zero-crossings*: they are more numerous for the vowels, diphtongs, liquids, glides and nasals for all the female speakers and in all contexts;
- *spectral tilt*: in the low frequency bands, the spectral tilt increases for most of the vowels, the nasals and the liquids. On the other hand, it tends to decrease for these phonemes in the high frequency bands;
- *lowest frequency of the fricative energy*: it generally increases. This tendency is very obvious for the phonemes /S/ and /Z/ for all speakers and in all contexts;
- *norm of cepstral coefficients*: it decreases for the vowels and diphtongs (between 15 and 30%) but, depending on the phoneme, there is a slight speaker dependence and a context independence;
- *burst*: there is a decrease of the burst strength for all the plosives except /B/. The frequencies of highest energy of /T/ and /G/ increase (about 25%), while the ones for /K/ decrease.

We also studied two other phenomena that give us interesting characteristics of the Lombard effect : the insertions and the omissions of phonemes in the database. We observed that :

- the phoneme omissions concern some consonants placed at the end of words, particularly /T/ and /P/ and also /F/ in a more moderate way. These omissions are three times more numerous in Lombard than in normal speech. For the phoneme /P/, they take place for 15% of the database.
- the phonemes insertions concern some aspiration introduced at the end of the words M and N. They are twice more numerous in the Lombard speech. However, it is difficult to conclude taking into account the number of occurences manipulated.

6 CONCLUSIONS

These results can be compared to those of previous studies concerning the Lombard effect [3, 4, 5] for the common features extracted. They are in agreement for the pitch, the first formant and the spectral energy center. Regarding the energy, we can notice a difference in the intermediate frequency bands in which Stanton found a significant increase. The origin of that difference can be due to the heterogeneity of our database containing female and male voices (which seem to have different tendencies of energy variations in some frequency bands), and also to the inter-speaker variability. As a matter of fact, we noticed that the incidence of the Lombard effect, due to the increase of the vocal effort, is variable from one speaker to another; for some parameters, the degree of variation between normal and Lombard speech evolves up to a ratio of 3. Future work should correlate more precisely the influence of the Lombard effect with the increase of the vocal effort. To the best of our knowledge, this acoustic study of the Lombard effect is the only one conducted on both male and female speakers. It allowed us to show some interesting differences concerning the pitch, the zero-crossings, the energy in high frequency bands and the second formant. The analyses of the phonetic context shows that its influence is sometimes important, particularly for some consonants placed at the end of words. We are currently processing other features essentially derived from the perceptually-based linear prediction auditory model (PLP[8]). All these results attest the important influence of the Lombard effect on the speech acoustic parameters; our purpose now will be to exploit them in order to improve robustness of automatic speech recognizers in noise.

ACKNOWLEDGMENTS

The authors wish to thank Yves Laprie and Dominique Fohr from the C.R.I.N/I.N.R.I.A Laboratory for providing the software for some of the features extracted in the acoustic analyses.

BIBLIOGRAPHY

[1] P. Rajasekaran, G. Doddington, and J. Picone. Recognition of Speech under Stree and in Noise. In ICASSP-86, pages 733–736, 1986.

[2] J.C. Junqua, and H. Wakita. A comparative study of cepstral lifters and distance mesures for all pole models of speech in noise. In ICASSP-89, pages 476–479, 1989.

[3] W.V. Summers, D.B. Pisoni, R.H. Bernacki, R.I. Pedlow, and M.A. Stockes. Effects of noise on speech production: acoustic and perceptual analyses. J. Acoust. Soc. Am., 84(3):917–927, September 1988.

[4] B.J. Stanton, L.H. Jamieson, and G.D. Allen. Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions. In ICASSP-88, pages 331–334, 1988.

[5] Z. Bond, T. Moore, and B. Gable. Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask. J. Acoust. Soc. Am., 85(2):907–912, 1989.

[6] H.L. Alder, and E.B. Roessler, editors. Intoduction to Probability and Statistics. W.H. Freeman and Compagny, 1977.

[7] Y. Laprie. Notice d'utilisation de Snorri. Technical report, CRIN / INRIA, 1988.

[8] H. Hermansky, B. Hanson, and H. Wakita. Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain. Speech Communication, (4):181–187, 1985.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A COMPARISON BETWEEN MEL-SCALE CEPSTRUM AND AUDITORY MODEL REPRESENTATION FOR NOISY SPEECH RECOGNITION

Piero COSI*, Daniele FALAVIGNA**, Gian Antonio MIAN ***, Maurizio OMOLOGO**

*Centro di Studio per le Ricerche di Fonetica-C.N.R. P. Salvemini 13, 35131 PADOVA,ITALY
**Istituto per la Ricerca Scientifica e Tecnologica-Pante' di Povo, 38050 TRENTO,ITALY
***Dipartimento di Elettronica e Informatica-Via Gradenigo 6, 35100 PADOVA,ITALY

A joint synchrony/mean-rate auditory model, recently proposed by Seneff[6], is embedded into a classical DTW-based system for the recognition of Italian digits. Its performances are evaluated in both clean and noisy speech and compared with those of a system based on the Mel-cepstrum representation. Experimental results show that the Mel representation outperforms the auditory model. Problems encountered by the auditory model in noisy speech are outlined and suggestions for noise compensation techniques both inside and outside the model are given. Simple image processing techniques aiming to clean up the synchrony spectrogram in noisy speech are suggested and some promising preliminary results are presented.

1. INTRODUCTION

Speech recognition can suffer significant degradations in adverse environments, especially when test and training are performed on data pronounced in different noise conditions. Both robust distortion measures and different speech analysis features have been recently considered to overcome this problem [1] when white noise is added to clean speech. It turns out that non symmetric distortion measures are needed to cope with this problem. Furthermore perceptually significant distance measures could be well applied to this task.

Auditory model representations of speech signal have been successfully applied as front-end of speech recognition systems [2,3]. Moreover, auditory models have shown better recognition performance than classical signal processing techniques, when speech signal is significantly degraded by noise [4,5].

In this work a recently proposed joint synchrony/mean rate auditory model [6] was incorporated into a classical DTW(Dynamic Time Warping)-based [7] system for the recognition of the Italian digits, in both clean and noisy environmental conditions. The synchrony stage of this model was used, though it was explicitly designed for sonorant sounds and it could be unsuitable for general recognition purposes. For comparison, a standard Mel-cepstrum representation of the speech signal [8] was utilized together with the same DTW recognition scheme.

In the following a description of the system under study together with various experiments and comparison results will be given. Results have suggested the introduction of a noise compensation technique that has been applied as postprocessing to the model outputs. On the other hand, some modifications to the model appeare necessary to better adapt the resulting system to noisy conditions.

2. ANALYSIS METHODS

2.1 Ear Model

The proposed computational scheme for modelling the human auditory system follows the one proposed by S. Seneff [6,9,10,11]. The overall system structure is illustrated in Fig. 1 and includes three blocks: the first two blocks deal with peripheral transformations occurring in the early stages of the hearing process while the third one attempts to extract important information relevant to speech perception like spectral lines related to formants.

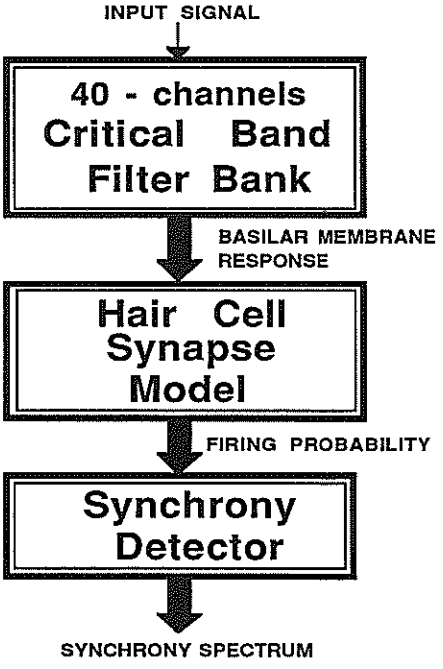


Fig. 1 - Block-diagram of the ear model

The speech signal, band-limited and sampled at 16 kHz, is analyzed by a 40-channel critical-band linear filter bank whose design was directly motivated by physiological considerations[10]. To each channel is then applied a nonlinear hair cell synapse model which is intended to capture prominent features of the transformation from basilar membrane vibration to probabilistic response properties of auditory nerve fibers. The "phase locking" property of nerve fibers [12] is implemented by the last block of the model called Generalized Synchrony Detector (GSD) [9,10], which enhances spectral peaks due to vocal tract resonances. A parallel output called Envelope Detector, can be obtained by averaging the discharge rate response.

2.2 Mel representation

A Mel based cepstrum analysis of the speech signal was performed according to [8]. Every 5 ms a set of 24 band-pass triangular filters, equally spaced in the Mel scale, is applied on the DFT transformed signal. The log-energies across the channels are then cosine transformed to obtain a vector of 8 coefficients.

3. DIGIT RECOGNITION TESTS

3.1 Speech material

The speech database used in the experiments described in the following consisted of ten repetitions of digits spoken in isolation by five male and five female speakers (i.e. 1000 digits). The speech was digitized at 16kHz with an accuracy of 12 bits.

3.2 Additive noise

In order to examine the impact of noisy speech on both representations, a Gaussian white noise was added to the input speech. Noise levels corresponding to an average SNR of 10 dB and 20 dB were adopted.

3.3 DTW-based recognition experiments

Speaker dependent recognition tests were performed to examine the robustness to noise of both the auditory model and the Mel representation. The recognition test is based on the application of a DTW algorithm with the following characteristics:

- the DTW optimal path is constrained to both the initial and the final frame. Since an accurate segmentation was previously performed on the described database, this constraint does not influence the recognition performance;
- for both representations an unweighted Euclidean distance measure is used;
- both global and local path constraints were introduced during preliminary experiments: the following results refer to a search area restricted to a range of seven frames centred around the locally optimum path [7];
- due to the insufficient available material, each reference consists of a single token; as a consequence, the estimated performance is expected to be less than the one corresponding to a more robust procedure for making up references.

3.4 Results

Preliminary recognition experiments were conducted by comparing reference and test digits with the same additive noise level. Actually, modeling the noise as a stationary process with a fixed level provides results that can be interpreted in many ways, as we shall discuss in the next section.

The error rates given in Tab. 1 show that the Mel representation outperforms the auditory model. However, some questions arise from these results. The main point is that the usage of an unweighted Euclidean distance on the 40 channel synchrony outputs may not be a suitable choice to fully exploit the auditory model properties.

On the contrary, the Cosine Transform(CT), used to compute the Mel coefficients, is known to approximate a principal component analysis thus allowing the use of both a more compact representation and the Euclidean distance. If applied to the auditory model, the CT introduces benefits with

SNR ref.	SNR test	Analysis	Error %
clean	clean	ear m.	2.1
clean	clean	ear m.+CT	1.4
clean	clean	mel	0.1
20	20	ear m.	2.5
20	20	ear m.+CT	1.4
20	20	mel	0.1
10	10	ear m.	9
10	10	ear m.+CT	4.5
10	10	mel	0.9

Tab. 1. - Recognition results with reference and test digits at the same noise level.

SNR ref.	SNR test	Analysis	Error %
10	clean	ear m.+CT	59
10	clean	mel	53
20	clean	ear m.+CT	25
20	clean	mel	14
10	20	ear m.+CT	12
10	20	mel	9

Tab. 2. - Recognition results with reference and test digits at different noise levels.

respect to the previous case, as reported in Tab.1. In this case, however, CT does not approximate a principal components analysis; further improvements are expected both by using a linear discriminant analysis and by combining the envelope and synchrony detector outputs [5].

When recognition experiments are performed by comparing test and reference digits subjected to different noise degradations, the performance of both representations decrease significantly. In all cases using noisy speech as reference provides better performance than using noiseless speech. In Tab. 2 results are given for experiments corresponding to reference digits more noisy than the test ones.

From the described experiments it turns out that a noise compensation technique is desirable to get some improvements.

4. EAR MODEL WITH NOISY SPEECH

When speech is corrupted by noise the synchrony detector performs well only during voiced segments. This effect can be attributed to the total absence of periodicity in the signal that causes flat excitation to each channel. As a consequence, higher channels show a "saturated" response that can be present even in voiced regions. In order to prevent this effect a compensation technique could be devised both inside the ear model itself [13] and outside it, as a postprocessor stage. Both strategies are being investigated.

A noise compensation technique inside the ear model requires some care not to degrade the global behaviour of the system with respect to noiseless speech, especially for unvoiced segments. Hence a noise estimation algorithm should be included to reduce the synchrony output level, when noise is supposed present in the speech itself. The noise estimation can be based either on the analysis of the input signal or on the synchrony outputs corresponding to previous frames. The last hypothesis is related to characteristics that are evident in the synchrony spectrogram for noisy speech, as illustrated in Fig. 2 c,d and as discussed in the following. Outside the model, a noise compensation technique can be conceived which exploits the previously described saturation effect on noisy speech.

Different noise compensation techniques have been proposed [14]. In the auditory model under investigation formant movements below 1000 Hz are generally preserved (Fig.2d), even with a high level of input noise. On the other hand, higher channels synchrony outputs show quite confused formant tracks, since the corresponding peak level is close to the noise output level. These formant peaks, however, preserve their connectivity.

Different algorithms were checked in order to clean the synchrony spectrogram by using simple image processing techniques. As a preliminary step, channel outputs were analyzed separately, by exploiting the output level and its derivative. But, in this case a formant with a constant center frequency can be degraded. As a second step, channel outputs were analyzed by windowing the spectrogram and subtracting a component which depends on the mean value, while preserving formant tracks. As a third step, both silence segments and stop closures were considered: the synchrony output of the lower channels can be well exploited as silence detector, even in the case of noisy speech. A broad threshold clipping algorithm has been defined, which subtracts a locally estimated noise level from higher channels when lower channels exhibit a low output level.

Digit recognition experiments (noisy test, noiseless reference) were repeated applying the last two compensation techniques to the synchrony outputs (Fig.2e). Preliminary results give an error rate of approximately 30%, to be compared with the corresponding 59% (see Tab.2), obtained with no compensation. This technique allows some improvements, though the overall performance is yet unsatisfactory. At present the described compensation technique requires to be improved in silence detection and in characterizing unvoiced sounds (e.g. strong fricatives, affricates) for which, however, an explicit modification of the auditory synchrony model should be looked for.

5. CONCLUSIONS

Comparing auditory-based recognition results with those obtained with Mel-cepstrum representation, leads us to conclude that noise compensation techniques together with a better DTW metric in the auditory space should be utilized in order to improve recognition performance. Moreover various modifications to the present auditory model could be conceived [6] to obtain better recognition performance, especially when speech is greatly degraded by noise.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. R. De Mori for some helpful discussions, G. Lovo for the initial development of this work, G. Antoniol and D. Giuliani for providing the Mel computation routines.

REFERENCES

[1] D. Mansour , B.H. Juang, "A Family of Distorsion Measures Based Upon Projection Operation for Robust Speech Recognition", IEEE Trans. on ASSP, vol.37, n.11, November 1989.

[2] J.R. Cohen (1989), "Application of an Auditory Model to Speech Recognition", JASA 85(6), June 1989, pp. 2623-2629.

[3] R, De Mori, Y. Bengio and P. Cosi, "On the Generalization Capability of Multi-Layered Networks in the Extraction of Speech Properties", Proc. IJCAI, August 20-25, 1989, Detroit, pp. 1531-1536.
[4] M. J. Hunt, C. Lefebvre, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model", Proc. IEEE ICASSP, April 11-14, 1988, New York, pp.215-218.
[5] M. J. Hunt, C. Lefebvre, "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech"", Proc. IEEE ICASSP, May 23-26, 1989, Glasgow, pp.262-265.
[6] S. Seneff (1988),"A joint synchrony/mean rate model of auditory speech processing", Journal of Phonetics (1988)16, pp.55-76.
[7] L. R. Rabiner, S. E. Levinson,"Isolated and Connected Word Recognition-Theory and Selected Application", IEEE Transactions on Communications, Vol.Com-29, No.5, May 1981.
[8] S. B. Davis, P. Mermelstein (1980), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuosly Spoken Sentences", *IEEE Trans. Acoust., Speech, and Signal Processing*, August 1980, vol. ASSP 28, n. 4, pp.357-366.
[9] S. Seneff (1984) , "Pitch and spectral estimation of speech based on an auditory synchrony model", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Diego, CA,1984.
[10] S. Seneff (1985), "Pitch and spectral analysis of speech based on an auditory synchrony model", *RLE Technical Report*, No. 504 , Mass. Inst. of Techn., 1985.
[11] S. Seneff (1986), "A computational model for the peripheral auditory system: application to speech recognition research", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Tokyo,1986, pp. 37.8.1-37.8.4.
[12] W.A. Yost, D.W. Nielsen (1985), *Fundamentals of Hearing. An Introduction (second edition)*, Holr, Rinehart and Winston, New York 1985.
[13] M. J. Hunt, C. Lefebvre (1986), "Speech Recognition Using a Cochlear Model", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Tokyo,1986, pp. 37.7.1-37.7.4.
[14] J. N. Holmes, N. C. Sedgwick, "Noise Compensation for Speech Recognition using Probabilistic Models", Proc. IEEE ICASSP, April 8-11, 1986, Tokyo, pp.741-744.

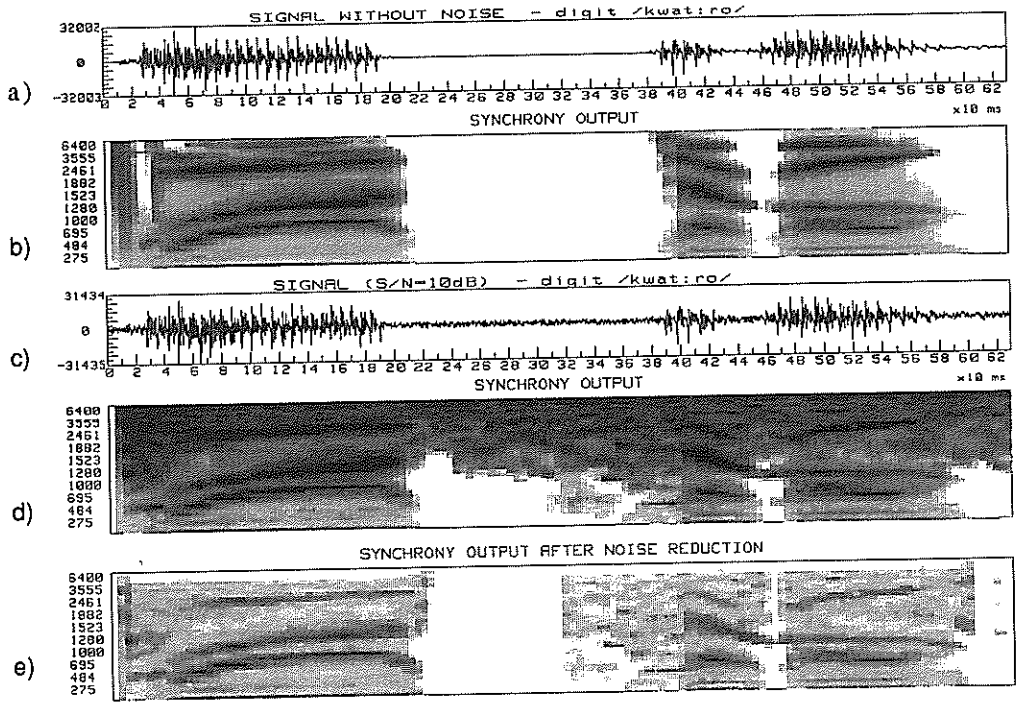


Fig.2. Speech signals and synchrony spectrograms for the digit "quattro"(four):
a) original speech; b) synchrony output; c) noisy speech (10dB); d) synchrony output; e)synchrony output after noise compensation.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

DESIGN OF AN ISOLATED WORD RECOGNITION SYSTEM OVER THE SPANISH TELEPHONE NETWORK

María Jesús Poza, José F. Mateos y Juan A. Siles

TELEFONICA, INVESTIGACION Y DESARROLLO. Emilio Vargas,6 28043 Madrid. Spain.

This paper describes the design of a system for automatic speaker independent recognition of isolated words spoken over the Spanish telephone network, in addition to the necessary steps to collect a data base with utterances of all the vocabulary words. The statistical features of these words are represented by hidden Markov models using multidimensional density functions. The recognition performance is evaluated for different LPC based feature vectors, for different model structures and under different conditions.

1. INTRODUCTION

Speaker independent recognition of small vocabularies, spoken over the telephone network has been demonstrated to be a viable technology, and Hidden Markov Modeling has become an increasingly popular technique in automatic speech recognition.

The use of such techniques allows us to develop new services in which the user can communicate with remote computers by voice. To offer these services we need a system able to recognize any speaker from the Spanish population, and, moreover, the system must be able to do it over the telephone network. As a first step we have concentrated our work on a small vocabulary: the ten digits and the 'si' and 'no' words.

With the goal of obtaining the recognizer structure that performs best on our vocabulary we have found it necessary to record our own data base (over 8000 Spanish speakers) from the telephone network. Also we have performed a set of experiments on isolated word recognition applying HMMs with a mixture of multivariate gaussian densities, and with these results, we have determined the final set of parameters for our recognizer.

The organization of this paper is as follows. In Section 2 we review all the subjects related with the data base: data base design, acquisition, verification, labelling and management, and the equipment we have used to collect and verify the recorded data files. In Section 3 we describe an isolated word speech recognizer, implemented with HMM ideas; first (in 3.1) its global structure, followed by an exhaustive description (3.2) of the experiments performed in order to select all the parameters for the recognition system. Section 4 summarizes the results from this set of experiments, concluding with the final structure of the complete recognition system.

2. DATA BASE

In order to train the twelve words of our recognition system to make it work properly, we need to collect a lot of utterances of the vocabulary words, spoken in an isolated manner over the telephone network. We must also control the origin of these utterances: we need to collect speech signals from several different speakers, and it would be good to cover all ages, both sexs, different cultural levels and physical origins. The recorded data base will be used in two ways: to make the recognizer learn the acoustic information contained, and afterwards, to test and evaluate it.

The technical equipment we have used to collect the files of the data base is composed of two identical 'collect stations', each one having a telephone line interfaz, converters A/D and D/A, an AT personal computer, a optical disk (WORM $5^{1/4}$") and the software needed for dialoging with the speaker and recording each utterance in a file.

We are also provided with two identical 'verification stations', each one having another AT computer, a DSP board, a D/A converter, headphones, an optical disk and the corresponding verification software, providing speech display, listening and edition facilities.

2.1. Data Base Acquisition

We have asked people from all the provinces in our country (one by one) to phone to a '900' number and to answer the questions our collect station poses to them. The software in the collect station includes pre-recorded messages that indicates to the person calling which word of the vocabulary theymust pronounce and when to do it. Each utterance of a word is recorded in a file and stored in the optical disk. Two more questions are asked to the person calling: his/her age and the postal code where he/she lives. In all, fourteen

files are recorded from each speaker.

Up till now, we have recorded utterances from almost 8000 different speakers.

2.2. Data Base Verification and Labelling

This is the most hard and expensive stage in the data base subject. The verification process can take a lot of time, as the task consists of accessing to each speech file stored in the acquisition phase and making an audiovisual analysis using the facilities provided by the verification stations. As a result of that study, the operator must take the following actions for each file:

- He must segment the samples, marking the end points of the speech in the file. There is a process that automatically marks these boundaries, but when the signal to noise ratio falls under a certain threshold, the operator must confirm and/or manually adjust the marks.

- He must label the speech signal contained in the file: sometimes the automatic labelling process fails.

- He must create a heading file with some information concerning the file, ie. noise level, connected/isolated speech, correct utterance, endpoints, etc.

So, we have two files for each utterance recorded, the speech file and the heading file. In a later process, both files will be merged in a new speech file, and transmitted from the optical disk to the final storage medium, and those files that happen to be bad files will be deleted.

2.3 Data Base Management

Once all the files had been verified and labelled, we built an ASCII file (the 'index' file), containing useful information per speech file, taken out from its heading. This information was:

- File name
- Speech endpoint
- Some characteristics of the speaker, ie. age, sex and postal code of his/her town
- Word/s uttered
- Isolated/connected speech
- Presence of special noises (tones from the telephone box ...)
- Noise level

And we have developed a 'handling program' to manage all this information. It is written in C language, and it provides facilities to access all the files fitting a group of given conditions (for example, "give me all the files containing the 'tres' word spoken in a isolated way by middle-aged men from Galician towns).

3. DESIGN OF THE RECOGNITION SYSTEM

In parallel with the data base acquisition, we have focussed our attention on the design and construction of a system for speech communication between man and machine. As a part of the system, we have designed a speech recognizer for our twelve word vocabulary spoken over the telephone line.

Fig. 1 shows the model used in the majority of isolated word speech recognition systems; there are four basic steps in the model:

- endpoint detection
- feature extraction
- score calculation
- decision rule.

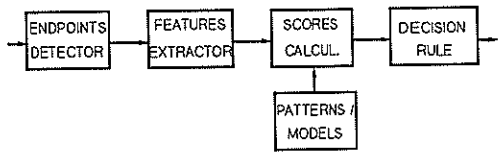


Figure 1. Recognition System.

The input to the model is the acoustic waveform of the spoken word embedded in background noise; the output is the word in the vocabulary the recognizer selects as the most probable one. Let us now consider our particular implementation of this model block by block.

3.1. Recognizer Structure

3.1.1. Endpoint Detector

The endpoint detection problem is nontrivial for nonstationary backgrounds where artifacts may be introduced by the speaker, the recording environment and the transmission system. The technique for endpoint detection we have used locates the speech boundaries prior to the recognition phase of the system; the algorithm follows [2], using five energy thresholds to find the real beginning and ending frames of energy pulses.

We have found that this algorithm works properly when the signal to noise ratio is over 12 dBs, ignoring high level spurious noises, and allowing plosive sounds in the words without detecting endpoints during the occlusion.

3.1.2. Feature extraction

The next step in the recognition process is the conversion of the flow of speech samples in a set of vectors of parameters that efficiently represent the acoustic information of the speech signal. Experience shows that

the most useful parameters for speech recognition are related to some property of the short-time power spectrum of the speech signal. We have studied a number of combinations of derived-LPC parameters, including some information about the energy, and we have also tested differential parameters, related with the transitional information of the spectrum.

Fig. 2 shows a more detailed description of the feature extractor block, which includes a pre-process module (preemphasis filter) to spectrally flatten the speech signal and to reduce some undesired effects [1].

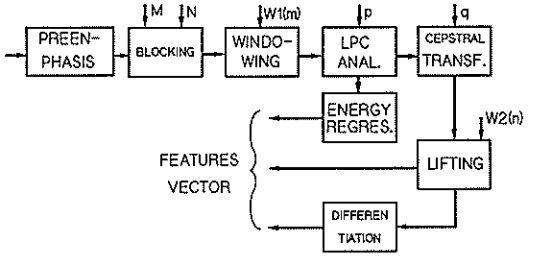


Figure 2. Feature Extraction Block.

The overall system is a block processing model in which a frame of N samples is processed giving a vector of features O_t. Consecutive frames are spaced M (M<N) samples apart [1]. Each N samples frame is windowed (using a Hamming window) to minimize the adverse edges effect.

The next step is to perform an autocorrelation analysis with each frame, and then, from each autocorrelation vector, a vector of 'p' LPC parameters is computed using a Levinson recursion method, and a LPC derived cepstral vector is then computed up to the q^{th} component. This q-coefficient cepstral vector is then weighted by a lifting window.

The final step in the system is to take the time derivative of the sequence of weighted cepstral vectors, obtaining the 'delta cepstrums' [1].

The observation vector O_t used for recognition will be the concatenation of the regression of the energy and some cepstral parameters.

3.1.3. Scores Calculation

As mentioned before, we have built our recognizer with HMM ideas, modeling each vocabulary word as produced by a Markov chain of N_e states.

It is well known that an HMM is defined by the following parameters:
- N_e : number of states of the model.
- A : transition matrix; its elements are a_{ij}, the probability of being in the state i, and to make a transition to the state j. We have tested bi-diagonal and tri-diagonal transition matrixes.
- $f_j(O_t)$, $0<j<N_e$, the p.d.f. that models the outputs from each state. We have used continuous HMMs with mixtures of weighted Gaussians as p.d.f.:

$$f_j(O_t) = \Sigma_i \; c_{ij} \cdot N(m_{ij}, s^2_{ij})$$

so, for each state j we need:
- Vector C_j of $\{c_{ij}, 0<i<NMIX\}$.
- Vector MU_j of $\{m_{ij}, 0<i<NMIX\}$.
- Vector $SIGMA_j$ of $\{s_{ij}, 0<i<NMIX\}$.

We have all that information per each vocabulary word in a file of models, and the recognizer will compute $P(O/WORD_i)$ via the Viterbi algorithm for each model of the vocabulary to give the twelve scores which will be used later.

3.1.4. Decision Rule

The final block in our recognition system modifies the twelve scores computed with the Viterbi algorithm, including durational models in the scoring in a post-processing way, and selects the recognized word as the one with a higher final punctuation.

3.2. Experiments Description

In Section 3.1 we have described the structure of our recognition system without specifying the value of its parameters (e.g. N_e, NMIX, n, m, etc) and the final length and composition of the feature vectors extracted from the speech signal. To select the set of parameters that performs best with our vocabulary, we have carried out the next experiments:

3.2.1. Selection of the Observation vector

The transformation of the waveform into a set of vectors may be seen as a mapping of the utterance into a set of points in a multidimensional parameters space. Different utterances of the same word spoken by different speakers will generate a set of points in the parameter space.The set of parameters that forms the observation vector will be good if the distributions of different words are concentrated at widely different locations in the parameter space. That is, we are interested in a set of parameters which are efficient in representing the speaker-independent information of the acoustic waveform, besides being easy to measure and stable over time. To find such a set of parameters, we have tested:

- Different window analysis length (25, 30, 35 mseg) with different overlap.
- 8,10 and 12^{th} order LPC analysis ('p').
- Extraction of 8 and 10 LPC-cepstrum coefficients (value of 'q').
- Use of the energy regression.

3.2.2. Parameters of the hidden Markov models

The second set of experiments shows the behaviour of the recognizer when using:

- 4, 6 and 8 state HMMs (value of 'N_e').
- Mixtures of 1, 2, 4 and 8 gaussians (NMIX).
- Temporal information modeled as a gaussian.
- Models with bidiagonal and tridiagonal transition matrixes.

The results of those sets of experiments are plotted in the figures 3 to 8. Figures 3, 4, 5 and 6 show the recognizer behaviour when varying the observations vector and the number of states (N_e) of the models:

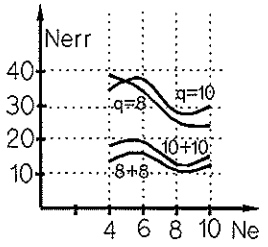


Figure 3

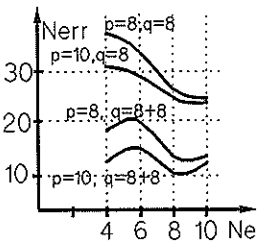


Figure 4

- Fig. 3 outlines the number of errors (N_{err}) versus N_e when the LPC analysis is of 10th order (p=10) and the observations vector contains 8 ceptrums (q=8), 10 cepstrums (q=10), 8 cepstrums plus 8 differential cepstrums (8+8), and 10 cepstrums plus 10 differential cepstrums (10+10).
- Fig 4 shows N_{err} versus N_e for different LPC orders (p=8,10) including or not including transitional information (here q=8 always).

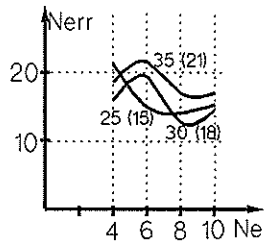


Figure 5

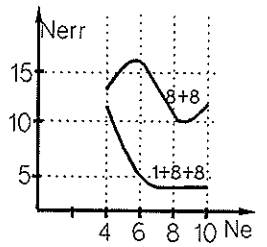


Figure 6

- Fig 5 draws N_{err} versus N_e for different analysis windows: lengths of 25, 30 and 35 mseg with overlaps of 15, 18 and 21 mseg respectively.

- Fig 6 shows the effect of including energy information in the observations vector. Windows are of 30 mseg, p=10, q=8 and differential information is included in the vectors.

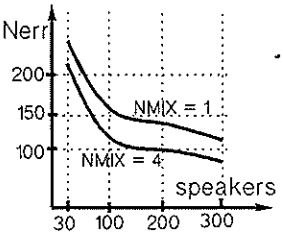


Figure 7

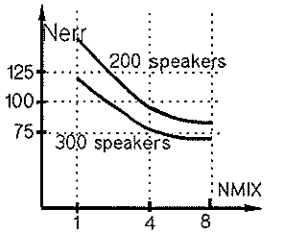


Figure 8

Fig 7 draws N_{err} versus the number of different speakers used in the model training phase for NMIX=1 and NMIX=4. Fig 8 shows N_{err} versus NMIX when the training has been done with 200 and 300 speakers.

4. CONCLUSIONS

With all our results in mind, the final structure of our system is:
- Feature vector composed by the concatenation of the energy regression, 8 cepstrums and 8 delta cepstrums (p=10, q=8).
- Analysis window of 30 mseg with overlap of 18 mseg.
- Each Markov model has 8 states (N_e=8).
- The pdf that models the outputs from each state is a mixture of 8 gaussians (NMIX=8)

The final error rate measured in our laboratory is over 4% when the training is performed with 300 speakers and the testing is done with men, women and children (we have found the error rate for children is about three times the error rate for men).

REFERENCES

[1]- Lawrence R. Rabiner : "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, Vol 77, No 2.
[2]- Lori F. Lamel et al.: "An Improved Endpoint Detector for Isolated Word Recognition", ASSP, Vol 29, No 4.
[3]- Sadaoki Furui : "A VQ-Based Preprocessor Ussing Cepstral Dynamic Features for Speaker-Independent Large Vocabulary Word Recognition", ASSP, Vol 36, No 7.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ISOLATED WORD RECOGNITION IN THE MOBILE-RADIO SYSTEM: EXPERIMENTS AND RESULTS *

L. Fissore, M. Codogno, G. Pirani

CSELT – Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - 10148 Torino, Italy

An isolated-word speech recognition system is presented with satisfactory performance in a noisy environment. In this sense it is effective for the application of voice dialling in the mobile-radio system which is operating in the car.
Experiments and results are reported, which are relevant to a speech database which has been collected directly in the application field.

1 Introduction

A challenging application of automatic speech recognition is to allow the driver to handle the mobile-radio terminal of the car through his own voice. Particularly, it is very useful to be able to dial the desired telephone number by voice, without taking one's own attention off the driving.
Automatic speech recognition in the car environment is a hard task owing to the noise which affects the signal to be recognized. Furthermore, as there is a great variety of noise sources (engine at different speeds, fans, jerks due to the road surface, external noises, etc.) it is quite difficult to give its characterization in order to devise a technique for "cleaning" the signal.
In this paper an Automatic Speech Recognizer (ASR) is described which recognizes isolated utterances from a vocabulary of 27 words, after a proper training of the system. The words that have been chosen for the experiment (10 digits and 17 commands, as reported in Table 1) should provide the user with a flexible means of voice dialling.
In order to design a speech recognition system in a noisy environment one or more of the following items are to be considered:

- choice of a suitable transducer (noise-cancelling microphone, microphone array, etc.) [8];
- use of speech enhancement algorithms to "clean" the signal on line (spectral subtraction, adaptive noise cancelling, etc.) [4];
- design of speech recognition algorithms which are robust enough for achieving high performance even in noisy environment [6,9].

As regards the first item a microphone has been chosen which is also suitable for the hands-free communication facility of the mobile-radio service. In particular, a noise cancelling back-electret microphone has been used with a 3dB cutoff at 10kHz.

*This work has been partially supported by EEC Esprit II project 2101 – ARS.

ZERO (zero)	NUMERO (number)
UNO (one)	MEMORIA (memory)
DUE (two)	AGENDA (agenda)
TRE (three)	VERIFICA (verify)
QUATTRO (four)	ESEGUI (call)
CINQUE (five)	SCRIVI (store)
SEI (six)	MEMORIZZA (store)
SETTE (seven)	CASA (home)
OTTO (eight)	UFFICIO (office)
NOVE (nine)	LAVORO (office)
COMPONI (dial)	SEGRETARIA (secretary)
SELEZIONA (dial)	AMICO (friend)
RICHIAMA (recall)	EMERGENZA (emergency)
CANCELLA (cancel)	

Table 1: Application vocabulary

As concerns the signal pre-processing for enhancing speech, the techniques proposed in literature have been already applied to speech coding, but they rather give benefit to the subjective quality of the signal than improve the performance of the ASR.
Therefore we have chosen to focus the experiments on the third item, i.e., on developing training and recognition algorithms which are particularly effective in the car-noise environment.
The general architecture of the ASR is described in Fig. 1.

2 Database collection

The speech database has been collected directly in the operative environment of the car.
In order to have a significant size, 4 speakers have been employed, 2 males (FC and GM) and 2 females (LC and LV). Each of them, seated in the passenger's seat, has uttered 10 repetitions of the 27 words of the vocabulary of Table 1 in three different conditions:

A) car off;

B) car travelling at approximately 50 km/h;

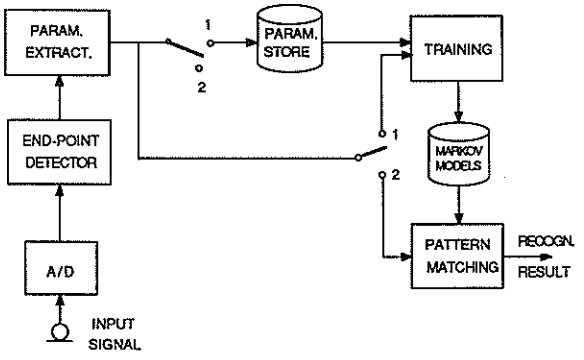


Figure 1: General architecture of the Automatic Speech Recognizer: (*1.Training mode; 2. Recognition mode.)*

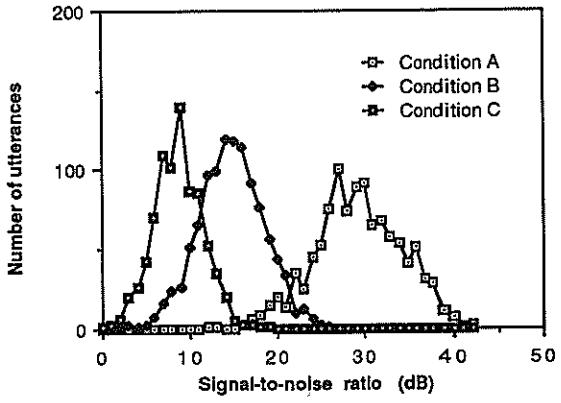


Figure 2: SNR value distribution for the uttered words.

C) car travelling at approximately 80 km/h;

The used car was a TIPO 1600 DGT FIAT and the utterances have been recorded on a NAGRA IV-S recorder through the above-mentioned back-electret microphone that was centered on the passenger's sunvisor.
Therefore the total number of the recorded utterances was 3240. These recordings have been filtered at 5 kHz and acquired with a 12 bit linear A/D converter with a sampling rate of 12 kHz.
The stored utterances have been segmented manually in order not to include at their beginning and ending more than 150 ms. of "silence"[1]. The signal-to-noise ratio (SNR) for the stored utterances has been computed for the three different conditions as

$$\mathrm{SNR} = \frac{E[s^2] - E[n^2]}{E[n^2]}, \tag{1}$$

where $E[s^2]$ is relevant to the segmented utterance and $E[n^2]$ is relevant to the segments of "silence" surrounding the utterance itself.

3 Parameter extraction

In order to obtain the parameters for the training and recognition algorithms an FFT is performed each 10 ms (128 samples) over the speech weighted by a 256-sample Hamming window with a 50% overlapping. A pre-emphasis is also performed through a filter $H(z) = 1 - 0.95z^{-1}$. The signal bandwidth is subdivided into 18 sub-bands not equally spaced according to the Mel scale [1]. For each window the signal energy relevant to the $i-th$ sub-band is computed as

$$E_i = \sum_{j=L_i}^{S_i} |\, X_F(j) \,|^2 \ , \ i = 1, \ldots, N_F, \tag{2}$$

where N_F is the number of sub-bands, L_i ed S_i are the lower and upper edge of the $i-th$ sub-band, respectively and $X_F(j)$ is the $j-th$ sample of the FFT of the input signal, $x(k)$, weighted with the Hamming window. For the experiments that will be reported, $N_F = 18$ has been chosen.

Finally the cepstral coefficients, C_i are computed as

$$C_i = \sum_{j=1}^{N_F} Log(E_j) \cdot \cos i(j - \frac{1}{2})\frac{\pi}{N_F} \ , \quad i = 1, \ldots, N_C \leq N_F - 1. \tag{3}$$

In the reported experiments, $N_C = 12$ has been chosen. For each frame, the logarithm of the energy is also computed as $\overline{E} = Log(\sum_{j=1}^{N_F} E_j)$. Furthermore, an approximation of the temporal derivative of the cepstral coefficients is evaluated by means of a first-order orthogonal polynomial:

$$\Delta C_i(l) = \sum_{k=-K}^{K} k \cdot C_i(l-k) \ , \ i = 1, \ldots, N_C, \tag{4}$$

where l is the examined frame. In the reported experiments $K = 2$ was used. In this way the speech is represented by a sequence of vectors $\mathbf{S}(l) = \{\mathbf{C}(l), \mathbf{\Delta C}(l), \overline{E}(l)\}$.

4 The recognition system

The developed algorithms have been tested after training the system with the speech data relevant to condition A. In all the experiments reported the values of $\mathbf{\Delta C}(l)$ are multiplied by the constant 0.398 in order to take into account their higher dynamic range.

Training

The vocabulary words are represented by Hidden Markov Models (HMM) with N states without skips. Their emission probabilities are described by Gaussian densities. The "silence" is represented by a Markov model with 1 state whose emission probability is a mixture of 4 Gaussian densities. The Markov model of a word is described by the following parameters:

[1] Quotation marks are used to indicate that silence is more properly background noise

1. number of states, N (in this case, $N = 8$);
2. transition probability matrix, $\mathbf{A} := \{a_{ij}\}, i = 1, 2, \ldots, N, j = i, i+1$, which gives the probability of transition from state i to state j ;
3. matrices $\mathbf{M} := \{m_{ij}\}$ and $\mathbf{\Sigma} := \{\sigma_{ij}\}, i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, N_C$, whose entries are the means and the variances of the Gaussian densities which approximate the actual densities of the parameters S.

The "silence" model is described by the matrices $\mathbf{M} := \{m_{ik}\}, \mathbf{\Sigma} := \{\sigma_{ik}\}, i = 1, 2, \ldots, N_C, k = 1, 2, \ldots, N_G$, and by the vector $\mathbf{\Lambda} := \{\lambda_k\}, k = 1, 2, \ldots, N_G$, where N_G is the number of Gaussians of the mixtures (in this case $N_G = 4$):

$$p(\mathbf{S}) = \sum_{k=1}^{N_G} \lambda_k N(m_{ik}, \sigma_{ik}, \mathbf{S}), \tag{5}$$

where $\lambda_k \geq 0, \sum_{k=1}^{N_G} \lambda_k = 1$, and $N(m_{ik}, \sigma_{ik}, \mathbf{S})$ is the multivariate Gaussian density.
To train the Markov models of the words the forward-backward procedure is employed, extended to the continuous densities Markov models [3], using the utterances of condition A.
A single Markov model of the "silence" has been obtained using approximately 20 minutes of signal recorded in the three different conditions (A,B,C). At first, the "silence" frames have been divided into four clusters by means of the LBG algorithm [5]. The mean and variance vectors, computed for each cluster, are those of the Gaussian densities of the mixture of equation 5. The weights λ_k are computed as the ratio between the number of frames belonging to the $k-th$ cluster and the total number of frames.
The transition probabilities are fixed as $a_{11} = a_{1A} = 0.5$. As the training is performed on the clean signal (condition A) an improvement has been experienced, replacing the emission probabilities of the low-energy states (mainly relevant to the occlusion interval of plosives) with those of the model of "silence".
In order to cope with the problems of statistical insufficiency of the training data the "grand variance" technique (GV) [7] has been adopted. It consists in evaluating *a priori* the variances of the components of vector S using all the training data and fixing to these values the variances of all the states of all the models.

Recognition

The recognition of the unknown input word is performed by searching for the model that generates the examined utterance with the highest likelihood score. This search is carried out using the Viterbi algorithm to find the best path in the trellis of the states. In order to improve the algorithm efficiency a beam-search strategy is introduced, by pruning the unpromising paths when their likelihood score goes below a given threshold with respect to the best path.
Two different types of models have been tested: the former uses only the Markov models of the words, the latter "augments" them linking one optional model of "silence" at their beginning and ending.

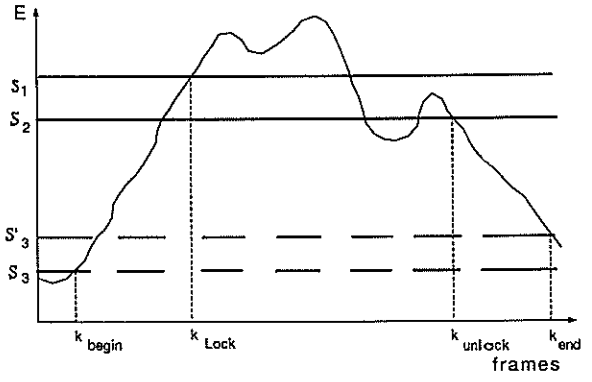


Figure 3: Scheme of the EPD operation

5 Automatic end-point detection

The developed end-point detection (EPD) algorithm relies only upon the values of the energy of the input speech, $E = \sum_{j=1}^{N_F} E_j$. The values of E(k) are stored in a buffer of 40 locations in order to cope also with the long silences before bursts in geminate plosives (like in OTTO, SETTE, etc.). The actual beginning and ending points are searched for after the activation of the "lock" and "unlock" flags, respectively. A detailed description of the algorithm is contained in [2].
As described in Fig. 3, EPD exploits four adaptive thresholds: S_1, "lock" threshold, S_2, "unlock" threshold, S_3 e S_3', refinement thresholds for the beginning and the ending of the word. The threshold adaptation is performed on the basis of the estimate of the noise power, P_N, according to the following equations

$$\begin{aligned} S_1 &= \alpha(P_N) \cdot P_N, \\ S_2 &= \delta S_1, \\ S_3 &= \gamma P_N, \\ S_3' &= \gamma' P_N, \end{aligned} \tag{6}$$

where δ, γ, and γ' are properly chosen constants and α is a function of P_N as shown in Fig. 4.
When applied to the collected spech data base the EPD algorithm detected correctly all the 3240 utterances without any insertion or deletion.

6 Results and comparisons

The experiments have been carried out with reference to three different schemes.

1. A manual end-point is performed to extract the utterance to be recognized.
2. The automatic EPD described in Section 5 is used as described in Fig. 1.
3. No EPD is used and the word to be recognized is spotted and extracted from the background noise.

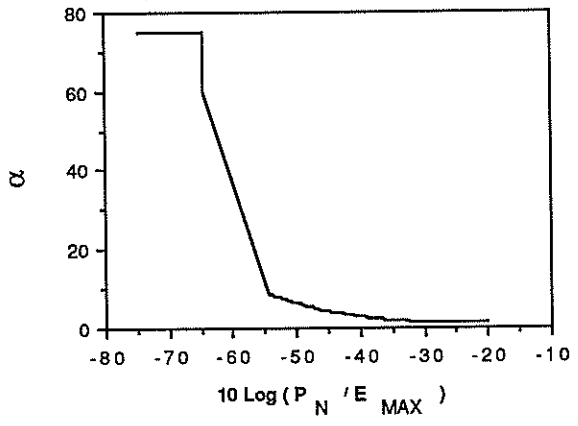


Figure 4: α vs. P_N

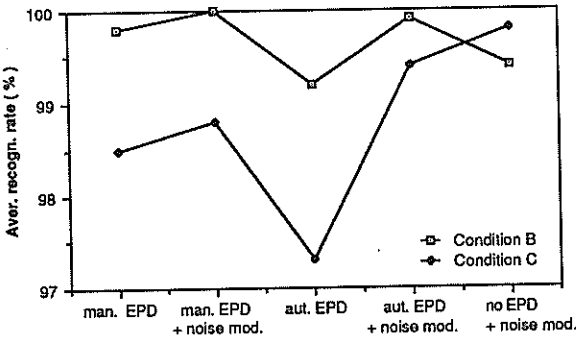


Figure 5: Recognizer perfomance (10 repetitions for training)

Of course only schemes 2 and 3 refer to real-life operation, while the "ideal" operation mode of scheme 1 gives only a reference performance value. In particular scheme 3 represents the mode where the user utters only a word after a prompt and within a timeout interval.
The performance of the ASR are sketched in Fig. 5 in terms of average correct recognition percentage. The results reported in this figure show that high performance is reachable even if the training of the system is carried out using the "clean" data of condition A (car off) that are quite different from the operating test conditions (car travelling). This is an outstanding advantage because condition A is the most comfortable for the user to train the system.
It is also worth underlining the substantial improvement obtained through the addition of the average noise model for all the operating conditions.

7 Conclusions

An isolated word recognition system has been presented which shows satisfactory performance in the noisy environment of the car. The experiments have been carried out on a speech database of more than 3000 utterances and point out that correct recognition rate higher than 98% can be achieved under heavy noise conditions, even if the training is performed in a relatively quiet environment.
It has also been shown that both the system using a properly designed automatic end-point detector and the system relying upon a HMM-based word spotting yield results perfectly comparable with those relevant to the manually end-pointed database.
The forthcoming research efforts will be devoted to the design of a faster training procedure and to the development of a speaker independent system with the same features of the one described in the paper.

Acknowledgments

The authors would like to thank Mr. E. Bracalente for his contribution to the design and experimentation of the automatic end-point detector algorithm.

References

[1] K.H. Davis and P. Mermelstein. **Comparison of parametric representation for monosyllabic word recognition in continously spoken sentences.** *IEEE Trans. Acoust., Speech and Signal Processing*, 28:357–366, 1980.

[2] L. Fissore, M. Codogno, and G. Pirani. **Riconoscimento di parole isolate nel sistema radiomobile: esperimenti e risultati.** *Rapp. Tecn. CSELT*, 1989.

[3] B. H. Juang and L. R. Rabiner. **Mixture autoregressive Hidden Markov Models for speech signals.** *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-33(6):1404–1413, 1985.

[4] J. S. Lim. **Speech Enhancement.** Prentice-Hall, 1983.

[5] Y. Linde, A. Buzo, and R.M. Gray. **An Algorithm for Vector Quantizer Design.** *IEEE Trans. on Communications*, 28:88–95, 1980.

[6] R.P. Lippmann, E.A. Martin, and D.B. Paul. **Multistyle training for robust isolated-word speech recognition.** In *Proc. of the ICASSP*, 1987, sect. 17.4.

[7] E.A. Martin, R.P. Lippmann, and D.B. Paul. **Two-stage discriminant analysis for improved isolated word recognition.** In *Proc. of the ICASSP*, 1987, sect. 17.5.

[8] J. R. Sank. **Microphones.** *J. Audio Eng. Soc.*, 33:514–547, 1985.

[9] J. G. Wilpon and L. R. Rabiner. **Application of hidden Markov models to automatic speech endpoint detection.** *Computer Speech and Language*, 2:321–341, 1987.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

SIMPLIFICATION AND IMPROVEMENT OF THE BINARY CODED EXCITED LINEAR PREDICTION (BCELP) FOR SPEECH CODING

R. BOITE. H. LEICH and GAO YANG(*)

Faculté Polytechnique de Mons, Belgium
(*) on leave from Nanjing Aeronautical University, R.P. China

In a recently proposed CELP method for low-bit rate speech coding with high quality [1], the non-zero amplitudes of the codewords are +1 or-1, regularly spaced in each word; a scalar gain is also computed for each word: this method, called BCELP, is theoretically equivalent to the original CELP, with less computation, moreover it does not require the storage of an excitation codebook, and it is robuste against channel errors. The objective of the present paper is to still reduce much the computation complexity while improving the quality of the coding, the "SBCELP" (Simplified BCELP).

1.INTRODUCTION:THE SIMPLFIED STRUCTURE

The original stucture of a CELP coder is well known [4], and it does not need to be recalled. If the polynomials A(z) and P(z) correspond respectively to the short term and the long term predictors (LTP):

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i} \qquad (1)$$

$$P(z) = 1 - \beta z^{-P} \qquad (2)$$

and if $W(z)=A(z)/A(z/\alpha)$ is the weighted filter, the modified structure is given at figure 1. s(n) is the speech signal, E is the error to be minimized,and G is the associated gain.

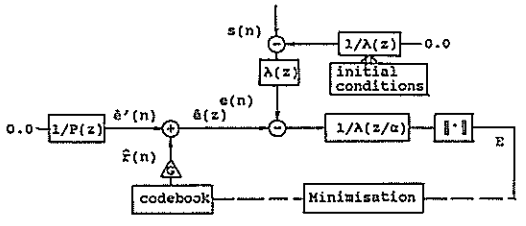


Fig.1 The structure of synthsis-analysis

Initial conditions for the short term synthesis filter 1/A(z) should be introduced as shown, with zero initial conditions for the inverse filter A(z):the weighted filter is combined with the short term synthsis into the shaping filtre 1/A(z/α);and the long term filter 1/P(z) can be taken out of the closed loop if the long term delay P is greater than the excitation frame length L (P≥L);in this way the complexity of the computations can be reduced by a factor 3 [2].

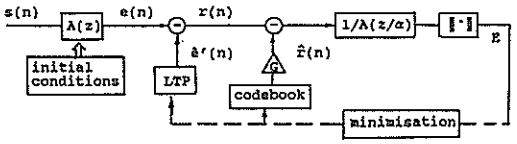
The following table will define the parameters and their chosen values in the present work.

TABLE 1 The chosen parameters

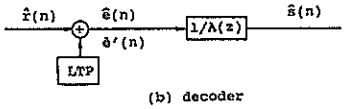
Sampling frequency:	f_s= 8 kHz
Short-term prediction:	Autocorrelation
Order:	p=10
Long term prediction:	closed loop design
Hamming window:	N_r=220
Analysis frame shift:	N=160
Excitation frame size:	L=40
Non-zero ampl/frame:	Q=10
Weight factor:	α=0.85
Codebook size:	1024
No pre-emphasis	

2. THE SIMPLE METHOD TO ELIMINATE THE INFLUENCE OF THE INITIAL CONDITIONS FOR THE SHORT TERM SYNTHESIS FILTER

The structure of figure 1 can be redrawn as given at figure 2. The influence of the initial conditions for the short term synthesis filter can be efficiently eliminated with more simple method: the initial conditions $\hat{s}(0)$,



(a) coder



(b) decoder

Fig.2 The redrawn structure of synthesis-analysisThe

$\hat{s}(-1),\ldots,\hat{s}(-p+1)$, are directly introduced into the inverse filter A(z) for obtaining the same reference excitation e(n).

3. COMPUTATION OF THE LONG TERM PREDICTOR

Generally, the long term parameters are first determined before computing the excitation parameters. The initial conditions: $\hat{e}(0),\hat{e}(-1),\ldots,\hat{e}(-P+1)$ of the long term filter with zero input are used to estimate the periodical part $\hat{e}'(n)$ of the residue e(n). The long term delay P and the coefficient β are determined by the so-called closed-loop method with the minimisation of the expression:

$$E_1 = \sum_{n=1}^{L} [(e(n)-\hat{e}'(n)*h'(n)]^2 \qquad (3)$$

where h'(n) is the impulse response of the shaping filter 1/A(z/α) and $\hat{e}'(n)$ is given by:

$$\hat{e}'(n) = \beta\hat{e}(n-P),\ (n=1,2,\ldots,L) \qquad (4)$$

The value of P is limited in the range (L,L+63). If the speech pitch period is smaller than the excitation frame length L, the optimal long term delay P will be equal to twice as much as the pitch period. Sometimes, it is better to use two pitch periods for the long term prediction.

The structure of figure 2 can be used as reference for the classical CELP, for the BCELP [1] and for its simplified version SBCELP proposed in this paper.

4. THE CODEBOOK FOR SBCELP

The codebook for SBCELP is designed as shown at figure 3. with a possible shift of two samples defined by the index k: the non-zero amplitudes b(i), (i=1,2,..,Q),of the codewords are +1 or -1,regularly spaced in each word; on the other hand,as the gain G is signed, the first amplitude b(1) can be fixed to +1 so sparing one bit. In reality the codes are computed during the encoding process so that the codebook does not need to be stored; moreover, a binary code is robust against the channel errors errors [1].

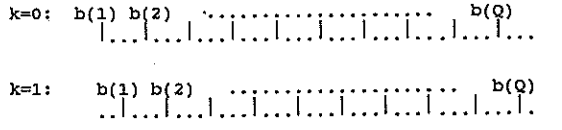


Fig.3 The binary codebook for SBCELP

5. COMPUTATION OF THE OPTIMAL EXCITATION SEQUENCE

The optimal excitation sequence is determined by a non exhaustive method in three steps. During the first step, continuous values $u^k(i),(i=1,2,\ldots,Q)$, (k=0,1),of the non-zero amplitudes of the excitation sequence are computed in a way similar to that used in the RPE algorithm [3]; during the second step they are approximated by binary values b(i). In the last step these binary values are improved using a very simple rule.

The position matrix $M_k=[m_{i,j}]$ of order (Q*L) with:

$m_{i,j}=1$, if $j=4i+(2k-3)$; $i=1,2,\ldots,Q$; $j=1,2,\ldots,L$; $k=0,1$
$=0$, otherwise

and the vector $U^k=[u^k(1),u^k(2),\ldots,u^k(Q)]$ are defined. The excitation sequence $\hat{P}^k=[\hat{P}(1),\hat{P}(2),\ldots,\hat{P}(L)]$ can be written as:

$$\hat{P}^k = U^k M_k \qquad (5)$$

If h(n) is the impulse response of the filter 1/A(z), that of the shaping filter h'(n) is given by $\alpha^n h(n)$, so that one has $h'(n)\approx 0$ for $n>L=40$ and the

related matrix can be approximated by the TOEPLITZ matrix of order [L*2L]:

$$H = \begin{bmatrix} h'(0) & h'(1) & \ldots & h'(L-1) & 0 & 0 & & 0 \\ 0 & h'(0) & h'(1) & \ldots & h'(L-1) & 0 & & 0 \\ & & & & & & & \\ 0 & & & & 0 & h'(0) & h'(1) & \ldots & h'(L-1) \end{bmatrix}$$

If the vector e^k with 2L components is the response of the shaping filter to excitation $(r-\hat{P}^k)$ of length L, where $r=[r(1),r(2),\ldots,r(L)]$ is the reference excitation vector, one has:

$$\begin{aligned} e^k &= (r-\hat{P}^k)H \\ &= rH - U^kM_kH \\ &= rH - U^kH_k \end{aligned} \qquad (6)$$

For a given value of k, the optimal amplitudes U^k are determined by minimizing the energy $E^k = e^ke^{kt}$ to E^k_{min}. As suggested in [3], the apparent complexity of the computations can be greatly reduced. Let:

$$R(i) = \sum_{n=0}^{L-1-i} h'(n)h'(n+i) \qquad (7)$$

and observe that $R(i)\ll R(0)$ for $i\geq 4$, so that one has $H_kH_k^t \approx R(0)I$; then:

$$\begin{aligned} u^k(i) &\approx R^{-1}(0) \sum_{j=-4i-2k+4}^{L-4i-2k+3} R(|j|)\, r[(4i+2k-3)+j], \\ &\qquad i=1,2,..,Q;\ k=0,1 \\ &\approx R^{-1}(0)\, v^k(i), \end{aligned} \qquad (8)$$

where

$$v^k(i) = \sum_{j=\max\{-10,-4i-2k+4\}}^{\min\{10,L-4i-2k+3\}} R(|j|)\, r[(4i+2k-3)+j],$$

$$i=1,2,..,Q;\ k=0,1; \qquad (9)$$

only 20 terms in the sum are used. Finally, as one has:

$$\begin{aligned} E^k_{min} &\approx rHH^tr^t - R(0)U^kU^{kt} \\ &\approx rHH^tr^t - R^{-1}(0)\, V^kV^{kt} \end{aligned} \qquad (10)$$

the optimal value of k is chosen to maximize the expression:

$$V^kV^{kt} = \sum_{i=1}^{Q} |v^k(i)|^2, \qquad (11)$$

The second step consists in the determination of the binary sequence:

$$\begin{aligned} b(i) &= \mathrm{sgn}[u^k(1)]\ \mathrm{sgn}[u^k(i)], \\ &= \mathrm{sgn}[v^k(1)]\ \mathrm{sgn}[v^k(i)], \\ &\qquad i=1,2,..,Q \end{aligned} \qquad (12)$$

Remember that b(1)=+1 and that a signed gain G is associated to the sequence.

During the last step, the sign of each binary pulse is changed one after the other, while controlling the resultant modification or the criterium; the modified sign is kept each times the criterium is lowered.

This process requires Q-1 computations of the criterium; however if the following expression:

$$\begin{aligned} E &= \sum_{n=1}^{L} [(r(n)-\hat{P}(n))*h'(n)]^2 \\ &= ||(r-\hat{P})H'|| \\ &= ||(P-GY)|| \end{aligned} \qquad (13)$$

where

$$H' = \begin{bmatrix} h'(0) & h'(1) & \ldots & h'(L-1) \\ 0 & h'(0) & \ldots & h'(L-2) \\ \cdot & & & \\ \cdot & & & \\ 0 & & \ldots & h'(0) \end{bmatrix} \qquad (14)$$

$P = rH'$, $\hat{P} = GbM_k$, $Y = bM_kH'$. $b = [b(1),b(2),\ldots,b(Q)]$,

is minimized with respect to the gain G, the optimal gain is found to be:

$$G = PY^t/YY^t \qquad (15)$$

After subsitution in E , one obtains:

$$\begin{aligned} E &= PP^t - (PY^t)^2/YY^t \\ &= PP^t - GPY^t \\ &= PP^t - D \end{aligned} \qquad (16)$$

and minimisation of E is equivalent to maximisation of D. Each modification of D does not requires recomputation of Y, which can be modified in this simple way:

$$Y=Y+\{0,..,-2\mathrm{sign}[b(i)],..,0\}M_kH', \qquad (17)$$

as one has:

$$b=b+\{0..,-2\mathrm{sign}[b(i)],..,0\}, \qquad (18)$$

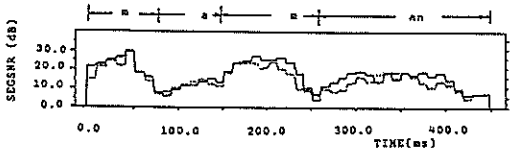
The new sign of the binary pulse is maintained if corresponding D is greater. The final excitation sequence is that used for the proposed SBCELP.

6. SIMULATION

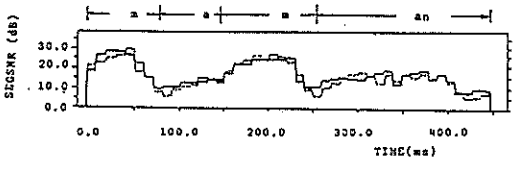
The french phrase "Attention,appel à tous les voyageurs, le train aura pour

terminus la station Schumann" containing 1600 excitation frames of length L=40 has been used for a first comparison of BCELP and SBCELP.

This phrase has been coded according the upper described SBCELP and also according the method with exhaustive search. It was observed that 93,5 % of the excitation vectors for SBCELP are identical to those obtained by the method of exhaustive search and that 4.2 % of them differ for one pulse only.The computation complexity for determining the excitation sequence with SBCELP has been reduced by a factor 240 comparatively to the classical CELP.



(a) with CELP (dashed line) and BCELP (solid line)



(b) with BCELP (dashed line) and SBCELP (solid line)

Fig.4 Comparison of the objective segmental SNR (dB) (10 ms frame) for CELP, BCELP and SBCELP.

TABLE 2 Comparison of the objective segmental SNR (dB) averaged on 450 ms of the figure 4 and the number of operations.

Méthode	SNR(dB)	Multi./sec.	Add./sec.
CELP	11.43	90,192 K	98,392 K
BCELP	13.46	17,258 K	57,344 K
SBCELP	15.39	395 K	397 K

An objective comparison was done by coding the french word "maman" in CELP, BCELP and SBCELP. Figure 4 shows the segmental SNR (10ms frame). Table 2 indicates for the three methods the SNR averaged on 450ms and the number of operations per second for determining the index and the gain of the excitation vectors. As seen in table 2,the SNR with SBCELP is a little greater than the others, because there is one possible shift in the codebook for SBCELP. The subjective test shows also the similar results.

CONCLUSION

The proposed method SBCELP compares favourably with the original CELP and BCELP: the computation complexity is considerablely simplified, improving a little the synthesis quality,while the advantages of BCELP are retained.

REFERENCES

[1]. R. A .SALAMI,"Binary Code Excited Linear Prediction (BCELP):New Approach to CELP Coding of Speech Without Codebooks", ELECTRONICS LETTERS, 16th March 1989, Vol.25, No.6.

[2]. P. SIEBERT, H. REININGER, and D. WOLF, "The Efficiency of Long-Term Prediction in CELP Schemes", PROCEEDINGS OF SPEECH'88, 7th FASE Symposium,Edinburgh, P.1033.

[3]. P. KROON, E. F. DEPRETTERE, and R. J. SLUYTER, "Regular-Pulse Excitation --- A Novel Approach to Effective and Efficient Multipulse Coding of Speech", IEEE Trans. on ASSP, Oct.1986, Vol.ASSP-34, No.5, P.1054.

[4] Manfred R.SCHROEDER,and B. S. Atal, " Code-Excited Linear Prediction (CELP):High-Quality Speech at Very Low Bit Rates", IEEE,ICASSP'85,25.1.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

HIGH QUALITY SPEECH CODING AT 4.8 KB/S USING MULTI-GRID CELP CODERS

U. Kipper, H. Reininger, and D. Wolf

Institut für Angewandte Physik der Universität Frankfurt a. M.,
D-6000 Frankfurt a. M., Robert-Mayer-Straße 2-4, FRG

In this paper we propose a new kind of excitation sequence for CELP speech coding algorithms called Multi-Grid-Excitation (MGE). MGE consists of multiple pulse grids each with a different number of pulses. Applying strategies used in Multi-Pulse-Linear-Predictive-Coding, MGE offers the possibility to calculate an adaptive excitation together with an efficient way for encoding the pulse locations. The amplitudes of the pulses within a grid are efficiently quantized using vector quantization. A CELP scheme using MGE yields a speech quality at 4.8 kb/s which is comparable to a CELP scheme with 8 kb/s using stochastic excitation.

1. INTRODUCTION

CELP is a closed-loop "analysis by synthesis" LPC system as depicted in Figure 1. The innovation sequence i(n) produced by an excitation generator is filtered by F^{-1}, a combination of adaptive inverse prediction error filters, in order to synthesize a speech signal $\hat{s}(n)$. Minimizing the power E of the difference d(n) between the actual speech signal s(n) and the synthesized speech signal $\hat{s}(n)$ -either by iteratively improving the excitation sequence or by searching over all possible excitation sequences- leads to the optimum excitation for the speech signal s(n). This excitation is encoded and transmitted to the receiver together with the parameters specifying F^{-1}. Provided that no transmission errors occur, the speech signal $\hat{s}(n)$ can be reconstructed in the decoder.

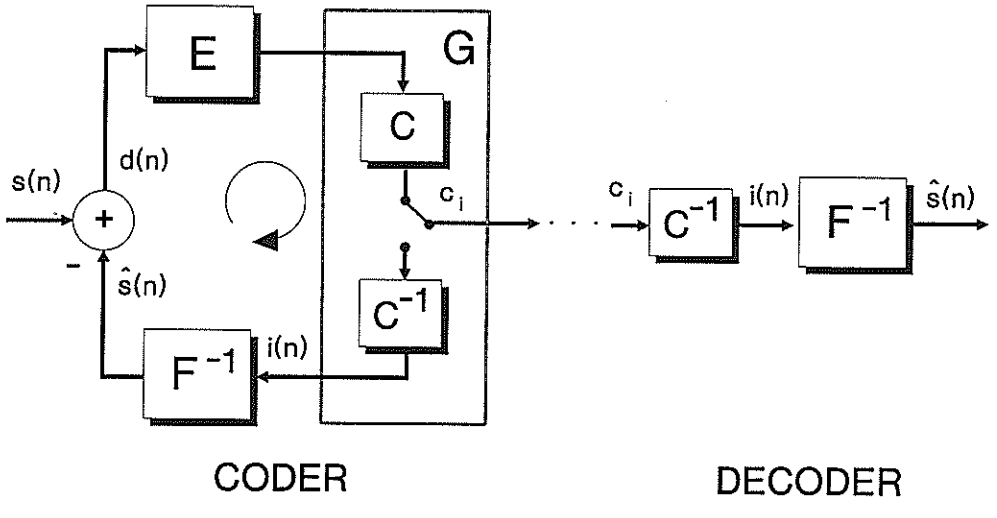


Figure 1: Structure of a closed-loop LPC-system

Multi-Pulse-Linear-Predictive-Coding (MPLPC) and CELP coding are two different methods for generating excitation sequences. In a MPLPC scheme a pulse generator produces a sequence of pulses with variable pulse distances and pulse amplitudes which are optimized in the encoding procedure [2,5]. The transmission of such an excitation sequence requires a high data rate. Thus, a good speech quality with MPLPC codecs can only be achieved at data rates above 10 kb/s. The conventional CELP scheme uses excitation sequences drawn from a small set of vectors, consisting of random numbers, stored in a codebook in combination with a gain factor [1]. The indices of the optimum codebook vectors and the quantized gain factors can be transmitted at a much lower transmission rate than the MPLPC excitation parameters. Therefore, with CELP it becomes possible to encode speech with a high quality at rates of about 8 kb/s. At a rate of 4.8 kb/s the resulting speech quality is strongly speaker dependent and for some speakers even not satisfactory. This effect is mainly due to the fixed excitation codebook containing inappropriate excitations for some speakers or waveforms.

Here we present a new kind of excitation sequence, called Multi-Grid-Excitation (MGE), which leads to an improved speech quality. The excitation is adaptive and can efficiently be encoded due to its regularity.

2. MULTI-GRID-EXCITATION

The MGE generator produces excitation grids with different numbers of regularly spaced pulses. The set of innovation grids is classified into NG different grid classes according to the number of pulses in the grids, as shown in Figure 2. Optimum amplitudes for each grid can be calculated by solving a set of linear equations [6]. In order to achieve a constant data rate for the excitation the quantization accuracy of a pulse amplitude must be decreased with increasing number of pulses in a grid.

For encoding a speech frame all possible grids out of one class each with their optimum pulse amplitudes are used to excite the synthesis filters F^{-1}. For each grid class that grid is determined which minimizes the weighted mean

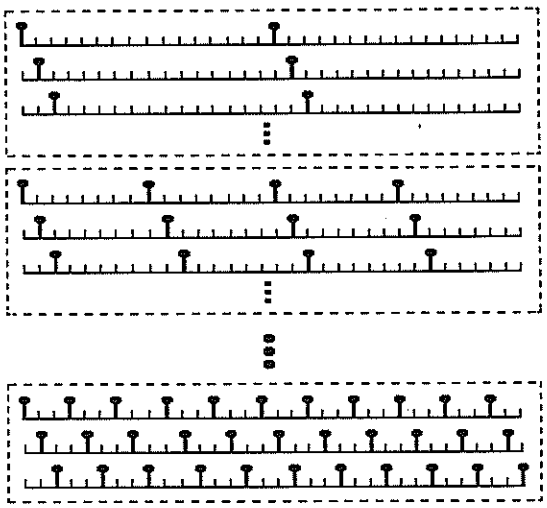


Fig. 2: Grid classes of MGE

squared error between the original speech signal and the output of the synthesis filters. The optimum excitation is found by comparing the squared errors produced by the best grids of all classes.

Thus, an excitation is uniquely defined by the index of the optimum grid together with the number of the grid class and the amplitudes.

3. QUANTIZATION OF PULSE AMPLITUDES

For quantizing the pulse amplitudes with an adequate accuracy at very low bit rates vector quantization is applied. Results obtained with MPLPC in earlier studies have shown that the maximum amplitude in an excitation has the major influence on the resulting speech quality [7]. According to these results a gain shape vector quantizer (GVQ) is used to quantize the amplitudes of the excitation in the following manner. Firstly, the maximum of the optimum

amplitudes is determined and quantized by means of an optimum scalar quantizer. Secondly, the excitation is normalized and rotated such that the maximum comes into the first position. Finally, the (NP-1) pulses, starting at the second element of the normalized and rotated excitation, are quantized using a (NP-1) - dimensional vector quantizer with a quadratic distortion measure. The quantized amplitudes are therefore specified by the index of the optimum codebook vector, the rotation factor R, and the quantization level M of the maximum amplitude.

4. EXPERIMENTS AND RESULTS

Different methods for generating MGE sequences were analyzed in simulation experiments carried out on a digital signal processing system based on a DEC workstation. For the simulations speech data of German male and female speakers were used. All utterances were band-pass limited from 0.3 to 3.4 kHz and sampled at 8 kHz with an amplitude resolution of 16 bit per sample.

To investigate the performance of MGE a codec was configurated for a data rate of 4.8 kb/s. Each speech segment consisting of 192 samples with an overlap of 54 samples at both sides was weighted by a Hanning window. The synthesis filter F^{-1} is a cascade of an inverse prediction error filter 1/A(z) and an inverse long-term prediction error filter 1/P(z). The filter coefficients of the 8-th order short-term predictor 1/A(z) were obtained by using the autocorrelation method [3,4]. The filter coefficients were vector quantized by applying a full search codebook employing the Itakura-Saito distortion measure [8]. The long-term predictor 1/P(z) of order 1 was adapted every 48 samples; the filter coefficient and the pitch period, limited to a range of 48 to 176 sample intervals, were evaluated with the closed-loop method [7] and quantized using optimum scalar 1-bit- and 7-bit-quantizers, respectively. The resulting transmission rates for the predictor parameters are summarized in Table 1.

Table 1: Bit allocation of the predictors

	Adaptation Period ms	Code Rate bit	Data Rate bit/s
Short-Term Predictor	24	10	417
Long-Term Predictor	6	1	167
Pitch Period	6	7	1167

Table 2 shows the excitation configuration of the 4.8 kb/s MGE codec. NG denotes the grid class, NP the number of pulses within a grid, and NG the number of different grids in one class.

Table 2: Excitation configuration

G	NP	NG
1	12	4
2	16	3
3	24	2
4	48	1

With the bit allocation shown in Table 3 the total data rate is 3.0 kb/s for transmitting the excitation sequence. B_G and B_R denote the numbers of bits required for quantizing the index of the grid class and the rotation factor, respectively.

Table 3: Bit allocation for 18 bit per excitation

G	B_G	Id NG	Id M	B_R	Id K_G
1	2	2	5	4	5
2	2	2	5	4	5
3	2	1	5	5	5
4	2	0	5	6	5

In informal listening tests the quality of the evaluated MGE codec with 4.8 kb/s was rated much better than a 4.8 kb/s CELP codec with stochastic excitation. Most listeners judged the quality even better than that of a conventional CELP codec operating at 8 kb/s.

The histogram of the grid classes in Figure 3 demonstrates the preference of the different grid classes. In 50% of all cases excitation sequences with the maximum possible number of pulses achieved the lowest mean squared error.

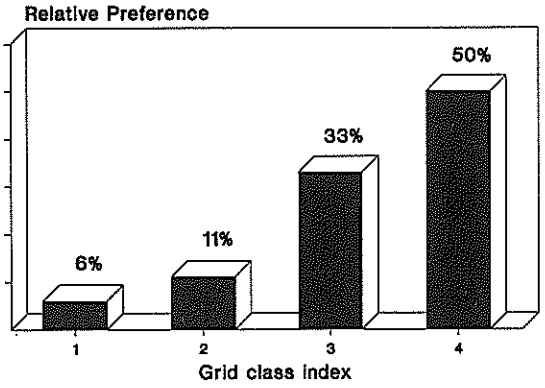


Fig. 3: Histogram of class indices obtained with a 4.8 kb/s MGE codec

In the other cases a more accurate quantization of pulse amplitudes is required. Simulations using only one grid class have confirmed that the intelligibility of the resulting quality deteriorates significantly.

5. CONCLUSION

With the proposed MGE for CELP schemes the performance of a conventional CELP scheme can be drastically improved. At 4.8 kb/s the processed speech using MGE sounds much better than that of the conventional CELP scheme. Furthermore, due to the adaptation of the excitation to the speech signal, the quality is almost independent of the speaker.

REFERENCES

[1] Schröder, M. R., Atal, B. S.: "Code Excited Linar Prediction (CELP): High-quality speech at very low bit rates, Proc. ICASSP 1985, Tampa 1985, pp. 937-940

[2] Atal, B. S., Remde, J. R.: "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates", Proc. ICASSP, Paris 1982, pp. 614-617

[3] Markel, J. D., Gray Jr., A. H.: "Linear Prediction of Speech", Berlin, Heidelberg: Springer, 1976.

[4] Jayant, N. S., Noll, P.: "Digital Coding of Waveforms", Englewood Cliffs, N.J.: Prentice Hall, 1984

[5] Singhal, S., Atal, B. S.: "Improving Performance of Multi-Pulse LPC Coders at Low Bit Rates", Proc. ICASSP, San Diego 1984, pp. 1.3.1-1.3.4

[6] Kipper, U., Reininger, H., Wolf, D.: "On the Performance of MPLPC Schemes at Rates Below 8kb/s", Proc. Speech'88, Edinburgh 1988, pp. 593-597

[7] Kipper, U., Reininger, H., Wolf, D.: "Optimization and Efficient Encoding of Excitation Pulses in Low-Bit-Rate MPLPC schemes", Proc. URSI-ISSSE, Erlangen 1989, pp. 824-827

[8] Siebert, P., Reininger, H., Wolf, D.: "The Efficiency of Long-Term Predictors in CELP Schemes", Proc. Speech'88, Edinburgh 1988, pp. 1033-1038

[9] Siebert, P., Reininger, H., and Wolf, D.: "Quantization of Predictor Parameters and Gain Factors in a CELP Codec for a Transmission Rate of 4.8kbit/s", Proc. EUSIPCO-88, Grenoble 1988, pp. 1027-1030, 1988

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

IMPROVED REGULAR PULSE CELP CODING FOR NARROW BAND SPEECH TRANSMISSION

M. LEVER, C. GRUET and M. DELPRAT

MATRA Communication, Rue J.P. Timbaud, B.P. 26
78392 Bois d'Arcy Cedex, FRANCE

Regular Pulse CELP coding of speech has been recently introduced as an efficient technique for narrow band transmission of speech. The codebook structure together with a convenient error criterion enables a very fast search procedure while maintaining high quality output speech. The paper analyses the error minimization criterion, introducing a generalized autocorrelation formulation that reduces block edge effects by allowing overlapping between blocks. Then a very fast algorithm for closed-loop long term prediction analysis is described and we show that this new method can be extended to a fast joint optimization of long term prediction and excitation parameters. The improved RPCELP coder has been real time implemented at 6 and 8 kbps on a TMS320C25 with an overall complexity around 5 Mips.

1. INTRODUCTION

In narrow band communications, low bit rate coding techniques are of major interest. For a given bit rate, the design of a speech coder is a difficult trade-off between two main features: speech quality and complexity. For general public applications, speech quality must be high and the coder must be robust to adverse transmission conditions (background noise, transmission errors...) [1]. But in mobile telephony, a low complexity speech coder is required for the realization of low cost and low power consumption mobiles.

Code-Excited Linear Prediction (CELP) is an efficient technique for producing good quality speech at low bit rate, but the basic complexity is very high. Several related schemes with a reduced computational load have been proposed [3,4,5] and improved quality CELP coders have also been presented [7,11,13]. However, efforts are still necessary to achieve both low complexity and high quality at low bit rate.

Regular Pulse CELP (RPCELP) coding has been recently introduced as an efficient technique for narrow band speech transmission [5,6]. It takes advantage of the particular codebook structure together with a convenient error criterion to achieve a very low complexity while maintaining good quality reconstructed speech.

The paper presents the design and implementation of an improved RPCELP coder. Section 2 gives a brief overview of the RPCELP technique. Then we focus on the error minimization criterion, introducing a generalized autocorrelation formulation that reduces block edge effects by allowing overlapping between blocks. Section 4 presents a new algorithm for long term prediction analysis that enables a very fast computation of parameters in closed-loop. We also discuss complexity reduction methods for the joint optimization of long term prediction and excitation parameters. Finally the paper describes the design and real time implementation of improved RPCELP coders at 6 and 8 kbps.

2. REGULAR PULSE CELP CODING

In CELP coding, synthetic speech is produced by filtering succesive sequences c_k, scaled by a gain factor G, through long and short term predictors, respectively 1/B(z) and 1/A(z). For each block, the optimum vector is selected from a codebook of waveforms using an analysis by synthesis procedure with a perceptual error criterion. The perceptual weighting filter is $W(z) = A(z)/A(z/\gamma)$, with γ around 0.8.

The codebook search comes down to minimize the weighted error signal energy, expressed as

$$E(k) = \| H.r - G.H.c_k \|^2 \quad (1)$$

where r represents the residual signal with the contribution from past excitations subtracted and H is the impulse response matrix of the weighted synthesis filter $1/A(z/\gamma)$. The basic analysis process is quite complex since each codeword c_k must be filtered through $1/A(z/\gamma)$.

The RPCELP technique relies on a modified error criterion together with the regular pulse codebook structure. First, we use an autocorrelation approach [6,8,10,11] for the minimization of E(k), instead of the traditional covariance approach. The best codeword is now selected by maximizing

$$P_w(k) = r^t.R.c_k / \| H.c_k \| \quad (2)$$

where $R = H^t.H$ is a symetrical Toeplitz matrix whose i^{th} diagonal term is the autocorrelation coefficient R_{i-1} of the impulse response of $1/A(z/\gamma)$. The signal $y = R.r$ can be efficiently computed as the result of a particular filtering operation [5,8]. Moreover, we use a different perceptual weighting filter of the form

$$W'(z) = A(z) / C(z/\gamma) \quad (3)$$

where 1/C(z) is an average short term speech predictor with time-invariant coefficients [5,8], so the matrix R also becomes time-invariant. The above modifications of the error criterion have been analysed in [6] in terms of quality and complexity. Some recent studies on this point are reported in next section.

The Regular Pulse (RP) codebook [5,6] is constituted of sequences with q equidistant pulses separated by D-1 zeros. The first pulse (initial phase p) is at one of the locations 0 to D-1. A "binary" RP codebook, built from the 2^q binary words of length q, is particularly efficient to reduce both computational load and storing requirements. Besides, this kind of excitation is intrinsically robust to transmission errors [4]. Regular Pulse excitation has already been used at higher rates in RPE coders [8,9]. Introducing a RP codebook in a CELP coder can be seen as a RPE technique in which the pulse amplitudes are optimally vector quantized.

Using a RP codebook and with a reasonable approximation on R, the weighting term $\| H.c_k \|$ in equation (2) become independent of k [5,6]. So the search procedure comes down to maximize the inner product $P(k) = y^t.c_k$, which is very fast. With a binary RP codebook, it does not even require an

exhaustive search. The RPCELP algorithm involves a very small amount of computations and the perceived speech quality has been found equivalent to that of the original CELP.

3. ERROR MINIMIZATION CRITERION

There are two traditional ways of minimizing the error E(k). In the covariance approach, minimization is performed over the current block and does not make any assumption on error signal samples outside the block. The impulse response matrix H is a LxL lower triangular Toeplitz matrix.

In the autocorrelation approach, the error signal is weighted by a window and the minimization involves all weighted error samples from $-\infty$ to $+\infty$. When a rectangular window of length L is applied over the block, the memory terms of LP filters can be neglected and H becomes a (L+J)xL Toeplitz matrix, assuming that the impulse response h is practically zero after J samples. Both approaches have been compared in [6] and found to give similar quality performances. But the autocorrelation is of particular interest since it yields low complexity algorithms, as discussed in section 2.

In the autocorrelation method, block edge effects can be reduced by performing pre or post-windowing (overlapping). Pre-windowing makes the window begin J_b samples before the block. It is a way of taking into account the memory terms [4]. With $J_b = 7$, it gives an improvement of 0.2 dB on both global and segmental SNR values (averaged over 50 seconds of clean speech from 4 speakers). Post-windowing extends the window by adding J_e error samples at the end of the block. It is a way to evaluate the influence of the choice of a codeword on the next block. The best result is obtained with a window applying a decreasing weight to these samples (trapezoidal window). With $J_e = 7$, an improvement of 0.3 dB is further added to the SNR values. A similar result is obtained for $J_e = 2$, but with a rectangular window. So this generalized autocorrelation formulation is an efficient way to reduce block edge effects.

4. LOW COMPLEXITY CLOSED-LOOP METHODS

4.1. Closed-Loop Optimization Techniques

The relatively high quality achievable in CELP coding relies on the analysis by synthesis procedure used to determine the optimum innovation sequence. Accordingly, speech quality is significantly improved by determining the long term prediction (LTP) parameters with a closed-loop procedure [12], as compared to a classical open-loop method [6]. However, the basic closed-loop procedure results in a huge computational load. Alternate methods [11] have been proposed, but the compexity is still high. This section introduces a new algorithm which enables a very fast computation of LTP parameters in closed-loop.

More generally, the error signal energy to minimize is a function of LTP and innovation sequence parameters. These parameters can be sequentially computed, but here again speech quality is improved when all the parameters are jointly optimized [13]. In the latter case, the basic complexity is quite prohibitive, but we show that the efficient approach mentioned above can be extended to a fast joint computation of the LTP gain together with the innovation sequence parameters.

The LTP and innovation sequence gains can be quantized within the search procedure [4] (i.e. in closed-loop), or they can be quantized afterwards. The closed-loop method is generally required to get an optimal set of parameters after quantization, but it is also much more complex. Though, we will see that in the particular case of RPCELP, an "open-loop quantization" method may lead to the same result as the closed-loop procedure.

4.2. A Fast Algorithm for Closed-Loop LTP

The contribution of the single-tap long term predictor can be considered as a scaled output of an adaptive codebook constituted of past excitation sequences e_T [11], as shown in Fig. 1. Then the determination of LTP parameters appears to be quite similar to the innovation codebook search procedure. Thus, similar observations as those used to derive the RPCELP analysis structure [5,6] can be applied to LTP analysis in closed-loop.

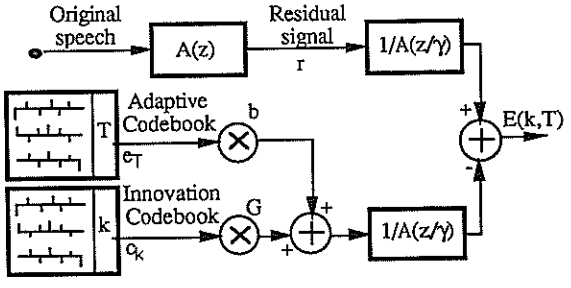


Figure 1: **Block Diagram of a Closed-Loop Analysis**

Assuming that the residual signal r takes into account the memory terms of both filterings through 1/A(z/γ), they can be considered as memoryless filterings with an associated impulse response matrix H. The error signal energy E is then expressed as

$$E = (H.r - b.H.e_T - G.H.c_k)^2 \quad (4)$$

Setting the excitation signal from the innovation codebook to zero, E is first minimized with respects to b and T. The optimal LTP parameters are then determined in two steps:

1) Find the lag T_o which maximizes

$$Q(T) = (r^t.H^t.H.e_T)^2 / \| H.e_T \|^2 \quad (5)$$

2) Compute the related gain b_o

$$b_o = (r^t.H^t.H.e_{T_o}) / \| H.e_{T_o} \|^2 \quad (6)$$

The above procedure is quite complex because all the possible excitation sequences e_T must be filtered through 1/A(z/γ) to get $H.e_T$ and compute Q(T) for each T. The problem is easily solved for the numerator of Q(T) which can be conveniently re-written [3,6]

$$N(T) = (r^t.H^t.H.e_T)^2 = ([H^t.H.r]^t.e_T)^2 \quad (7)$$

The vector $y = [H^t.H.r]$ is determined once per analysis block and only inner product computations are further required to get N(T) for each T. Though, the problem still

remains for the computation of the denominator. In a first approximation, the denominator can merely be neglected, which leads to a very fast sub-optimal method. However, the efficiency of closed-loop LTP analysis is significantly reduced with this simplification, especially when the block size is small. Another solution consists of determining the lag with an open-loop procedure and still computing the gain with the closed-loop formulation of equation (6). The underlying idea is that the long term predictor is essentially efficient for voiced sounds, where the LTP lag is closely related to the pitch period. So open or closed-loop method will often lead to the same value. In fact, the results with this "mixed" procedure are intermediate between open and closed-loop performances.

These experiments leads to the conclusion that to achieve an accurate LTP analysis, T_o must be determined in closed-loop using a normalized cross-correlation function like Q(T). The normalization is essential because the energy of e_T may vary a lot with T. But all these sequences are extracted from the same signal, covering a small duration, so their spectral properties are very similar. Then filtering the sequences (i.e. computing $H.e_T$) is expected to have a small impact on the distribution of their energies. So the lag T_o is now determined by maximizing

$$Q'(T) = N(T) / \| e_T \|^2 \qquad (8)$$

and the related gain is still computed using equation (6). Indeed, this fast procedure has been found to give equivalent performances to standard closed-loop analysis, both with objective measures (SNR) and informal listening tests.

The method can be further simplified when the error criterion uses an autocorrelation approach (see sections 2 and 3). In that case, $y = H^t.H.r$ is efficiently computed with a filtering operation [8] based on the matrix R, as defined in section 2, just like for the determination of the innovation sequence in RPCELP [5,6]. Using the same convenient perceptual weighting filter as in equation (3), R can even be time-invariant without any loss of subjective quality.

The division in the computation of Q'(T) can be avoided, using cross products to compare successive values. Moreover, each energy term $\| e_T \|^2$ can be efficiently computed from the previous one $\| e_{T-1} \|^2$ using the particular structure of the adaptive codebook: successive sequences differ in only two samples. With these remarks, the above closed-loop LTP analysis has a very low complexity.

4.3. A Fast Joint Optimization of LTP and Innovation Sequence Parameters

When LTP and innovation sequence parameters are to be jointly optimized, the error signal energy E expressed in equation (4) is minimized by setting its derivatives with respects to b and G to zero. It comes [14]

$$b = (R_1.E_2 - R_2.R_3) / (E_1.E_2 - R_3^2) \qquad (9)$$

and $$G = (R_2.E_1 - R_1.R_3) / (E_1.E_2 - R_3^2) \qquad (10)$$

with $$R_1 = r^t.R.e_T\ ,\ R_2 = r^t.R.c_k\ ,\ R_3 = e_T{}^t.R.c_k\ , \qquad (11)$$

$$E_1 = \| H.e_T \|^2 = e_T{}^t.R.e_T\ ,\ E_2 = \| H.c_k \|^2 = c_k{}^t.R.c_k\ .$$

Replacing b and G in (7) with expressions (12) and (13), optimal indexes T_o and k_o are now determined by maximizing

$$C(T,k) = (R_2^2E_1 + R_1^2E_2 - 2R_1R_2R_3) / (E_1.E_2 - R_3^2) \qquad (12)$$

The complexity is very high if T_o and k_o are to be jointly computed. A good strategy [13] is to first determine T_o, and then search for k_o which maximizes $C(T_o,k)$ in equation (12). The lag T_o can be efficiently determined with the fast closed-loop procedure described above. Then $T = T_o$ being fixed, R_1 and E_1 are known. The computation of R_2 do not require any additional filtering since $r^t.R$ have already been computed for the determination of T_o, and the computation of R_3 only involves one filtering per block to get $e_{T_o}{}^t.R$. The quantity E_2 must be computed once per frame, which generally requires the filtering of each codeword c_k in the innovation codebook. Structured or sparse codebooks [4,6,14] may be used to significantly speed up the filterings and inner products in the computation of E_2 and R_3. Moreover, with a RPCELP structure this term turns out to be constant [6]. In that case, the joint optimization of LTP and innovation sequence parameters becomes feasible on low cost DSP chips, though it remains slightly more complex than a sequential optimization.

4.4. Closed-Loop Gain Quantization

For practical applications, the LTP and innovation sequence gains b and G have to be quantized to discrete values b_i and G_j. The gains may be computed as described above using equations (6) or (9) and (10), and quantized afterwards. But once the gains are quantized, the determination of the sequence indexes T_o and k_o might not be optimal anymore. In general, the gains must be quantized during the search procedure to get an optimal set of parameters. In that case, b and G must be replaced by their quantized values b_i and G_j in the expression of the error signal energy given in equation (4) and an optimal search procedure requires an exhaustive search through all possible values of T, k, i and j. This would result in a quite prohibitive computational load. However, when LTP and innovation sequence parameters are determined sequentially, an efficient approach is to compute and quantize b after the search for T ("open-loop quantization"), and to still quantize G during the search for k ("closed-loop quantization"). Here the argument is that the choice of an optimal set of parameters (k_o,j_o) for the innovation sequence can compensate for small inaccuracies in LTP parameters.

Furthermore, for a RPCELP coder the energy terms E_2 are constant and for this reason it can be shown [15] that quantizing G afterwards leads exactly to the same result as the exhaustive search, under a simple and non-restrictive condition on the gain quantizer. This property can be extended to any CELP coder through a normalization of each codeword by its "perceptual" energy and this technique is refered to as "Normalized CELP" [15].

5. DESIGN OF 6 AND 8 KBPS RPCELP CODERS

The design of a 6 kbps RPCELP coder has been presented in [6]. Here we describe a modified 6 kbps coder and a new 8 kbps coder, that both incorporate the improvements introduced in this paper. Real time implementation of these coders is then detailed.

5.1 Quantization Issues and Bit Allocation

Both coders uses 8 kHz sampled speech and their bit allocation is summarized in Table 1. The quantization of LTP parameters is a function of the subframe. For the 6 kbps coder, in subframe 1, T is chosen in [40,103] and quantized with 6 bits and b is restricted to [0,1] and quantized with 2 bits. For subframes 2,3 and 4 we use an absolute or differential 6 bits vector quantizer for (b,T) or (b,ΔT), depending on the values of b and of the differential lag ΔT. This method uses the fact that in voiced sounds ΔT is generally small and b close to 1. For the 8 kbps coder, in subframes 1 and 3, T is chosen in [20,147] and quantized with 7 bits while in subframes 2 and 4, T is delta coded relative to the previous delay with 5 bits. In all subframes the LTP gain is scalar quantized with 4 bits.

Table 1: **Bit Allocation of 6/8 kbps Coders**

6/8 kbps	LPC filter	LTP	Codebook
Update	20 ms	20/4= 5ms	20/4= 5ms
Order	8/10	1 tap	Decim. factor= 4/3
Analysis	autocorrelation, no pre-emphasis	closed loop	closed loop Codebook 12/15 bits
Bits/ frames	30/40 bits reflexion coefficients	26/40 bits for index and gain	index= 48/60 bits gain = 4/5 bits*4
Rate	1500/2000	1300/2000	3200/4000

5.2 Real Time Implementation

Fixed point simulations have been realized to prepare real time implementation on a 16 bits DSP chip (TMS320C25). Scaling factors have been optimized to achieve the maximum precision, using dynamic scaling and 32 bits computations when necessary. The time budget was estimated for each coder and is shown in Table 2. The overall complexity of both 6 and 8 kbps coders is around 5 Mips for the whole coder / decoder process and the memory requirements are 4 Kwords of program and 2 Kwords of data.

Table 2: **Computational Requirements**

Bit Rate	6 kbps	8 kbps
LP Analysis	1.2 ms	1.9 ms
Residual Computation	4 * 0.45 ms	4 * 0.50 ms
LTP Analysis	4 * 0.80 ms	4 * 0.85 ms
Excitation Search	4 * 0.23 ms	4 * 0.25 ms
Coder	7.2 ms	8.3 ms
Decoder	2.9 ms	3.4 ms
TOTAL (MIPS)	5.05	5.85

CONCLUSION

Improved Regular Pulse CELP coders have been presented. Block edge effects are reduced using an efficient error minimization criterion. Speech quality is also significantly improved with the implementation of closed-loop optimization techniques. New complexity reduction methods have been introduced for the computation of LTP parameters in closed-loop and for the joint optimization of LTP and excitation parameters. Improved RPCELP coders at 6 and 8 kbps are real time implemented on a TMS320C25. They produce high quality speech with a very low complexity and are therefore particularly suitable for narrow band speech transmission.

REFERENCES:

[1] I. Lecomte, M. Lever, L. Lelièvre, M. Delprat, A. Tassy, "Medium band speech coding for mobile radio communications," in Proc. ICASSP, Apr. 1988.

[2] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," in Proc. ICASSP, Mar. 1985.

[3] J. P. Adoul, P. Mabilleau, M. Delprat and S. Morissette, "Fast CELP coding based on algebraic codes," in Proc. ICASSP, Apr. 1987.

[4] A. Le Guyader, D. Massaloux, J. P. Petit, "Robust and fast code-excited linear predictive coding of speech signals," in Proc. ICASSP, May 1989.

[5] M. Lever and M. Delprat, "RPCELP: A high quality and low complexity scheme for narrow band coding of speech," in Proc. EUROCON, June 1988.

[6] M. Delprat, M. Lever and C. Gruet, "Efficient excitation model and fast selection in CELP coding of speech," in Proc. EUROSPEECH, Sep. 1989.

[7] J. P. Campbell, V. C. Welch, T. E. Tremain, "An expandable error-protected 4800 bps CELP coder (U.S. Federal Standard 4800 bps voice coder)," in Proc. ICASSP, May 1989.

[8] P. Kroon, E. F. Deprettere, R. J. Sluyter, "Regular pulse excitation: a novel approach to effective and efficient multi-pulse coding of speech," IEEE Trans. ASSP, vol. 34, Oct. 1986.

[9] P. Vary, K. Hellwig, R. Hofmann, R. J. Sluyter, C. Galand, M. Rosso, "Speech codec for the European mobile radio system," in Proc. ICASSP, Apr. 1988.

[10] M. Berouti, H. Garten, P. Kabal, P. Mermelstein, "Efficient computation and encoding of the multipulse excitation for LPC," in Proc. ICASSP, Mar. 1984.

[11] W. B. Kleijn, D. J. Krasinski and R. H. Ketchum, "An efficient stochastically excited linear predictive coding algorithm for high quality low bit rate transmission of speech," Speech Communication, vol. 7, 1988.

[12] S. Singhal and B. S. Atal, "Improving performance of multi-pulse LPC coders at low bit rates," in Proc. ICASSP, Mar. 1984.

[13] P. Kabal, J. L. Moncet and C. C. Chu, "Synthesis filter optimization and coding: applications to CELP," in Proc. ICASSP, Apr. 1988.

[14] L. Cellario, G. Ferraris and D. Sereno, "A 2 ms delay CELP coder," in Proc. ICASSP, May 1989.

[15] R. Di Francesco, M. Delprat, M. Lever, "A fast procedure for speech coding: Normalized code excited linear prediction," submitted to IEEE Trans. Com.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

CONSIDERATIONS FOR REAL-TIME IMPLEMENTATION OF A 4.8 KBPS CELP CODER

L.A. Hernández-Gómez*, F.J. Casajús-Quirós*, A. Pena-Gimenez*, C. García-Mateo** and E. López-Gonzalo*

* ETSI Telecomunicación-UPM, Cdad. Universitaria, 28040 Madrid, SPAIN
** ETSI Telecomunicación-USC, Aptdo. 62, 36280 Vigo (Pontevedra) SPAIN

[1] *The aim of this contribuition is to discuss some considerations related to a real-time implementation of the Proposed Federal Standard 1016 4800 bps voice coder developed by the U.S. DoD and AT&T Bell Laboratories. The main specific procedures we have used in order to have a real-time implementation on a commercial board with a single DSP32C processor are described. We have focussed our attention specially on those procedures where most of the computation is concentrated: the adaptive codebook search and the stochastic codebook search.*

1. INTRODUCTION

A Code-Excited Linear Predictive (CELP) coder jointly developed by the DoD and AT&T Bell Laboratories has been recently proposed as the U.S. Government Standard 4800 bps voice coder [1]. This coder is based on an enhaced version of the coder selected after a complete evaluation of 4800 bps voice coders [2]. The proposed coder is very flexible providing different possibilities for real-time implementations depending on different computation power offered from different commercially available DSP chips. Moreover the Proposed Federal Standard (PFS-1016) presents a wide range of applications [1] including Land Mobile Radio, Mobile Satellite communications, the possibility of replacing many existing systems now based on 16000 bps CSVD, etc.; CELP being also very suitable for applications where voice must be digitized prior to encryption.

In this contribution we will refer to the main results related to our work directed towards a real-time implementation of the 4800 bps standard using a commercial board by Loughborough Associates.

Because CELP's most computational effort is dedicated to search procedures of the adaptive codebook and the fixed stochastic codebook we have foscussed our attention on the different possibilities these two processes present. Nevertheless we have also considered the main issues related to the evaluation and quantization of the parameters which represents the spectral envelope, in particular the evaluation and quantization of the Line Spectrum Pairs (LSP).

The remainder sections of the paper are organised as follows. Firstly, we present (section 2) a brief description of the basic CELP algorithm. Section 3 discuss some general topics related to the search procedures for both the adaptive and the stochastic codebook. The particular adaptive codebook search we use is discussed in Section 4. Section 5 presents the search procedure for the fixed stochastic codebook. Some issues related to the spectral envelope representation are presented in Section 6. Details about our preliminar implementation on a commercial board by Loughborough Associates based on a single AT&T DSP32C can be found in Section 7.

2- CELP ALGORITHM

The Code-Excited Linear Predictive (CELP) algorithm can be considered as a two-stage Vector Quantizer (VQ) for the spech signal. The first VQ stage is refered to as *adaptive codebook contribution*, and a codebook adaptely constructed from the past excitation to the synthesis filter is used. The second stage is refered to as *stochastic codebook contribution*, and it is based on a fixed codebook randomly generated.Both stages, adaptive and stochastic contributions, need a synthesis filter in order to represent the speech signal. This synthesis filter represents the spectrum envelope of the original speech and is evaluated from short-delay prediction. A 10th order linear prediction filter is estimated by means of the autocorrelation method, using a 30 ms Hamming window. Linear prediction coefficients are transformed to LSP coefficients as will be considered in Section 6. This coefficients are linearly interpolated in order to form an intermediate parameter set for each of the excitation subframes.

After linear prediction the optimum excitation sequences for the synthesis filter are obtained from both the adaptive and the stochastic codebook. This operation is

[1] This work has been supported by the PLANICYT under the Project "Tratamiento Avanzado de la Información".

made by using exhaustive search and thus most of the computational load depends on how this procedures are implemented. The particular procedures we have used in our real-time implementation will be described in the next sections. The adaptive codebook search is performed with a delta searching and coding of even frames every 7.5 ms and, according to PFS-1016, 128 integer delays and (optionally) 128 non-integer delays can be used. The stochastic codebook search is also performed every 7.5 ms, and again according to PFS-1016 a subset of a maximum of 512 codevectors can be used. Our present implementation scheme includes 256 codevectors in order to meet the final requeriments of our real-time prototype.

3. SEARCH PROCEDURES

A basic representation of the search procedure for both the adaptive and the stochastic codebook can be expressed as a search for the minimum error:

$$\sum(x-Gy)^2=\sum x^2+G^2\sum y^2-2G\sum xy \qquad (1)$$

where x is the signal to be represented at each stage, y corresponds to the sucessive codevectors and G is the scale factor.

The evaluation of y requires to perform a filtering operation and thus most of the tasks involved in the search procedures of both adaptive and stochastic codebooks are convolution and energy calculations.

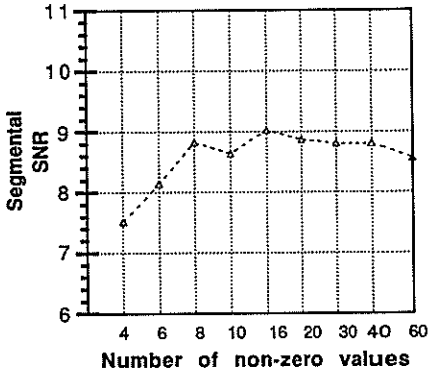
Our discussion for a final real-time implementation is strongly related to the behaviour of the two last terms in the right side of expression (1): i.e. the energy of y and the correlation between x and y.

The different behaviour of these two terms for the adaptive and the stochastic codebooks is the base for the different procedures we have used in our real-time implementation. In particular we found a small variance in the energy term for the stochastic codebook relative to the correlation term. So we have design a simplified two-step search procedure (described in Section 5). On the other hand a similar importance for both the energy and the correlation terms were found for the adaptive codebook and thus the simplified two-step search is not included here. Fortunately other simplified procedures can be applied for the adaptive search as will be discussed in the next section.

4. ADAPTIVE CODEBOOK SEARCH

The procedure we have included in our coding algorithm is based on a straight calculation of the energy factor related to every candidate (codevector) and an efficient strategy for the evaluation of convolutions. This strategy uses a truncated impulse response together with overlap and add operations. We have found that even with a reduced number of samples for the impulse response the final quality of the coder is maintained. Figure 1 shows different SNRs for different non-zero samples for the impulse response, and a small influence is perceived.

Figure 1. SNR versus Non-zero Impulse Response



Moreover in order to obtain additional reductions in computational cost, for those long-term delays shorter than subframe length, recursive and overlap operations are used to efficiently obtain the adaptive codebook contribution.

Finally using non-integer delays provides a smother variation of the prediction delay, which is particularly important for non-exhaustive search procedures based on prior information (i.e delay of the previous frame). In this way the total computation complexity for the adaptive codebook search can be controled by using an structurated search depending on both the delay values from the previous frame and the specific characteristics of the waveform representing the function to be maximized during the search procedure. According to our preliminar simulation results, this structured search together with the use of non-integer delays can lead to perceptually noticeable improvements in the final quality while providing efficient (low computational cost) procedures.

5. STOCHASTIC CODEBOOK SEARCH

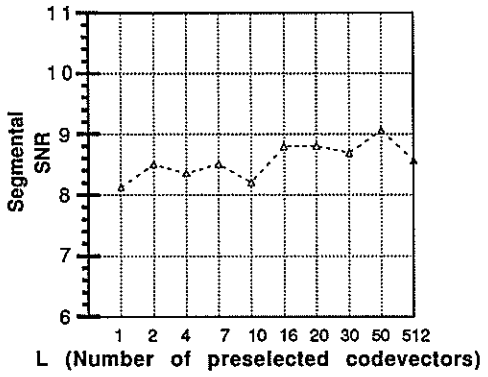
From the considerations in Section 3 a simplified two-step search procedure for the optimum codevector in the stochastic codebook can now be included. According to our simulation results a low complexity algorithm can be based on a *first* selection of L codevectors in the stochastic codebook using only the correlation information (see expression 1). This procedure results in an important reduction in computation because the energy factor (see (1)) is not evaluated in the first step. And the energy term is the most important factor in terms of computation.

It is important to point out that in the second step of the proposed simplified algorithm the energy factor can not

be recursively evaluated as in the original procedure. This is because the pre-selected L vectors generally do not correspond to shifted vectors (as it hapens in the original codebook). Therefore depending on the number of pre-selected vectors (L) and on the number of non-zero values for the impulse response the global reduction in computational cost can be quite different.

According to our preliminar simulation experiences good results can be obtained by using L=20 and a number of non-zero samples for the impulse response equal to 30. Figure 2 represents the SNR obtained for different numbers of pre-selected vectors, i.e. different values of L.

Figure 2. SNR versus L



An additional reduction in the complexity of searching in the stochastic codebook can be based on efficient procedures for the evaluation of correlation terms. In particular we have included a correlation procedure which exploits the high number of zero-valued elements in the excitation codebook based on the storage of the relative positions of the non-zero elements in the sucessive excitation vectors.

Finally our algorithm includes a joint optimization process for the codevector index and the quantized gain factor. This procedure is restricted to the L pre-selected vectors. Moreover the optimum gain factor for these L pre-selected codevectors is stored in order to check during the joint optimization of index and gain only those codevectors having the possibility of being selected in the final search. This procedure introduces a new reduction in computational cost for the stochastic codebook search because, according with our simulation results, the quantization is only performed in average three or four times per frame. Furthermore this joint optimization introduces an important increase in the final quality of the synthetic speech [1].

6. SPECTRAL ENVELOPE REPRESENTATION

The Proposed Federal Standard (PFS-1016) includes a spectrum analysis performed every 30 ms based on a 10th order autocorrelation linear prediction. The spectral envelope representation (synthesis filter) is obtained by transforming the linear prediction coefficients into Line Spectrum Pairs that must be quantized prior transmission.

An efficient procedure has been used for a fast conversion of linear prediction coefficients directly to quantized LSPs. The evaluation of LSPs is related to a search for the roots of two polynomials. The evaluation/quantization algorithm we have used (described in [3]) combines a search procedure with the quantization grid in order to provide a fast operation.

Our present work related to spectral envelope representations is directed to the comparission of the use of other coefficients different from the LSPs. In particular our still preliminar simulation results show similar quality for the use of log area ratios, inverse sine or LSP, while some of this coefficients can be evaluated with a reduced ammount of computation.

7. IMPLEMENTATION

The coding scheme we have described in the preceding paragraphs is amenable to real-time implementation. However some important remarks must be made:

The main problem considering a real-time implementation arises in the conversion from "arithmetic" MIPS to real "DPS chip MIPS". And of course it is crucial to take into account the "real-time overhead" due to memory constrains, loop setup, etc. There are some techniques we have used to maintain the DSP chip MIPS into reasonable values. In general those methods are based on:

* The use of circular buffers for recursive operations instead of data movements.
* Recursive convolutions during the codebook search greatly reduces the computational complexity relative to brute force techniques.
* Reducing the loop setup. This could be achieved by optimizing the sequence of operations, for example during convolution evaluations.
* Using different data representations, for example using pointers for representing the stochastic codebook. This can help both in memory storage and computational complexity.

Following these topics and using a commercial board by Loughborough Associates based on the DSP32C by AT&T, we have been able to realize our coder. There is a single processor chip on the board capable of performing

12.5 MIPS, including single-cycle floating-point multyply-add. There are 128 Kbyte of slow memory and 8 Kword of fast memory readily available. The board also features two analog input/output channels with 16 bit A/D and D/A converters.

We started with the C-language version of the coder written according to some heuristic guidelines in order to produce efficient assembly code. After compiling this version with AT&T C-compiler we obtained an assembly code version capable of running in about three times real-time. A set of macros previously developed by us made it possible to substitute part of the C code so as to increase efficiency. A final hard-optimization of the assembler output was found to be necessary, thus reaching a stage in wich our coder was running in real-time.

Due to the architecture of the board we have used, our coding scheme has been divided into three independent processes:

- **Transmitter:** it is a set of routines wich operate in background mode, reading a frame of speech samples from the input buffer and yielding the bits corresponding to coded speech as output.
- **Receiver**: this process synthetizes speech by taking the set of bits received through a serial line as its input. The synthesized speech it produces is written to an output buffer, and the process runs in foreground mode.
- **I/O**: this set of routines manage the input/output buffers. There is an internal clock generating an interrupt to the processor every 0.125 ms. Whenever an interrupt is received the I/O routines read a sample from the A/D converter and pass it to the input buffer, after that, a sample is taken from the output buffer and it is written to the D/A converter.

Transmitter routines are by far the most complex of the coder. They can be divided into three functional blocks, which are:

- **Spectral analysis**: which taking a frame of speech samples produces the corresponding linear prediction parameters in the form of line spectrum pairs (see Section 6).
- **Long predictor**: it calculates the best codevector in the adaptive codebook (see Section 4).
- **Excitation selection**: it selects from the fixed stochastic codebook the codevector that produces the best synthetic speech (see Section 5).

Taking into account this subdivision, table I reflects the distribution of the time interval between analysis frames among the different tasks to be carried out by the coder. The use of data memory (0 wait-state) can be seen in table II.

8. CONCLUDING REMARKS

Some considerations for a real-time implementation of the PFS-1016 4800 bps voice coder on the AT&T DSP32C Digital Signal Processor using a board by Loughborough Associates has been presented. Our attention has been focussed on the use of reduced search procedures for both the adaptive and the stochastic codebook search. These two procedures have been found to be of crucial importance for real-time implementation of this kind of algorithms. The efficient representation of the spectral envelope has also be considered. Finally details about a preliminary implementation which includes the procedures and algorithm presented in the different sections of the paprer have been presented.

Our present work is directed towards the future imprivements of this preliminar real-time version, and also towards the research on future expansions of the basic CELP algorithm: in particular we are considering the possibility of using different coding strategies for frames of speech with different phonetic characteristics; thus providing better quality speech at 4800 bps or lower bit rates while keeping the present speech quality.

9. ACKNOWLEDGEMENTS

We gratefully acknowledge the colaboration of V. Molinero, M.A. Rodriguez Hernández and J. Alvarez Cercadillo who did the hard work.

10. REFERENCES

[1] J.P. Campbell, T.E. Tremain. and Jr; V.C. Welch "The DoD 4.8 Kbps Standard (Proposed Federal Standard 1016),". Advances in Speech Coding. Kluwer Academic Publishers 1990.

[2] D. Kemp, R. Sueda and. T.E. Tremain."An Evaluation of 4800 bps Voice Coders," Proceedings of the IEEE Intl. Conf. ASSP 1989, p. 200-203.

[3] F.J. Casajús Quirós, L.A. Hernández Gómez, and C. Garcfa-Mateo "Analysis and Quantization Procedures for Real-Time Implementation of a 4.8 Kbps CELP coder,". in proc. ICASSP, April 1990.

Table I

TASK	% Interframe interval
Spectral Analysis	7
Adaptive Codebook	42
Stochastic Codebook	40
Receiver	8
I/O	rest

Table II

USE	Memory 32 bit words
I/O buffers	960
Codebooks	300
Quantization tables	640
Miscellaneous	128

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

BI-FILTER LPC VOCODER

Dinei A. F. Florencio and Henrique S. Malvar

Dept. de Engenharia Elétrica, Universidade de Brasília
70910 Brasília, DF, Brazil

A new model for LPC speech synthesis is introduced, called bi-filter LPC, where the periodic and aperiodic excitation drive two distinct filters. The bi-filter vocoder is capable of producing synthetic speech that is free from the "buziness" of traditional LPC vocoders. The analysis procedure is presented, as well as experimental results.

1. INTRODUCTION

In low bit rate speech coding the standard LPC vocoder [1] is widely used, mainly when the voice quality is not a prevalent factor, e.g., in audio response units. There are other applications, like voice mail, where low bit rate is required but voice quality must be as high as possible. Therefore, in recent years there has been done a lot of research into techniques for improving the traditional LPC vocoder.

It has long been recognized that one of the major shortcomings of the standard LPC synthesis model [1] of Fig. 1 (a) lies in the excitation source, with the simple voiced/unvoiced switch being too simplistic for human speech production. Thus several models for the excitation source have been recently proposed, including selecting vectors from codebooks and multiple pulses per pitch period. These are both computationally intensive, and so other solutions must be found. In [2] an improvement in the speech quality was obtained by simply controlling the periodic and noise excitations with appropriate gain factors, thus avoiding the hard voiced/unvoiced switch. In [3] it was proposed that the voiced/unvoiced decision be made in the frequency domain, for each of the predefined subbands of the excitation signal.

In this paper we propose a new model for LPC speech synthesis, depicted in Fig. 1 (b). Instead of using a single filter that is driven by either a periodic (impulse train) or an aperiodic (noise) excitation, the two forms of excitation drive two different filters. The reason behind this bi-filter approach is that the physical sources of the two kinds of excitation are located in different regions of the human vocal tract, and are therefore subjected to two different filtering processes.

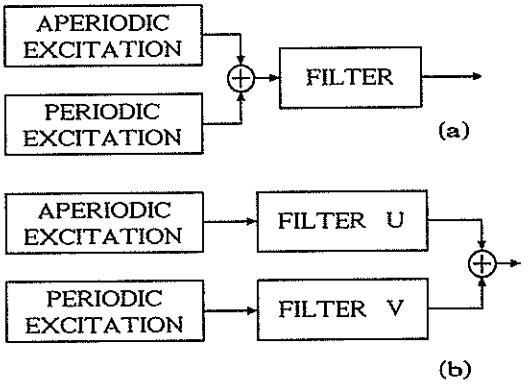


Fig. 1. Speech synthesis models: (a) standard LPC; (b) bi-filter LPC.

With the two filters in Fig. 1 (b) the periodic and aperiodic excitations can be independently processed, and so we expect that mixed sounds like /z/ and /v/ can be better reproduced. We refer to the first filter as the "voiced filter" **V** and the second as the "unvoiced filter" **U**, as shown in Fig. 1. In the following we will describe the simplest vocoder that can be implemented using such a model.

2. ESTIMATION OF THE MODEL PARAMETERS

For the bi-filter synthesis, there are five sets of parameters that must be estimated: the pitch period, the periodic and aperiodic excitation energies and the two sets of filter coefficients. The procedure that we have adopted is depicted in Fig. 2.

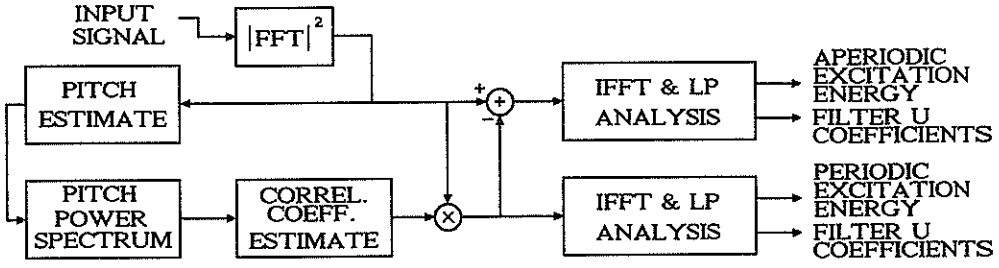


Fig. 2. Analysis steps in the bi-filter vocoder.

The main problem is that of separating the input spectrum in its periodic and aperiodic parts. To do that, let us start by recalling that the Fourier transform of an infinite periodic signal is an impulse train. If we window the signal (as we have to do in the analysis procedure), then the effect in the frequency domain will be the convolution of the impulse train with the window transform, generating the sidelobe pattern that is typical of power spectra of windowed periodic signals. We can model this power spectrum as a product of an spectral envelope $V(\omega)$ and a regular sidelobe pattern $P(\omega)$, which corresponds to the transform of a windowed impulse train. If the signal in non-periodic, the Fourier transform will not have any regular pattern. We may also model this power spectrum as a product of another spectral envelope $U(\omega)$ and a constant irregular pattern $N(\omega)$, which in fact corresponds to the transform of a windowed white noise.

Thus, supposing that the periodic and aperiodic parts from which the input speech segment is composed are uncorrelated, we can approximate its power spectrum $S(\omega)$ by

$$S(\omega) = S_V(\omega) + S_U(\omega) = V(\omega)P(\omega) + U(\omega)N(\omega) \ . \quad (1)$$

The factor $P(\omega)$ can be easily obtained once the pitch period is known, and $N(\omega)$ is the best estimation for the power spectrum of a windowed white noise segment. Given $S(\omega)$, $P(\omega)$, and $N(\omega)$, we want to estimate the power spectra $U(\omega)$ and $V(\omega)$, in order to obtain the correspondent filter coefficients.

Using the orthogonality principle, we can prove that the $V(\omega)$ that gives the best m.s.e. approximation to $S(\omega)$ is given by

$$V(\omega) = \frac{E\{S(\omega)[P(\omega) - E\{P(\omega)\}]\}}{E\{[P(\omega) - E\{P(\omega)\}]^2\}} \ , \quad (2)$$

where $E\{.\}$ is the operation of finding the envelope. Of course, $E\{P(\omega)\}$ is a constant, equivalent to the average level of $P(\omega)$. Similarly, the best estimate of $U(\omega)$ is obtained from

$$U(\omega) = E\{S(\omega)\} - V(\omega) \quad (3)$$

In practice, we do not know $S(\omega)$ exactly, and so it must be estimated from a segment of the input speech. In our simulations we have used a 16 ms frame, on a 8 kHz sampled speech, with an overlapping factor of 50%. Therefore we have 256 signal points per frame. The frame is Hamming windowed, zero padded to 512 points, and $S(\omega)$ is obtained from the magnitude squared FFT (the box labeled $|\text{FFT}|^2$ in Fig. 2).

Because of the windowing operation, the spectrum separation procedure must be slightly modified. Using the convolution with a Hamming window (with $512/p$ points, where p is the pitch period) as the $E\{.\}$ operator, we compute the correlation coefficient $C(\omega)$ between the signal power spectrum $S(\omega)$ and the pitch power spectrum $P(\omega)$ as

$$C(\omega) = \frac{E\{S(\omega)P(\omega)\}}{\sqrt{E\{S^2(\omega)\}E\{P^2(\omega)\}}} \ . \quad (4)$$

This correlation coefficient $C(\omega)$ needs some real world adjustments, which are described below, resulting in an adjusted correlation coefficient $C'(\omega)$, from which we obtain

$$\begin{aligned} V(\omega) &= C'(\omega)S(\omega) \ , \text{ and} \\ U(\omega) &= [1 - C'(\omega)]S(\omega) \ . \end{aligned} \quad (5)$$

Computing now the inverse FFT for each of these power spectra we obtain separated estimates for the "periodic part" autocorrelation and "aperiodic part" autocorrelation sequences. Using these autocorrelation estimates and the Levinson-Durbin's

algorithm [1] we obtain the two sets of filter coefficients and the two excitation energies.

The analysis procedure described above is sensitive to pitch precision; even small errors in the pitch estimation may lead to the classification of the higher frequency part of a periodic signal power spectrum as being aperiodic. To ensure a good accuracy in the pitch estimation we select the pitch period that maximizes the correlation coefficient between the signal power spectrum and the pitch train power spectrum. We have obtained excellent pitch estimation using this method. In order to reduce the computational burden, we limit the search to a small neighborhood of an initial estimate.

3. EXPERIMENTAL RESULTS

We have simulated the bi-filter LPC vocoder using the analysis procedure of the previous section. Although it is capable of producing highly intelligible speech, some adjustments have shown to improve the quality of the synthetic speech. These adjustments are primarily intended to compensate for the following observed effects:

- We cannot always count on a perfect pitch estimate.
- The true pitch period may vary within the frame.
- In aperiodic frames the power spectrum may occasionally coincide with the pitch sidelobe pattern, suggesting that it would be a periodic component.

To compensate for these inaccuracies, we have introduced the following empirical corrections to the correlation coefficients $C(\omega)$:

- Give an arbitrary gain to all coefficients to compensate the imprecision; we have used a gain of two.
- Disregard negative coefficients, since they do not make sense.
- Compute an average of the correlation coefficients. If this average is greater than one, consider the frame 100% periodic and set the aperiodic excitation energy to zero, otherwise limit the coefficients to one.

In Fig. 3 we have a frequency domain comparison between the bi-filter and standard LPC vocoders models for a mixed phoneme /z/ in the Portuguese word "fazer" spoken by a woman. In Fig. 3 (a) we have the 256-point FFT of the original speech, in Fig. 3 (b) the synthesized spectrum for the standard LPC vocoder, and in Fig. 3 (c) the synthetic bi-filter spectrum. We note that with the standard LPC the periodicity of the synthesized signal extends over the whole frequency range,

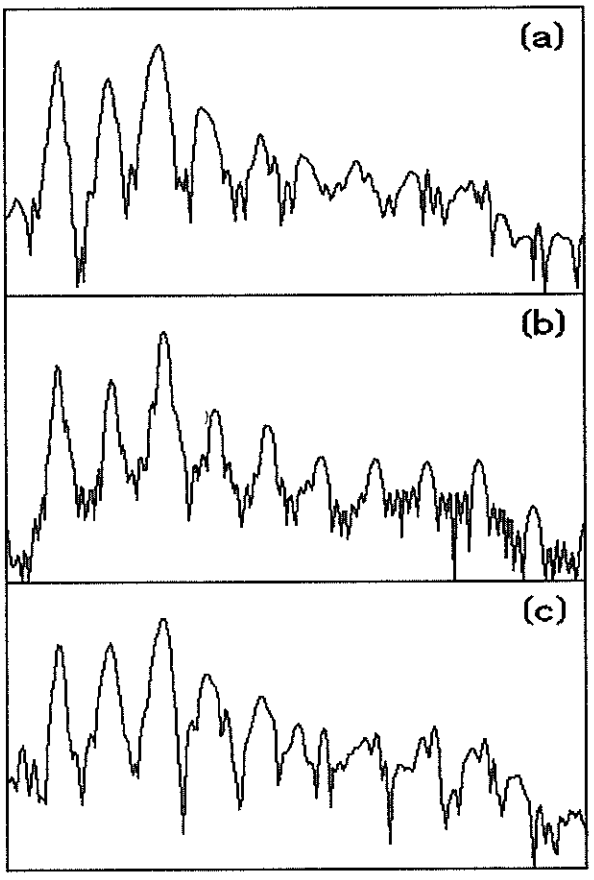


Fig. 3. Speech spectra (70 dB vertical range): (a) original /z/ sound; (b) synthesized with standard LPC; (c) synthesized with bi-filter LPC.

whereas the bi-filter method has produced a spectrum the becomes noise-like at higher frequencies, thus resembling more closely the input spectrum. Informal listening tests have demonstrated the ability of the bi-filter LPC vocoder to produce more naturally sounding speech than the standard LPC vocoder, with the "buzziness" effect [3] being completely eliminated.

We have observed that about 80% of the frames are classified either as 100% periodic or as 100% aperiodic, cases in which the bi-filter vocoder is reduced to a standard LPC vocoder. This suggests that the bi-filter vocoder can be used, presumably with better results, at about the same bit rates as the LPC vocoder. We are now working on strategies for efficient coding of the bi-filter parameters.

4. CONCLUSION

We have presented a new synthesis model that is capable of improving considerably the speech quality in an LPC vocoder. This new bi-filter vocoder is based on the idea that the periodic and aperiodic excitations should be processed by different filters. The corresponding analysis algorithm is simple enough to permit real-time implementations with a single DSP processor chip. With appropriate coding of the parameters, we believe that the bi-filter vocoder can operate at bit rates in the range of 2.4 to 4.8 kbps. We are currently working on the fully-coded bi-filter vocoder.

REFERENCES

[1] J. D. Markel and A. H. Gray, Jr, Linear Prediction of Speech. New York: Springer-Verlag, 1976.

[2] S. Y. Kwon and A. J. Goldberg, "An enhanced LPC vocoder with no voiced / unvoiced switch", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, pp. 851-858, Aug. 1984.

[3] D. W. Griffin and J. S. Lim, "Multiband Excitation Vocoder", IEEE Trans. Acoust., Speech, Signal Processing, vol. 36, pp. 1223-1235, Aug. 1988.

[4] J. L. Flanagan, "Speech Analysis Synthesis and Perception", 2nd ed. New York: Springer-Verlag, 1972.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

6.55 KBIT/S SPEECH CODING FOR APPLICATION IN THE PAN-EUROPEAN DIGITAL MOBILE RADIO SYSTEM

Rosario DROGO DE IACOVO and Daniele SERENO

CSELT - Via G.Reiss Romoli, 274 - 10148 Torino ITALY Phone +39.11.21691 - Telex 220539 CSELT

This paper describes our work in designing a speech coder for application in the GSM half-rate channel. The study has been oriented first at the definition of some schemes that have been compared to select a unique candidate. Then we considered the impact of channel errors on the transmitted bit-stream in order to optimize the channel coding scheme. The evaluation of the distortion introduced by channel errors has been carried out by measuring both the spectral distortion and the segmental SNR and by extensive listening tests. As a result we evaluated the subjective importance of the various bits and grouped them into five classes that have been used to test the speech coder performance with different channel coding schemes.

1. INTRODUCTION

The GSM pan-European Digital Mobile Radio (DMR) system has been defined with the capability to use alternatively both full-rate or half-rate channels. This allows the TDMA frame structure to accommodate both N channels at the nominal rate or 2N channels at half-rate.

The study presented in this paper has been carried out with the aim to design a coder suited for application in the half-rate subsystem and trying to achieve a speech quality comparable to the one provided by the full-rate coder.

The total bit-rate available for one half-rate channel is 11.4 kbit/s and by resorting to the experience gained in the definition of the full-rate scheme, the net bit-rate expected for the source coder at half-rate is around 6.5 kbit/s.

At this bit-rate the most suited speech coding technique seems to be the CELP coding [1,2]. However, several different schemes can be implemented, depending on the actual bit-rate, the considered sub-frame length, the parameters update rate, the technique used for the computation of the Long Term (LT) information, the quantization technique adopted for the side-information, the innovation codebook structure and so on. All these possibilities led us to the definition of four schemes with bit-rate ranging from 6.55 kbit/s to 7.45 kbit/s.

Then the selection among these four candidates has been carried out by comparing their performance with and without channel errors. To take into account channel errors we considered three conditions corresponding to a C/I ratio of 4,7 or 10 dB. After this, the efforts were concentrated on the determination of the design parameters for the channel coder. To do so, we computed the sensitivity to channel errors for each bit in the frame. Performance were measured by computing the spectral distortion and the segmental SNR as in [3]. From these results and by means of intensive listenings to the speech material we arrived at the definition of five classes of importance that have been used to optimize the Forward Error Correcting (FEC) algorithm.

Finally a further mechanism to cope with highly corrupted frames has been considered. It consists in replacing, at the decoder, the received parameters by others evaluated exploiting the information carried by the previous frames. The decision on when adopting this strategy is taken considering a Bad Frame Indicator (BFI) flag provided by the channel decoder.

2. CODER DESCRIPTION

The various schemes we have studied in this work, all belong to the class of CELP coders and differ for the operating bit-rate, the evaluation of the long term filter parameters and the sub-frame length.

The block diagram of a CELP transmitter can be represented as in Fig. 1. The innovation signal, scaled by the proper gain factor, is fed into the cascade of two synthesis filters which carry the information of the long term and short term correlation respectively, to produce the output signal. The optimum innovation vector is selected among a finite collection (codebook) by minimizing the spectrally weighted error between the original and the output signals. The information to be transmitted to the receiver consists of the filter coefficients, the scale factor and the innovation index.

The four CELP schemes we considered in our simulations are based on a frame length of 20 ms which is the same framing of the GSM full-rate coder. Due to the interleaving employed in the TDMA system, a shorter frame length could allow only a marginal reduction of the system delay and thus it has not been considered here. The spectral information is computed on this frame basis and represented in terms of 10 Line Spectrum Pairs (LSP) that are quantized with 30 bits. In addition the LSP parameters are interpolated between contiguous frames, in order to smooth spectral transitions.

The long term correlation information, sometimes called adaptive codebook, can be computed by means of an open loop procedure starting from the residual signal, or minimizing the error between the weighted original signal and the contribution

of the filter memories (closed loop procedure). In both cases the information consists of a delay (lag) value and some coefficients (typically one or three). In our simulations we considered both cases in order to assess which approach, at the same bit-rate, produces better performance especially when considering high channel error rates. We also considered one or three taps and scalar or vector quantization. Vector quantization provides better performance in comparison to scalar quantization and the complexities involved, in this case, are comparable. This is because the scalar quantization approach requires the additional computations needed to ensure stability to the three taps filter, while by using VQ this step can be performed off-line when designing the codebook.

Concerning the vector length of the innovation signal we considered two cases: a short vector length of 20 samples and a longer vector length of 32 samples. The innovation codebooks used in our experiments have been designed by means of the keywords approach we described in [4]. This method allows to define two pulses codebooks starting from a small set of keywords and avoids the need to store all the codebook entries. Moreover, the sparse structure of the codebook allows a noticeable reduction in complexity.

In order to achieve different bit-rates we considered both the possibilities to use different codebook sizes or to avoid the transmission of the innovation index for a single vector in a frame. When the innovation is discarded, the excitation signal consists only of the adaptive codebook contribution and the scheme can be seen as a self-excited vocoder [5]. We noticed in our experiments that this approach is sometimes preferable, but when the number of "holes" in the innovation becomes too high a smaller codebook dimension but always transmitted, produces better performance.

This result prompt us to consider the possibility to dynamically allocate the available bit-rate for the innovation signal among the various vectors in a frame. For the determination of the optimum bit-allocation we used first the energy of the residual signal and then tried an optimum exhaustive search. As expected this last approach provided the best performance, but the increase in quality was judged to be only marginal and considering the heavy increase in complexity we decided to discard this possibility.

Concerning the scale factor we used the following quantization procedure: we compute the maximum scale factor among the vectors in a frame and quantize it with 6 bit, then each singular scale factor is normalized to the quantized value and coded with 3 bit. This procedure allows a saving in the number of bit needed, but has the drawback that the excitation parameters, with the exception of the first vector, have to be computed considering unquantized values. To cope with the degradation introduced by this suboptimal criterion, we introduced a reoptimization of the maximum quantized scale factor at the end of the frame.

Finally we considered the possibility to introduce a post-filter [6] at the receiver to improve the subjective quality of the output speech. We experimented that there is actually an improvement when considering a single encoding and decoding process. However, in the case of mobile to mobile communications, where the connection requires the cascade of two speech coders and consequently two post-filtering operations, the distortion introduced on the signal is unacceptable. Notice that we optimized the post-filter parameters in order to minimize the audible distortion when only one coder is considered. The distortion on the speech introduced by two post-filtering can be mitigated by modifying the parameters of the post-filter, but in this case its effectiveness is also reduced. It follows that a post-filtering technique can be effectively employed only if it is provided the facility, from the network, to signal at the receiver to insert or not the post-filter.

3. CODER SELECTION

As a result of our simulations we optimized four schemes whose characteristics are summarized in Tab.1. The schemes have been compared by means of informal listening tests on the basis of pair comparisons. The speech material included both male and female voices and we considered error-free and error distribution typical of the mobile channel corresponding to C/I of 10,7 and 4 dB. To consider channel errors we adopted a FEC scheme similar to the one used in the full-rate system, that is we divided the total number of bit into two classes: the first protected with a convolutional code of rate 1/2 and the second unprotected. The allocation of the bit into two classes was driven by the considerations reported in [7].

The results of the comparison under error-free condition was that coder n.2 was often preferred to coder n.4, while coder n.1 and n.3 were judged to be almost equivalent with a slight preference for n.3.

When considering the impact of errors, in every condition, but especially at 7 dB of C/I, the quality provided by coders n.2 and n.4 was judged to be too poor to balance the better performance in error-free condition. As a final decision we retained coder n.3 for further investigations.

4. CHANNEL ERROR SENSITIVITY

To optimize the channel coding algorithm, we investigated carefully the impact of errors on each bit of the speech coder frame. To do so we first computed the spectral distortion and the segmental SNR when corrupting systematically each bit in the bit-stream.

Fig. 2 shows the results of this experiment. Each line refers to one particular parameter and indicates the range of performance achieved modifying the bit used to represent the parameter itself. Looking at these figures it appears that there are some bit that can tolerate very low or even none protection.

In the next step we ordered the whole bits by using both the criteria and after some listenings it was clear that the spectral distortion correlates better with the subjective quality but not always. Thus we used a combination of both measures to produce a preliminary figure of bit importance. This curve has been then modified comparing, by means of extensive listen-

ings, each point with the adjacent. As a result of this work we were able to divide the output bit-stream into five classes of importance as represented in Tab. 2.

Then we evaluated, for each class, the bit error rate of random distributed errors, which determines an increase in the distortion of the output speech. These values are shown in the last row of Tab. 2. Considered separately, these values strongly depend on what the listener perceives to be the threshold for the beginning of the degradation due to errors, but our purpose was to determine the relative differences among the five classes in order to design a non-uniform protection scheme.

We have then exploited this information to test the performance of different FEC algorithms. We started from the convolutional code with rate 1/2 defined for the full-rate system and compared it with other codes obtained starting from rate 1/3 and using a proper puncturation to shape the achieved error probability according to the characteristic evaluated before.

Considering the number of bits available for protection, it appeared quite difficult to ensure a low BER to class 1. Moreover errors in this class produce a degradation in the output speech which is often worst than the one obtained using the same bits of the previous frame. Thus, in addition to the FEC scheme, we have used three parity bits to protect the bits in class 1 as in the full-rate system. These bits are used at the receiver to provide a Bad Frame Indicator (BFI) flag which allows to substitute the received corrupted parameters with an extrapolation of the ones received in the previous frames.

For this purpose we tested different alternatives and adopted different strategies during voiced and unvoiced periods. In particular we noticed that a smooth decreasing of the energy of the output signal is necessary when more than one speech frame has to be substituted and this can happen, in the GSM system, not only due to burst of errors, but also when one or more speech frames are stolen for signalling purposes.

Preliminary results on these experiments indicate that the BFI rate influences more the speech quality than the shape of the BER in the protected classes, suggesting to conceive more sophisticated techniques for the detection of high errors conditions.

5. CONCLUSIONS

The paper reported on the optimization of a speech coding scheme at 6.55 kbit/s to be used in a mobile environment. Several experiments have been conducted to assess the channel error sensitivity of the transmitted bits. The results of this investigation have been used to test different channel coding alternatives. The use of a BFI flag at the receiver to cope with strong burst errors has been found to be necessary to maintain acceptable performance. A post-filtering technique improves the reproduced speech quality but can only be used if adequate signalling is provided by the network.

The algorithm has a reasonable complexity and its implementation is under way on a AT&T DSP-32 processor.

ACKNOWLEDGMENTS

We wish to thank F. Muratore and V. Palestini for their work in the simulation of the different alternatives of the channel coding scheme.

REFERENCES

[1] B.S. Atal and M.R. Schroeder, "Stochastic coding of speech signals at very low bit rates", Proc. Int. Conf. Commun. , May 1984, part 2, pp. 1610-1613

[2] M. Copperi, D. Sereno, "Inproved LPC excitation based on pattern classification and perceptual criteria", Proc. Int. Conf. on Pattern Recognition, Montreal 1984, pp. 860-862

[3] R.V. Cox, W. B. Kleijn, and P. Kroon, "Robust CELP coders for noisy backgrounds and noisy channels", Proc. of ICASSP-89, pp. 739-742, Glasgow.

[4] L. Cellario, G. Ferraris and D. Sereno, "A 2 ms delay CELP coder", Proc. of ICASSP-89, paper S2.9, Glasgow.

[5] R.C. Rose and T.P. Barnwell III, "The self excited vocoder- An alternate approach to toll-quality at 4800 bps", Proc. of ICASSP-86, paper 9.6, Tokyo.

[6] N.S. Jayant, "ADPCM coding of speech with backward-adaptive algorithms for noise feedback and postfiltering", Proc. of ICASSP-87, paper 29.10, Dallas.

[7] M. Omologo and D. Sereno, "A Comparison Between CELP and MPLPC at 8 kbit/s for Mobile Communications", Proc. of ISCAS-88, pp.1823-1826, Helsinki.

# CODER	BIT-RATE (bit/s)	LT PARAMETER EVALUATION	VECTOR LEGTH (samples)	CODEBOOK SIZE (bit/vector)	PERCENTAGE OF BIT PROTECTED	SEGMENTAL SNR (dB)
1	6600	CLOSED-LOOP EVERY VECTOR	32	10*	62%	11.26
2	7300			10	46%	12.28
3	6550	OPEN-LOOP OVER THE FRAME	20	7	63%	11.46
4	7450			9	44%	12.69

* *Innovation computed only on 4 vectors out of 5*

Tab. 1 - Coders characteristics

CLASS	1				2				3		4		5
PARAMETERS	LSP	Σ	τ	μ	LSP	Σ	S	μ	σ	Inn	σ	LSP	Inn
NUMBER OF BITS	12	4	7	7	16	2	8	1	8	18	16	2	30
TOTAL	30				27				26		18		30
BER THRESHOLD	$5 \cdot 10^{-4}$				$5 \cdot 10^{-3}$				$8 \cdot 10^{-3}$		$1 \cdot 10^{-2}$		$5 \cdot 10^{-2}$

LSP – *Line spectrum pair coefficients (30 bit)*
Σ – *Maximum scale factor (6 bit)*
τ – *Long Term lag (7 bit)*
μ – *Long Term coefficients index (8 bit)*
σ – *Normalized scale factors (24 bit)*
Inn – *Innovation index (48 bit)*
S – *Sign of innovation vectors (8 bit)*

Tab. 2 – Classes of bit importance

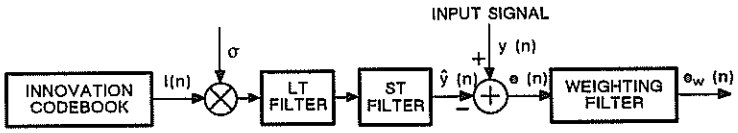


Fig. 1 – Block diagram of CELP transmitter

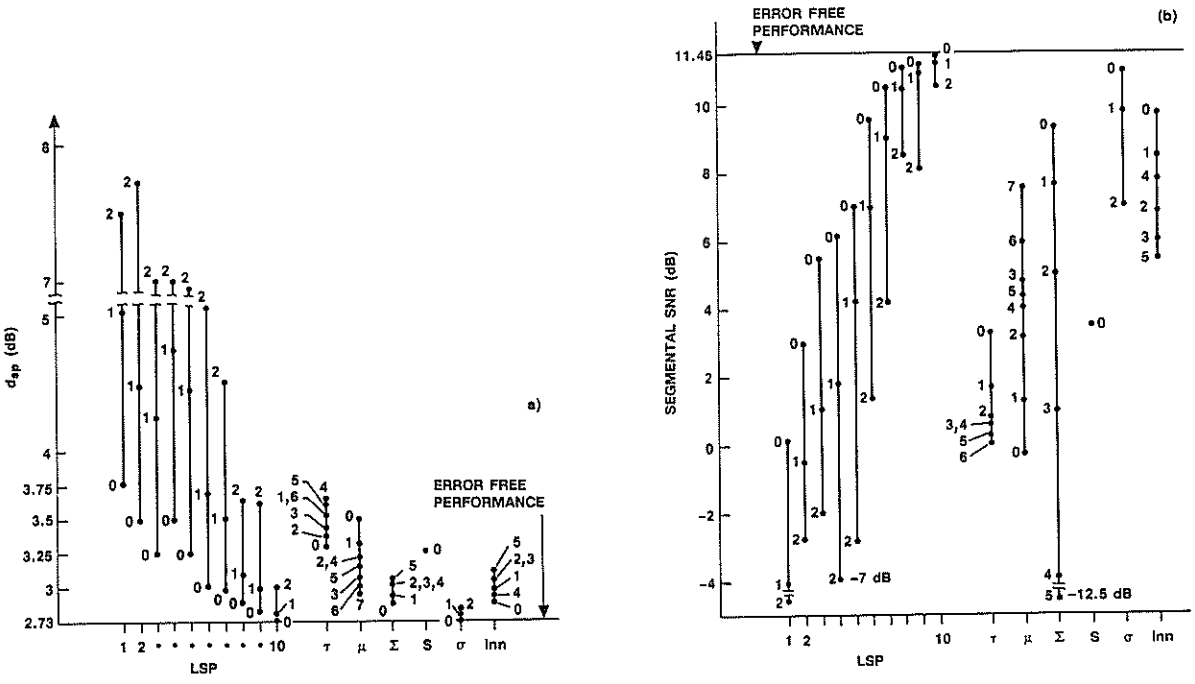


Fig. 2 – Spectral distortion (a) and segmental SNR (b) performance achieved when modifying one bit in the bit-stream. Each point identifies the bit of the papameter which has been corrupted. Bit 0 indicates the LSB

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

8 KBPS SPEECH CODER FOR DIGITAL CELLULAR MOBILE APPLICATION - PRINCIPAL AXIS EXTRACTING VECTOR EXCITATION CODING -

Y.Tanaka, T.Taniguchi, Y.Ohta, F.Amano, K.Utsugi, and Y.W.Sun

Speech Signal Processing Section, Fujitsu Laboratories Ltd.
1015 Kamikodanaka, Nakahara-ku, Kawasaki, 211 Japan

We propose Pricipal Axis Extracting Vector Excitation Coding (PAVXC) for mobile telephony applications. In this method, an efficient two-step codebook search and an improved excitation model using principal axis extraction of the stochastic code vector are employed. To cope with channel bit error, forward error correction by Reed-solomon code and parameter smoothing in combination with CRC error detection is employed. High quality and good robustness were obtained in our prototype codec.

1. INTRODUCTION

Progress in digital mobile radio systems is creating a great demand for a high-efficiency speech codec that can make efficient use of the radio frequency spectrum. Transmission rate should be about 8 kbps or less to secure the same channel capacity as current analog systems, but the coder should maintain a comparable reproduced speech quality. The coding algorithm which employs vector quantization of the excitation signal known as Code Excited LPC(CELP)[1] or Vector Excitation Coding(VXC)[2], is the most promising coding algorithm to meet the requirements. Research is currently directed at improving the reproduced speech quality[3] and reducing the amount of computation[4]. Moreover, in mobile telephony enviroments, a large amount of bit error will occur at deep attenuation of the received power, due to Rayleigh fading, which causes severe degradation in reproduced speech quality, so that robustness against transmission error is required for the speech coder for this application.

We propose Principal Axis Extracting Vector Excitation Coding (PAVXC)[5], in which an efficient codebook search and improved excitation modeling are employed. First, two-step codebook search is employed to reduce the computational complexity of the codebook search. In this method, instead of finding the optimum vector out of a large size codebook of 40 dimensions, the optimum vector for the first half and that of the second half are sequentially selected out of a small size codebook of 20 dimensions. Second, we introduced multiple vector excitation coding, using a linear combination of impulse and random gaussian vectors as an excitation signal in order to improve accurate excitation modeling, especially in transient and voiced regions.
To improve the robustness against channel error, forward error correction (FEC) and parameter smoothing techniques in combination with error detection are employed.

In the following section, the main features of the PAVXC speech coding algorithm are described. The error protection procedure employed in PAVXC is described in section 3. Then, we describe the implementation of our prototype codec in section 4, and give some experimental results under the simulated mobile channel conditions in section 5.

2. PRINCIPAL AXIS EXTRACTING VXC (PAVXC)

2.1. Multiple vector excitation

The VXC coding system represented by CELP generates the LPC excitation by adding the past excitation signal (pitch vector) and stochastic excitation signal. However, using only these two components to model the excitation signal for a voiced or transient speech period, which usually resembles a pulsive sequence, can be a problematic. Each pitch vector from the adaptive codebook has its origin in a random gaussian sequence read from the stochastic codebook. To generate a pulsive sequence from the random code vector, the phase of the random code vector must be shifted so that, when the vectors are added together, some of their respective noise components will be cancelled out. Therefore, at the transient between voiced and unvoiced periods, the accuracy of the excitation signal estimate is sometimes degraded, which results in the degradation of the reproduced speech quality.
To solve this problem, our PAVXC coder adds an impulse vector as the third component of the synthesized excitation, so that it models the LPC excitation signal using three vectors (shown in Figure 1.), instead of the conventional two. To determine the optimum reconstructed LPC excitation vector for 1/A', all the shape combinations of pitch, stochastic, and pulse vectors (P, C, U) must be searched, which will require more computation than synthesis using two vectors. In addition, the transmission information compression efficiency will decline, since information about the pulse phase and gain (U, r) must be transmitted.
However, if the pulse phase is made to uniquely correspond to the code phase by using the method described in the next section, the computation for the codebook search can be

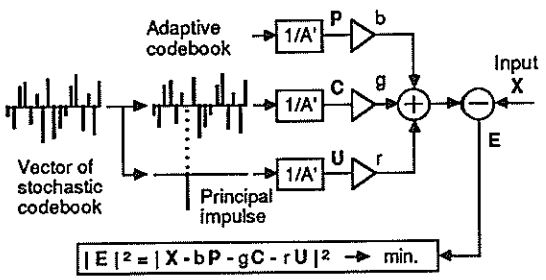


Figure 1. Multiple vector excitation

maintained at the level of two-vector synthesis, and the increase of transmission information can be suppressed to only the pulse gain (r). The advantage of three vector synthesis over the conventional two vector synthesis is that the characteristic of the excitation signal can be varied between random gaussian and impulsive by adjusting the gain ratio (g/r) of the stochastic and pulse vectors. The reproduced speech quality is subjectivery much better than that of conventional VXC, thanks especially to the improvement in the voiced regions.

2.2. Principal axis extraction

As mentioned in the previous section, in our PAVXC coder, the most significant pulse is extracted from each stochastic code vector, and its gain is coded separately from the gain for the original stochastic vector. Figure 2 shows the procedure for extracting this pulse vector from a stochastic code vector. Using a K-dimensional codebook, there are K types of corresponding pulse phase, which can be thought of as the vectors ($e^{(1)}$~$e^{(K)}$) representing the axis of the K-dimensional residual space. The most significant pulse, here, is the pulse that makes the greatest contribution to the code in the reproduced signal space after perceptual weighting of these K types of pulse. To select the optimum pulse, the angle (θ) between the perceptually weighted vector of each axis pulse ($U^{(i)}$: i = 1 to K) and the code vector (C) is evaluated using the inner-product given in equation (1), then the pulse vector in the direction with the minimum angle is extracted. Extracting the most significant impulse can also be referred to as principal axis extraction.

$$\cos^2\theta^{(i)} = (U^{(i)}, C)^2 / (|C|^2 |U^{(i)}|^2) \qquad (1)$$

This extraction method makes the main component pulse (U) uniquely correspond to the code vector (C), so that the decoder is able to determine the pulse shape from the stochastic code vector. This makes it unnecessary to transmit the shape of the principal axis vector.

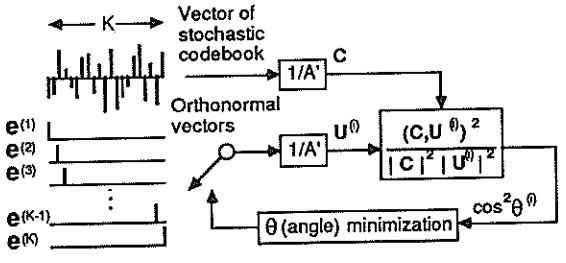


Figure 2. Principal axis extraction

2.3. Two-step codebook search

This section describes the computational complexity reduction scheme in the codebook search in the PAVXC coder. To reduce the codebook size while keeping an effective amount of codebook shape information, we propose a two-step search which halves the dimension of the codebook, using a 20-dimensional codebook with a size of 64 to determine separately the shapes of the first and last half of the 40-dimensional stochastic code vector, which is shown in Figure 3. In the first step, the shape (C_1) of the first half of the code vector is searched by evaluating the mean-square error (MSE) for a 40-dimensional code vector consisting of the 20-dimensional stochastic code vector in the first half, followed by a 20-dimensional zero vector. The shape and gains (C_1, b', and g') of each vector are determined to minimize the MSE of the error vector (E'). In the second step, the shape (C) of the entire code vector and gains (b,g) are decided to minimize the MSE using a 40-dimensional codebook consisting of the 20-dimensional stochastic codebook in the second half, preceded by the previously selected vector C_1.
This two-step search realizes a 40-dimensional codebook with an equivalent codebook size of 4096 entries with only a slight degradation in VQ performance, while reducing the amount of computation by 1/64 (N: 64/4096) and the amount of memory by 1/128 (KxN: (20x64)/(40x4096)).

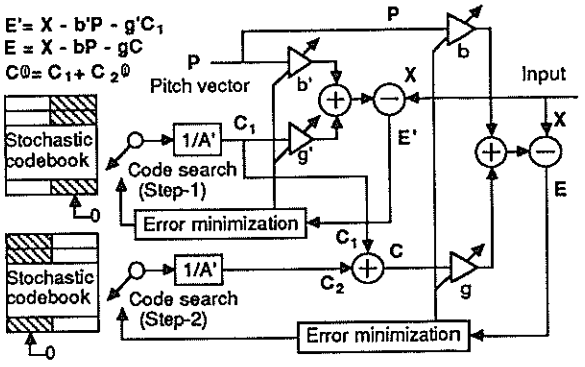


Figure 3. Two-step codebook search

2.4. Bit allocation

The bit allocation of PAVXC is shown in Table 1. LPC analysis is performed every 20 ms using 10th order autocorrelation. The pitch and codebook search are performed every 5 ms subframe. Each parameter is directly scalar quantized except for the code gains. We define a gain vector (g,r) whose elements consist of the stochastic code vector gain (g) and the pulse vector gain (r). This is first transformed to its power component ($s = g^2+r^2$) and the ratio (r/g), then each value is scalar quantized, which results in improvement of the quantization efficiency, because of correlation between the gain amplitudes, and also localizes the bit error sensitivity into the power component (s), which is more sensitive than the ratio (r/g), allowing an improvement in the channel error robustness.

Table 1. Transmission bit allocation

		Subframe #1	#2	#3	#4	bits/ frame
Adaptive codebook	Shape: P	7	7	7	7	28
	Gain: b	4	4	4	4	16
Stochastic codebook	Shape: C_1C_2	6 6	6 6	6 6	6 6	48
	Gain: s	5	5	5	5	20
Impulse	Gain: r/g	3	3	3	3	12
LPC parameters		5, 5, 4, 4, 4, 3, 3, 3, 2, 2				35
Synchronization		1				1
Total						160

3. ERROR PROTECTION

In the mobile radio transmission enviroment, severe bit error will occur due to Rayleigh fading. To mitigate the effect of channel errors, PAVXC employs forward error correction and parameter smoothing techniques, combined with CRC error detection.

3.1. Forward error correction and detection

Transmission data consist of several different parameters. The bit error sensitivity of each parameter and each bit is different, so that to use bit-selective error protection according to the bit error sensitivity is an effective scheme for speech coding[6]. In a long burst error at deep attenuation of the received power, an entire frame of data would be corrupted and can not be corrected by any powerful FEC. In this case, error detection by CRC coding is more effective, when combined with the parameter smoothing process which follows it.

Figure 4(a)(b) show the bit error sensitivity of 8 kbps PAVXC in S/Nseg and LPC cepstrum distance(CD) respectively. We perturbed a particular bit in each frame. It has to be noted that each parameter is gray coded, that is the same condition in the actual implementation. As shown in these figure, errors in low order LPC coefficients (reflection coefficients) cause large distortion, especially in the spectrum, and produce speech with a whistling sound. An incorrect

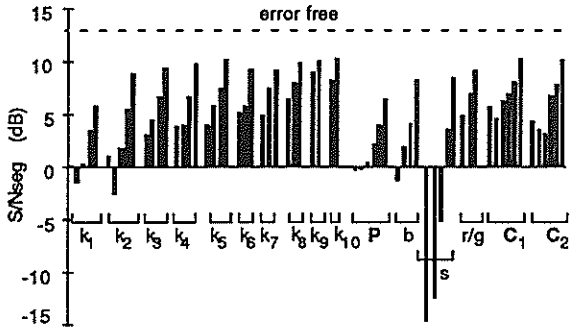


Figure 4(a). Error sensitivity in S/Nseg

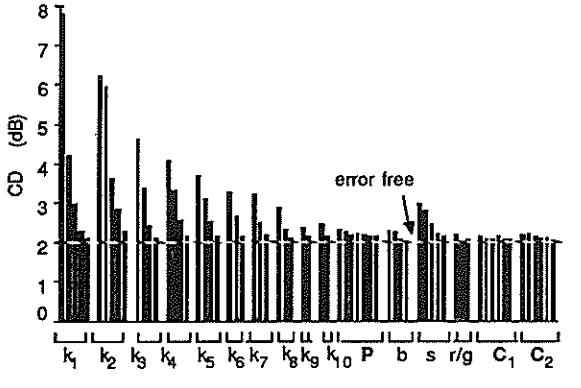


Figure 4(b). Error sensitivity in CD

adaptive codebook index value causes a noticeable distortion in S/Nseg in spite of a small spectral distortion, because the main components of this error are phase and waveform distortion.The reprodeced speech sounds rough. Errors in gain vector power (s) produce an abrupt increase in signal magnitude, which makes listening by handset very uncomfortable, while the ratio of the gain vector (r/g) is less sensitive to errors. The codebook in PAVXC employs a one-sample overlapped structure to reduce the perceptual weighting computation for each code vector, so that the error sensitivity in the upper bits of the codebook index is slightly larger than that of the lower bits, but still has low error sensitivity. According to this error sensitivity profile, we found that the gain vector power (s) is the most sensitive parameter to error, followed by the LPC coefficients (k), the adaptive codebook gain (b), the adaptive codebook index (P), the gain vector ratio (r/g), and the codebook index (C1,C2).

We have classified the transmission bits into three ranks according to the error sensitivity profile, in which rank-1 is the most significant bit group and rank-3 is the least significant bit group, as shown in Table 2. The bit groups of rank-1 and rank-2 are protected by (31,25,7) and (31,29,3) Reed-solomon codes on GF(2^5), respectively. The bit group of rank-3 is not protected by any FEC. As shown in this table, most of the parameters in rank-1 and rank-2 have 4-bit CRC check bits to detect the occurrence of incorrectable error in the parameter. The CRC coding of separate parameters enables individual control of smoothing of each parameter, and prevention of error enhancement by FEC miscorrection.

Table 2. Classification of protection bits

rank	Protection bits		CRC	Applied FEC
rank-1	s	b4 - b2	CRC-4	RS(31,25)
	r/g	b2 - b0	none	
	b (#1,#3)	b3 - b1	CRC-4	
	b (#2,#4)	b3 - b1	CRC-4	
	P	b6 - b0	CRC-4	
rank-2	k1 - k4	all bits	CRC-4	RS(31,29)
	k5 - k10	all bits	CRC-4	
rank-3	s	b1 - b0	none	No FEC
	C_1	b5 - b0	none	
	C_2	b5 - b0	none	

3.2. Error recovery

To maintain the reproduced speech quality on the fading channel, in which FEC has little impact, recovery procedures of incorrect parameters such as replacement, smoothing and muting are indispensable[7]. Generally, these recovery procedures cause quality degradation in error free conditions, so that the indication of error occurence from the above-mentioned CRC codes is used to control these recovery process in PAVXC. We describe the recovery strategy of each parameter respectively.

LPC coefficient : 4-bit CRC codes are added to the low order coefficient set (k1~k4) and the high order coefficient set (k5~k10) individually. Only the erroneous coefficient set is replaced by the corresponding set of the previous frame.

Codebook gain : A large magnitude change in the decoded code gain due to bit error causes click and pop noise in the reproduced speech. We restricted the transition of quantization code indices to within 1 in the erroneous frame. This method realizes nonlinear smoothing according to the magnitude of the gain, since a nonlinear quantizer is employed.

Adaptive codebook gain : This gain tends to exhibit random behavior, which degrades the smoothing performance of this parameter. To improve this smoothing property, 1st and 3rd subframe's coefficients and 2nd and 4th subframe's coefficients are separately checked by different CRC codes. The erroneous coefficient set is interpolated and extrapolated using the correct coefficient set if error was detected in only one set. If errors are detected in both coefficient sets, all coefficients are replaced by the last subframe's value of the previous frame, in which the coefficient is truncated to 1.0 when it exeeds 1.0, to stabilize the pitch filter.

Adaptive codebook index : The erroneous frame's values are simply replaced by the value from the last subframe of the previous frame.

The codebook indices and the code gain ratio (r/g) are not protected by any recovery procedures because of their lower error sensitivity. Gray coding is applied to all parameters as redundancy-less error protection.

4. IMPLEMENTATION

A speech codec for mobile telephony must be implemented in a small size and at low power consumption. However, in our prototype codec, a redundant configuration was introduced for the easy debuging and algorithm modification. The encoder is implemented using four AT&T 32C DSPs for speech coding and a microprocessor for FEC. The decoder is implemented using one DSP and one microprocessor. The required resources to implement PAVXC, including FEC, are 16.7 Mips operations, 6.6 kword RAM, and 6.3 kword ROM, all of which is easy to implement in one high-speed fixed-point DSP.

5. PERFORMANCE

Table 3 shows the simulation results for PAVXC and VXC. Short Japanese sentences, spoken by three male and three female speakers, were used in the evaluation. In the VXC, the computationary-reduced two-step codebook search degraded the S/Nseg by 0.9 dB from that of the conventional full-search method (K=40,N=4096). In PAVXC, however, adding a pulse as one of the excitation signals, we get a S/Nseg improvement of 1.0 dB, with an additional transmission rate of 1.0 kbps required to transmit the pulse amplitude information. In listening test, the reconstructed speech produced by 8 kbps PAVXC, compared to that produced by conventional VXC, was confirmed to be noticeably clearer, in spite of the comparable S/Nseg value.

We have evaluated our prototype codec on a simulated mobile channel conditions. We have simulated a Rayleigh fading (40 Hz fading pitch) channel using two-channel space diversity. At the error rate of 1%, noticeable degradation was not observed in the informal listening test. In the severe channel condition at the error rate of 3%, the degradation was audible, but the large pop noise was effectively suppressed by the recovery procedures.

Table 3. Simulation results

	Conventional VXC (K=40, N=4096)	2-step search VXC (K=20, N=64)	PAVXC (K=20, N=64)
S/Nseg	14.4	13.5	14.5
S/N	13.5	12.5	13.6
LPC-CD	2.14	2.28	2.10

(CD: Cepstrum distance, Codebook dimension: K, size: N)

6. CONCLUSION

We proposed Principal Axis Extracting Vector Excitation coding (PAVXC) for mobile telephony applications. In this method, the most significant pulse is extracted from the stochastic code vector, and its gain is coded independently from the gain of the stochastic code vector. This improves the accuracy of the excitation estimate, and the quality of the reproduced speech. A two-step stochastic codebook search is employed, which results in a large-scale reduction in the computational complexity. Forward error correction using Reed-solomon codes and parameter smoothing techniques, in combination with CRC error detection, improved the reproduced speech quality even in severe channel error conditions. Our prototype codec was implemented using five DSPs. However, the required computation to realize PAVXC is 16.7 Mips, which will be implemented in a one-chip high-speed fixed-point DSP.

ACKNOWLEDGEMENTS

The authors wish to thank K.Murano, S.Unagami, and M.Johnson of Fujitsu Laboratories Ltd. for their valuable suggestions and consistent encouragement.

REFERENCES

[1] M.R.Schroeder, B.S.Atal, "Code-Excited Linear Prediction(CELP): High Quality Speech at Very Low Bit Rates", Proc. ICASSP'85, pp.937-940, Mar.1985.
[2] G.Davidson, A.Gersho, "Complexity Reduction Methods for Vector Excitation Coding", Proc. ICASSP'86, pp.3055-3058, Apr.1986.
[3] P.Kroon, B.S.Atal, "Strategies for Improving the Performance of CELP Coders at Low Bit Rates", Proc. ICASSP'88, pp.151-154, Apr. 1988.
[4] J-P.Adoul, P.Mabilleau, M.Delprat, S.Morissette,"Fast CELP Coding Based on Algebraic Codes", Proc. ICASSP'87, pp.1957-1960, Apr. 1987
[5] T.Taniguchi, Y.Tanaka, A.Sasama, Y.Ohta, "Principal Axis Extracting Vector Excitation Coding: High Quality Speech at 8 kb/s", Proc. ICASSP'90, S4b.8, Apr.1990.
[6] H.Suda, T.Miki, "An Error Protected 16 kbit/s Voice Transmission for Land Mobile Channel", IEEE journal on SAC, pp.346-352, Vol.6, No.2, Feb. 1988.
[7] R.V.Cox, W.B.Kleijn, P.Kroon, "Robust CELP Coders for Noisy Backgrounds and Noisy Channels", Proc. ICASSP'89, pp.739-742, Apr. 1989.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

FAST PITCH TRACKING ALGORITHM FOR LTP BASED SPEECH CODERS

C.Galand
IBM Thomas J.Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

M.Rosso, C.Arnaud
IBM Laboratory
06610 La Gaude
France

This paper presents an algorithm which allows to decrease the complexity requirement of the long-term prediction (LTP) analysis of speech coders. This algorithm has been applied to the GSM coder and has been showed to maintain the speech quality at the specified level. The proposed algorithm is based on a low cost and adaptive evaluation of a rough value for the pitch period. This estimation is then used to compute the exact period with a reduced number of cross correlation products. The complexity requirement of the speech coder is then reduced from 4 to 2.4 MIPS, which allows us to process 30 full duplex channels on a single card server which is based on a master/slave arrangement of high throughput digital signal processors.

Introduction

The major requirement for a voice data product is to provide a good speech quality. This is why such products usually support standards for the encoding of speech signals. Another requirement is the cost per port. When considering a single server to compress a large number of input voice ports, the system granularity is an important factor of cost. In other words, the greater the number of ports per single card adaptor the lower the cost per port.

In our study, we have considered a 30 channel-granularity, corresponding to a full CEPT (2,048 Mbps) as the design point of a speech compression server.

The standard recently adopted by the GSM for the encoding of speech in the European cellular radio network specifies a compression rate at 13 kbps while maintaining the speech quality at an acceptable level. It however requires a processing capability of 4 MIPS (Millions of Instructions Per Second) per voice channel. This requirement, along with the associated buffer requirements does not allow a cost effective implementation of a 30 channel server on a single card.

This paper presents a modification of the GSM algorithm which allows to decrease the requirement to 2.4 MIPS per voice port while maintaining the speech quality at the specified level. The modifications of the GSM algorithm mainly concern the computation of the pitch period to be used in the long-term prediction loop. The proposed algorithm is based on a low cost and adaptive evaluation of a rough value for the pitch period. This estimate is then used to compute the exact period with a reduced number of cross correlation products.

The paper is organized as follows. In next section, we briefly remind the GSM algorithm and its processing requirements. We then describe the fast pitch tracking algorithm, and we report formal voice quality tests. In last section we show that the modified algorithm can be used to process 30 full duplex channels on a single card server which is based on a master/slave arrangement of high throughput digital signal processors.

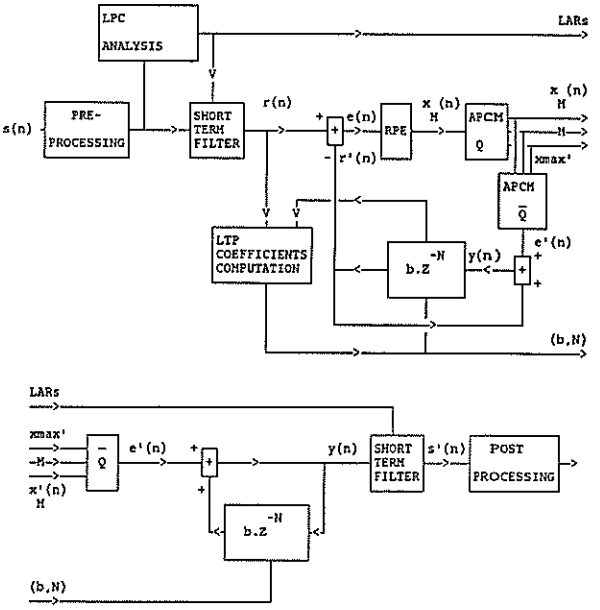
1 - GSM algorithm

In this section, we remind the operation of the 13 kbps RPE/LTP algorithm specified by the CEPT Group Special Mobile (GSM) in his recommendation GSM 06.10 dated July 1988 for application to the Pan European Cellular Radio System [1], and we give the processing load for each function, with an emphasis on the LTP function.

Algorithm overview

Fig.1 represents the general block diagram of the RPE/LTP encoder. In the pre-processing block, offset compensation is applied to prevent a DC component being translated into an annoying side tone by the process of high frequency regeneration in the decoder.

Fig.1 GSM coder general block diagram



In the segmentation buffer, the speech signal is divided into non overlapping segments having a length of 20 ms (160 samples). A LPC analysis is performed for each segment by calculating 8 reflection coefficients using the Schur recursion algorithm. The reflection coefficients are then converted to log area ratio (LAR) coefficients which are coded and transmitted. The LAR's are interpolated linearly within a transition period of 5 ms to avoid spurious transients, and then used to produce 160 samples of error (or residual) signal r(n) with a lattice filter.

The LTP loop is used to compute the estimated r'(n) of the error signal r(n) from the reconstructed excitation signal y(n). The LTP filter is characterized by the gain b and the delay N which are calculated every 5 ms (40 samples). Each segment of 160 samples of the residual r(n) is subdivided into four sub-segments. For each sub-segment, the cross-correlation function between the signal r(n) and the previously reconstructed excitation y(n) is computed as follows:

$$(1) \quad R(n) = \sum_{i=0}^{39} r(i).y(i-n) \;\; ; \;\; n = 40,...,120.$$

The optimum delay N_i for each sub-block (i = 1,...,4) is determined as the maximum value of the cross-correlation function:

$$(2) \quad R(N_i) = \max(|R(n)|; n = 40,...,120)$$

The corresponding gain b_i is calculated by:

$$(3) b_i = \frac{R(N_i)}{\sum_{n=0}^{39} y^2(n-N_i)}$$

The LTP parameters b_i and N_i are encoded with 2 and 7 bits respectively. Note that, due to the sub-block processing by RPE, the cross-correlation lag n takes values from 40 to 120. As a result, the value N_i does not correspond necessarily to a pitch period of the signal, but at least to a multiple of this period.

A FIR block filter algorithm is applied to each sub-segment of 40 samples of the second residual signal e(n). For notational convenience, the block filtered (smoothed) version of each sub-segment is denoted $x(n), n = 0,...,39$.

In the next step, the signal $x(n)$ is down-sampled by a ratio of 3, resulting in three interleaved sequences of lengths 14,13,13, which are divided again into 4 sequences x_m of length 13: $x_m(i) = x(m + 3i)$ with $m = 0,1,2,3$:*fe.*, $\wedge$:*f.i* = 0,...,12. The optimum candidate sequence $x_M(n)$ is selected as being the one with the maximum energy.

Finally, the selected RPE-sequence $x_M(n)$ is quantized by block adaptive PCM (APCM). Each block of 13 samples is normalized by its block maximum absolute value X_{max}. The samples are then quantized uniformly with 3 bits, the block maximum X_{max} is encoded logarithmically with 6 bits, and the grid position M is coded with 2 bits. The 3-bit quantized samples are denoted $x'_M(n)$, $n = 0,12$. The 6-bit quantized maximum is denoted X'_{max}.

At the decoder, the RPE parameters are decoded and used to reconstruct the excitation e'(n) of the long-term synthesis filter which produces the excitation signal y(n) for the short-term synthesis filter. The sample rate of the re normalized RPE samples is increased by a factor of 3 by inserting zero samples and by placing the nonzero samples in the correct temporal grid position M. The so-obtained signal e'(n) is used to excite the long term prediction filter. The STP excitation signal y(n) is then forwarded to the short-term filter. The filtered speech signal s'(n) is finally sent to the post-processing block which includes proper re-scaling.

Computational complexity

The GSM algorithm has been implemented on several 16 bit DSP's [1]. The complexity of the implementation slightly varies from one DSP to the other but requires about 4 MIPS. Table I sums up for each function, the cycle requirements of the GSM algorithm implemented on the IBM DSP. The data are given in number of cycles per 20 ms blocks, and in corresponding MIPS.

Table I: GSM processing load

	Cycles	MIPS
Analysis		
Pre-processing	3500	
LP analysis	7000	
STP analysis filter	8000	
LTP analysis	24000	
RPE and LTP filters	16500	
Sub-total	59000	2.95
Synthesis		
RPE and LTP filter	6000	
STP synthesis filter	11000	
Post-processing	1000	
Sub-total	18000	0.90
Total	**77000**	**3.85**

One can see from Table I that the analysis requires about 3 MIPS and the synthesis about 1 MIPS. Two of the most consuming functions are the LTP analysis (1.2 MIPS) and the LPC filters (0.95 MIPS). For the reasons we mentioned, we searched a mean to reduce this computational load while strictly keeping the speech quality at the specified level, and we found out that one could reach this goal by considering the following approach:

- Use a fast pitch tracking algorithm
- Use a direct form implementation of the LPC filters.

These two points are discussed in next section.

2 - Modifications of the GSM algorithm

Fast Pitch Tracking algorithm

The idea is to take advantage of the relatively slow variations of the pitch period in sustained voiced sounds, to make a rough estimate of this period from the value determined in the recent past. Once this estimation has been made, one merely apply a fine search of the period in a reduced domain around this rough estimation.

The rationale behind this idea is that the pitch loop has been shown to provide a significant gain only in sustained voiced

sounds. In transient sounds, the pitch loop -even with the correct values of the pitch parameters computed from relations (1) to (3)- does not have a high gain. In this case, a rough estimation of these parameters does not degrade the performances.

The algorithm works as follows. The cross-correlation function between the first 40 sample sub-segment of r(n) and the previously reconstructed excitation y(n) is computed according to relation (1). The optimum delay N_1 for the first sub-block is determined according to relation (2), and the corresponding gain b is calculated from relation (3).

For the next three sub-blocks, the optimum delays N_2 , N_3 , and N_4 are not determined according to relations (1) and (2). Rather a rough value of the delay at each sub-block is estimated from the delay at the previous sub-block, as following.

For the evaluation of N_2 the cross-correlation function R(n) as given by (1) is evaluated only for the lag n varying in the following intervals:

$$n = -(N_1 + 5),...,(N_1 + 5)$$

(4)

$$n = -(2N_1 + 5),...,(2N_1 + 5)$$

$$n = -(\frac{N_1}{2} + 5),...,(\frac{N_1}{2} + 5)$$

with the constraint 40 < n < 120

Then, the maximum value of $R(n)$ is detected, which gives the second sub-block optimum lag N_2 .

For the third sub-block, relation (1) is evaluated for the values of n given by relations (4) with N_2 replacing N_1 , and the maximum value of the R(n) function gives the optimum lag N_3 .

Finally, the same algorithm is again applied with N_3 replacing N_1 in relations (4), to give the optimum lag N_4 .

As we will see in next section, this algorithm has been shown to perform well as far as the speech quality is concerned. Let's now evaluate the complexity reduction it can afford.

From Table I, one can note that the determination of the LTP parameters using relations (1) to (3) requires an implementation complexity of 1.2 MIPS. The algorithmic complexity is of .64 MIPS, and mainly corresponds to 320 evaluations of relation (1) per block of 20 ms, that is 12800 multiply/additions or cycles. The difference (1.2-0.64 MIPS) corresponds to the evaluation of relations (2) and (3), and to data monitoring such as delay-line updating.

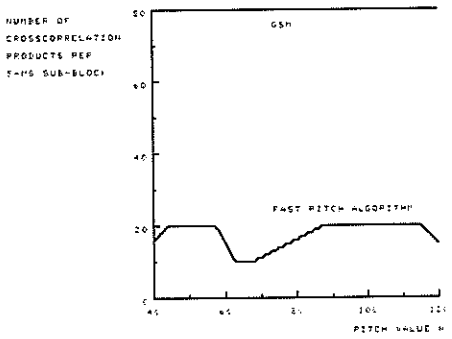
The modified search procedure as defined above allows relation (1) to be evaluated less often per 20 ms block. Roughly speaking, 80 times for the first sub-block and at most 33 times for the following sub-blocks.

In fact, this worst case is never reached because the multiples and sub-multiples of the pitch period as computed by relations (4) do not necessarily fall within the authorized range of variations (40,...,120). For example, if $N_1 = 48$, then $N_1/2 = 24$, and is not considered.

Fig.2 shows, for each value of N in the interval (40,120), the number of times relation (1) should be evaluated. One can see that this number is always less than 20, and can be as small as 10, with an average value of 17.

One can therefore estimate the complexity of the modified search procedure to 80 + 3 x 17 = 131 evaluations of relation (1). This shows a reduction in the complexity of about 60%. In fact, direct measurement on the microcode showed that the relative reduction in the implementation complexity is slightly lower (55%), resulting in a 0.65 MIPS saving.

Fig.2: Number of cross-correlation products as a function of N.



Direct Form filtering

The GSM algorithm specifies a lattice implementation for the LPC filters. In a DSP based implementation, this implementation of the filters is not efficient, since it requires several cycles per tap (6 to 7). This is why we decided to replace it by an equivalent direct form filtering which can be implemented with only one cycle per tap on a multiply/accumulator structure. This choice allowed us to save 0.8 MIPS.

Implementation summary

The current implementation of the modified RPE/LTP coder on the IBM signal processor [2] is summarized in Table II.

Table II: Modified RPE/LTP requirements

		DATA (HW)	CYCLES (MIPS)	PROG (K)
CODER	Input buffer	160	1.90	1.8
	Output buffer	16		
	State variables	143		
DECODER	Input buffer	16	0.50	0.6
	Output buffer	160		
	State variables	136		
COMMON	Interrupt area	7		
	I/O pointers	2		
	Working area	239		
	Constants	177		
TOTAL		1056	2.40	2.3

Speech quality

The speech quality of the modified RPE/LTP coder was evaluated by listening tests. The speech material used for the test consisted in 10 French sentences (5 male and 5 female voices). These samples were obtained by digitizing, through a 8 bit A-Law A/D converter, signals captured by a carbon microphone. Each of these test sentences was processed under simulation by the modified RPE/LTP coder, the GSM coder, and a reference 24 dB MNRU (speech artificially corrupted with a Modulated Noise Reference Unit). Then all the possible pairs were compared by standard pair-to-pair comparison tests. The pairs of sentences were randomly played to 10 listeners who were asked to mark their preference. The preference was averaged over all listeners and speakers.

On the average, 28% of the preference scores went to the modified RPE/LTP coder, 25% to the GSM coder, 31% to the MNRU, and 16% of the answers indicated no preference. From this evaluation test, we conclude that the modified RPE/LTP coder provides the same level of quality as the GSM coder.

These results were further confirmed by using prototype equipments in conversational full-duplex tests. The prototype could on request instantaneously switch between the GSM and the modified GSM algorithm. The persons were asked to mark their preference after a 5 minute conversation including several switches from one to the other algorithm. The results showed that no one was able to tell the difference between the GSM algorithm and the modified algorithm.

3 - Voice server

The reduction from 4 MIPS to 2.4 MIPS of the processing requirement for a complete speech coder allowed us to consider a single card prototype card capable of processing a full duplex CEPT (30 channels).

Figure 3 shows the architecture of the Voice Server Card. It is organized around a master digital signal processor (DSP) chip S0, and five slave digital signal processor chips S1,S2,S3,S4,S5. The slave chips, as well as the master data and instruction memories, are plugged on the data bus of the master chip.

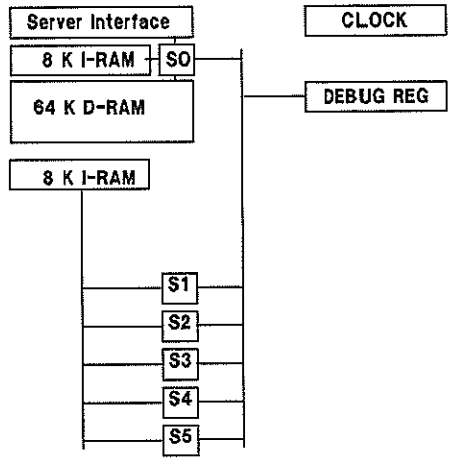
We used an IBM proprietary chip including a DSP macro implementing the architecture given in [2], a data RAM, an instruction RAM coupled with an instruction cache, and a communications port. The communications port is used on one hand to connect the slave processors to the master processor, and on the other hand to connect the master processor to the server interface.

Each of the slave chips implements the speech compression algorithms, and is monitored by the master DSP on a task basis. The signal processing algorithms are stored in an 8K common instruction RAM, which is shared between the slave DSP's thanks to the on-chip cache and to an external arbitrator.

The master signal processor has a 64 K data RAM and an 8 K instruction RAM.

The DSP cycle time is 50 ns. Therefore, the card has a 100 MIPS capability for the signal processing operations, and 20 MIPS for the data management. The throughput of the data bus is 320 Mbps.

Fig.3: Architecture of the voice server card



Conclusion

In this paper, we have presented an algorithm to decrease the complexity requirement of the long-term prediction (LTP) analysis of speech coders. This algorithm has been applied to the GSM coder and has been showed to maintain the speech quality at the specified level. In addition, we have changed the implementation of the lattice filters to a direct form implementation. As a result, the complexity requirement of the speech coder has been reduced from 4 to 2.4 MIPS, which allowed us to consider a single card server to process 30 full duplex voice channels. The card has a throughput of 100 MIPS and is loaded at 72%.

References

[1]K.Hellwig, P.Vary, D.Massaloux, JP.Petit, C.Galand, M.Rosso: 'Speech codec for the European mobile radio system'. pp.1065-1069, GLOBECOM, Dallas 1989.

[2]J.P.Beraud: 'Signal processor chip implementation' IBM Journal of Research and Development, Vol.29, N.2, April 1985.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ON THE USE OF ENERGY INFORMATION FOR SPEECH RECOGNITION USING HMM

Antonio M. Peinado *, Padma Ramesh, and David B. Roe

AT&T Bell Laboratories
Speech Research Department
Murray Hill, NJ

Abstract - *In the last several years, algorithms for high performance connected word recognition have been developed, achieving error rates below 3% (unknown length). For applications such as voice dialing of telephone numbers or automatic credit card entry, even higher accuracies would be required. One possibility is to incorporate new features (more information), such as energy and delta energy, to obtain better performance. In this paper, we investigate the use of energy and delta energy as additional features to improve the accuracy of speech recognition systems. Furthermore, we discuss ways of incorporating these and other features in a statistically unified manner into the Hidden Markov Model.*

1. Introduction

After several years of research, the HMM technique has become a well known method for speech recognition. Continuous density hidden Markov models have been successfully used for isolated and connected speech recognition [1] at AT&T Bell Laboratories, yielding accuracies of 97% for connected digits. It is necessary to improve the performance of connected digit recognition systems even more, for applications such as voice dialing of telephone numbers and automatic credit card entry.
Rabiner et al [1] have achieved an error rate of 2.94% for connected digit recognition (unknown length, 4 models per digit). This system uses delta cepstrum information along with cepstrum. Thus, the augmented feature vector characterizes both the short-time spectrum and the short-time spectrum derivative, reducing effectively the recognition error rate. This system is described in section 2.
Frame energy has been shown to improve the accuracy of isolated word recognition systems [2]. Recognition performance of connected digit recognition algorithms can be improved by using energy histograms [1]. In section 3, we present a way of including energy and delta energy of the spoken utterance as additional features for recognition. In fact, we wish to be able to incorporate these and other new features, in a statistically unified way. One of the problems of adding new features is the choice of the distance measure used for pattern matching. In section 4, we propose a new distance measure for these features. Section 5 discusses the experiments and the results obtained on the Texas Instruments connected digit database.

* Dpt. Electronica, Universidad de Granada
Granada, Spain

2. The HMM connected digit recognizer

This system uses the connected-word, continuous density, multiple Gaussian mixture, HMM connected word recognition algorithm [1]. Basically, the recognizer has three main functions:

1) Feature analysis: the speech signal is converted to LPC cepstrum and LPC delta cepstrum, plus energy and delta energy.
2) Level building pattern matching: the input sequence of feature vectors is matched with the set of Hidden Markov Models, getting a list of candidate strings (those with highest string probabilities).
3) Postprocessing: the most likely string is obtained from the list of candidates, using duration information.

This recognition system is trained by a segmental k-means algorithm, which provides a good estimation of the HMM parameters for each word. In addition, Level building algorithm is used in the training to provide a segmentation of each training sentence into individual words.

3. Feature analysis

The choice of features is a critical issue in speech recognition and in speech processing. A better characterization of the speech signal can provide improved recognition and coding. Finding the best set of features to represent the speech signal, is a key problem. Also, for template based systems, the distance measure to be used with the chosen features must be specified.
Rabiner et al [1] have shown that a feature vector consisting of the weighted cepstrum and the weighted delta cepstrum using Euclidean distance measure for clustering, yields high performance for connected digit recognition. Based on this work, we propose the following feature analysis:

1) Preemphasis: the digitized speech signal (sampled at 6.67 kHz) is processed by a preemphasis filter to

spectrally flatten out the signal.

2) Windowing: each feature vector is obtained from segments of N samples of speech. Consecutive frames are overlapped by M samples (we use N=300 and M=200) and Hamming window is used for smoothing the effects of finite frame size.

3) Autocorrelation and LPC/cepstral analysis: p+1 autocorrelation coefficients are first computed. From these, the p LPC coefficients are obtained by Levinson or Durbin recursion (p is the LPC order; we use p=8). Q cepstral coefficients are then computed from the LPC spectrum (we use Q=12). These coefficients are multiplied by a liftering window $w(m)$, to obtain a liftered cepstrum,

$$\tilde{c}_l(m) = c_l(m) \cdot w(m) \tag{1}$$

where

$$w(m) = \left[1 + \frac{Q}{2} \sin\left[\frac{\pi m}{Q} \right] \right] \quad 1 \le m \le Q \tag{2}$$

4) Delta cepstrum: the time derivative of the weighted cepstrum is approximated as follows,

$$\Delta\tilde{c}_l(m) = \left[\sum_{k=-K}^{K} k \cdot \tilde{c}_{l-k}(m) \right] \cdot G, \quad 1 \le m \le Q \tag{3}$$

where G the gain, is chosen such that the variances of $\tilde{c}_l(m)$ and $\Delta\tilde{c}_l(m)$ are about the same (in our system G=0.375).

5) Energy: the energy of frame l is computed in dB, from the zeroth autocorrelation coefficient, $V_l(0)$, and then normalized by the peak energy in the string,

$$E_l = 10 \cdot \log_{10}(V_l(0)) - 10 \cdot \log_{10}\left[\max_{1 \le l \le T} \left[V_t(0) \right] \right] \tag{4}$$

where T is the total number of frames in the string.

6) Delta energy: It is the time derivative of the energy contour and is obtained by the following equation.

$$\Delta E_l = \frac{\sum_{k=-K}^{K} k \cdot E_{l-k}}{\sum_{k=-K}^{K} k^2} \tag{5}$$

We use K=2 for both delta cepstrum and delta energy.

This analysis differs from the analysis performed in [1] in the two new features, energy and delta energy. It has been shown by Rabiner et al [2], that frame energy is useful for isolated word recognition. It was shown in [1], that energy histograms, in addition to the features of cepstral and delta cepstral coefficients, improved the recognition accuracy. Delta energy has already been successfully used in other research efforts at CMU, TI and NTT.
In this work, we use two different types of feature vectors:

1) A vector with D=26 coefficients: weighted cepstrum, weighted delta cepstrum, energy and delta energy,

$$\mathbf{O}_l = \left\{ \tilde{c}_l(m), \Delta\tilde{c}_l(m), E_l, \Delta E_l \right\} \tag{6}$$

2) A vector with D=25 features: weighted cepstrum, weighted delta cepstrum and delta energy,

$$\mathbf{O}_l = \left\{ \tilde{c}_l(m), \Delta\tilde{c}_l(m), \Delta E_l \right\} \tag{7}$$

4. Distance measure for HMM training

The basic problem in recognition is that we do not know *a priori* the model that best matches a specific word, so it's necessary to train it from speech data. The new features in the feature vector need to be modeled during the training process, in order to get upgraded models.
The HMM training consists of obtaining an *optimum* set of parameters for a word. Each state s_j is characterized by the following subset of parameters:

1) A state transition vector a_j.
2) A state observation density $b_j(\mathbf{O})$.
3) A state energy histogram, $p_j(\varepsilon)$.
4) A state duration histogram, $p_j(\tau)$.

To generate the model, it is, obviously, necessary to cluster the training data at some point. In the continuous HMM case, this is done in the computation of the observation densities. Our next problem is the choice of the distance measure for clustering. We have to choose a distance measure that best matches two acoustic events for the features used. Besides, in our case, we have the problem of including features of different types, with cepstrum, delta cepstrum, energy and delta energy.
In [3], Tohkura proposes a weighted cepstral distance measure,

$$d_{MCEP} = \sum_{i=1}^{Q} w(i) \left[c_t(i) - c_r(i) \right]^2 \tag{8a}$$

$$w(i) = 1/\sigma_i^2 \tag{8b}$$

where σ_i^2 is the variance of ith cepstral coefficient, $c_t(i)$ and $c_r(i)$ are ith cepstral coefficients of a test and a reference frame, respectively, and Q is the number of cepstral coefficients. This distance has the following advantages:

1) The contributions to the distance of the various cepstral coefficients are at the same level.
2) Also, the distance can be considered as windowing in the quefrency domain. If we take into account that σ_i is proportional to $(1/i)$, we can approximate $w(i)$ by a rectangular window.

Juang et al [4] proposed a new distance measure for speech recognition that improved significantly the recognition rate. The expression of this new distance is the same as Eq. (8a), and the expression for the weighting $w(i)$ is the same as Eq. (2). The equivalent window in the quefrency domain is a bandpass liftering window. The effect of this window on the LPC spectrum is to reduce the unnecessary sensitivity in

spectral comparison due to peaks in the LPC spectrum. This is done by smoothing the spectrum, without distorting the fundamental formant structure.
Finally, a Euclidean distance measure d_E, using liftered cepstrum (1) and delta cepstrum (3) in the feature vectors, was used to improve recognition [1], with

$$d_E = \sum_{i=1}^{Q}\left[\tilde{c}_t(i) - \tilde{c}_r(i)\right]^2 + \sum_{i=1}^{Q}\left[\Delta\tilde{c}_t(i) - \Delta\tilde{c}_r(i)\right]^2 \quad (9)$$

where $\tilde{c}_l(i)$ and $\Delta\tilde{c}_l(i)$ are cepstral and delta cepstral coefficients defined in (1) and (3), respectively. The variances of the delta cepstral coefficients are adjusted to be about the same as the cepstral coefficient variances.
From the above discussion, we can infer two important points:

1) Adjusting the variances of the different features provides a good equalization of the weighting of each feature.
2) Bandpass liftering provides a less sensitive spectral comparison.

Taking the above into consideration, we propose a *Multifeature Weighted Distance Measure* (MWDM), given by

$$d_{MW} = \frac{\sum_{i=1}^{Q}\left[\tilde{c}_t(i) - \tilde{c}_r(i)\right]^2 + \sum_{i=1}^{Q}\left[\Delta\tilde{c}_t(i) - \Delta\tilde{c}_r(i)\right]^2}{\sigma_c^2} + \sum_{l=1}^{L}\frac{\left[F_t^{(l)} - F_r^{(l)}\right]^2}{\sigma_l^2} \quad (10)$$

or equivalently,

$$d_{MW} = \sum_{i=1}^{Q}\left[\tilde{c}_t(i) - \tilde{c}_r(i)\right]^2 + \sum_{i=1}^{Q}\left[\Delta\tilde{c}_t(i) - \Delta\tilde{c}_r(i)\right]^2 + \sigma_c^2\left[\sum_{l=1}^{L}\frac{\left[F_t^{(l)} - F_r^{(l)}\right]^2}{\sigma_l^2}\right] \quad (11)$$

in order to avoid problems when σ_c=0. $F_l^{(l)}$ and $\sigma_{(l)}^2$ denote the new features (energy and/or delta energy, in our case) and their variances, L is the number of new features, and σ_c^2 is the weight for the cepstral terms. Clearly, Eq. (10) affords a simple and unified way of introducing new features in the distance.
The value of the weight for the cepstral terms, σ_c^2, has to be in the range of the cepstral coefficients variances. In Eq. (11), we can use σ_c=0 if the new feature set is highly correlated with the spectral shape. In the next section we determine the *optimum* value (in the sense of minimum error rate) of σ_c.

5. Experimental results

In this section we describe the results from the recognition experiments, with the different features, using MWDM (Eq. (11)). The training and the test database (a subset of the Texas Instruments connected digit database) consisted of 1674 strings from 22 male speakers (69-77 strings per speaker) and 1679 strings from 22 male speakers (65-77 strings per speaker), respectively. The cepstral and delta cepstral coefficient variances, and the variances σ_E^2 and $\sigma_{\Delta E}^2$ were calculated from the training database. In all cases, we used 1 model per word, 5 states per model, and 3 mixtures per state. These experiments are:

1) *Experiment 1*: The feature vector defined in Eq. (6), that includes energy and delta energy, is used with Euclidean distance and energy histograms.
2) *Experiment 2*: Same feature vector as in experiment 1, with MWDM ($F_l^{(1)} = E_l$, $F_l^{(2)} = \Delta E_l$) and energy histograms.
3) *Experiment 3*: Same as experiment 2, but without energy histograms.
4) *Experiment 4*: The feature vector defined in Eq. (7) using MWDM ($F_l^{(1)} = \Delta E_l$) and energy histograms.

Experiments 2, 3, and 4, were conducted for various values of σ_c, chosen to be in the range of cepstral coefficients variances, $0 \le \sigma_c \le 0.5$ plus σ_c=1.0. A graph of the Number of String Errors, NSE, (for 1679 strings) versus the cepstral terms weight, σ_c, for these experiments is shown in figure 1. Table 1 also shows a comparison between the number of string errors.

σ_c	Exp. 2	Exp. 3	Exp. 4
1.00	159	140	167
0.50	149	147	140
0.44	139	142	--
0.39	132	130	142
0.34	--	143	--
0.28	133	147	135
0.18	139	--	139
0.00	139	140	134

Table 1.- Number of strings errors (for 1679 test strings) for experiments 2, 3 and 4.

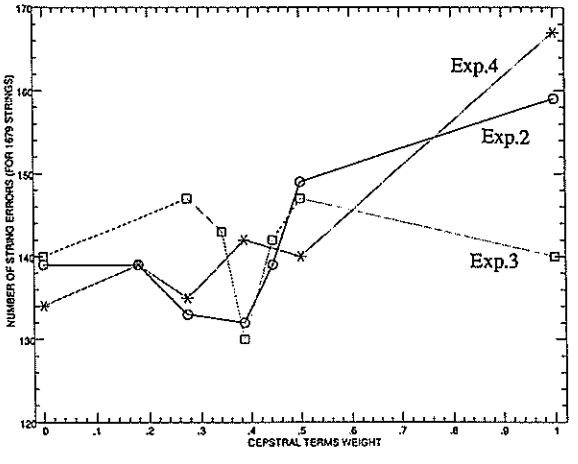


Figure 1.- NSE versus σ_c value for experiments 2, 3 and 4.

As a reference, we used the number of string errors

(unknown length) obtained for the system described in [1], with the above test database, NSE=161 (error rate 9.59%). The conclusions from these experiments are:

1) Experiment 1: The number of errors is NSE=173 (10.30%). This result indicates that energy and delta energy degrade the error rate if Euclidean distance is used for clustering. This is because the numerical values of energy and delta energy are far greater than the values of cepstrum coefficients, and, therefore, energy and delta energy are overweighted. Thus, we need a distance measure with more equal contribution from the different features.

2) Experiment 2: This yields a minimum error rate of NSE=132 (7.86%) for σ_c=0.387. In all the cases, NSE is less than 161 (9.59%) ($0 \le \sigma_c \le 1$). Even for σ_c=0.0, the recognition rate is better than that for the system described in [1].

3) Experiment 3: There is a very sharp minimum for σ_c=0.387 (NSE=130, 7.74%). This is the absolute minimum error rate in our experiments.

4) Experiment 4: In this case, the best result, NSE=134 (7.98%), is obtained for σ_c=0.0. As in the experiment 2, better recognition than the system in [1] is obtained for $\sigma_c < 1.0$.

From these experiments we conclude the following about the optimum value of σ_c.

a) The best results are obtained when the value of σ_c is chosen to be close to the standard deviation of all the cepstral and delta cepstral coefficients (σ_c=0.32). Even for a range of values near the optimum, string error rates are lower than that for the system in [1]. This means that, when features of different classes are used, variances can be used as good weightings for these features.

b) Even when $\sigma_c \to 0$ (the new features are not used for clustering), good results are obtained. This can be explained based on the correlation between the LPC shape and energy [2].

c) The absolute minimum error rate is achieved in exp. 3. This result indicates that we can get very good error rate using MWDM with energy and delta energy and without histograms, but in this case the error rate is very sensitive to the value of σ_c. Thus, it is safer to use energy and delta energy in the feature vectors along with energy histograms, for high performance connected digit recognition.

We carried out a fifth experiment with the complete TI database (8565 training strings and 8578 testing string from male and female speakers) using 1 model per word, 10 states per model, and 5 mixtures per state. This experiment was the same as experiment 2 (using the same features cepstrum, delta cepstrum, energy, and delta energy, and also energy histograms), but with the whole database. In this experiment, σ_c was chosen to be the the overall standard deviation of the cepstral and delta cepstral coefficients (σ_c^2, σ_E^2 and $\sigma_{\Delta E}^2$ were calculated from the whole TI database). The error rate obtained in this experiment was NSE=341 (3.975%). In comparison, the error rate for the system described in [1] with the whole database was NSE=404 (4.709%). This result shows that we can get an improved high performance connected recognition system by using energy and delta energy information along with cepstral and delta cepstral coefficients using the MWDM distance measure for clustering during the training process.

6. Summary

In summary,

1) The Euclidean distance has higher error rate when energy and delta energy are incorporated in the feature vectors, because of their large numerical values in comparison with the values of the cepstral coefficients.

2) The distance measure MWDM yields lower error rates when a suitable value of σ_c is used (close to the overall standard deviation of cepstral and delta cepstral coefficients), improving the system described in [1].

3) Best results are obtained when σ_c is chosen to be close to the overall standard deviation of cepstral and delta cepstral coefficients. Recognition rate is improved even when $\sigma_c \to 0$, because of the correlation between the LPC shape and the energy information. Finally, the error rate is very sensitive to the value of σ_c when energy histograms are not used.

4) MWDM distance measure used for clustering during the training, with cepstrum, delta cepstrum, energy, and delta energy, as the features for recognition and σ_c set at a value close to the standard deviation of the cepstral and delta cepstral coefficients, yields an improved high performance connected digit recognition system.

5) Indeed, the distance defined in (10) could be a simplified case of a more general distance measure when features of several classes are considered. This distance measure would involve the computation of the best weights for each class of features. The problem is that, so far, the only way to determine the weights is by experiments, and this becomes unmanageable when the number of different classes is too large.

Acknowledgement

Special thanks to Pedro J. Moreno (from Telefonica I+D), for their stimulating discussions throughout this work.

REFERENCES

[1] L. R. Rabiner, J. G. Wilpon, and F. K. Soong. "High Performance Connected Digit Recognition, Using Hidden Markov Models," *Proc. ICASSP 1988*, paper S3.6, pp. 119-122, April 1988.

[2] L. R. Rabiner, M. M. Sondhi, and S. E. Levinson, "A Vector Quantizer Incorporating Both LPC Shape and Energy," *Proc. ICASSP 1984*, paper 17.1.1, March 1984.

[3] Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-35, No. 10, pp. 1414-1422, October 1987.

[4] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-35, No. 7, pp. 947-954, July 1987.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A NEW PRE-PROCESSING FILTER FOR A NETWORK BASED SPEECH RECOGNITION

Hiroyuki Sugawara, Shogo Nakamura, Yoshihiko Horio and Masahide Yoneyama

Tokyo Denki University, Japan

In this paper we describe a new time spectrum pattern(TSP) processing method based on a state network(SNET) to be used for recognizing monosyllabic words. To evaluate performance of the modified feature extraction methods using the new pre-processing filter, a computer simulation of speech recognition for ten figures was carried out. As a result, the recognition rate was improved about 20% compared with the preceding one.

1. INTRODUCTION

Recently, new algorithms for speech recognition using neural networks and parallel distributed processing have been proposed and some advanced results have been reported [1][2]. However many unsolved problems remain. One of them, the recognition of utterances having short time spans such as phonemes and monosyllabic words, is caused by the difficulty of feature extraction, which is the separation of the phonetic differences within an utterance. Effective feature extraction from speech signals would be a very important means of enhancing speech recognition rate. In this paper we describe a new time spectrum pattern (TSP) processing method based on a state network(SNET) to be used for recognizing monosyllabic words.

2. OUTLINE OF STATE NETWORK BASED SPEECH RECOGNITION

First we will explain how a state network is built up from TSPs. The speech information used here is local spectrum peak (LSP) patterns obtained from TSP. As shown in Fig. 1, the speech input signal enters a 16ch digital Band-Pass Filter Bank (BPFB), and then 16ch envelopes are obtained as output of a Low-Pass Filter Bank (LPFB). These envelopes are sampled and held at 200[Hz]. The lifting and filtering section of the pre-processing filter performs the functions of smoothing the time readings with low-pass filters and emphasizing TSP frequency readings with a high-pass filter. The resultant TSP is called a modified TSP (MTSP) and an example is shown in Fig. 2(a). In the MTSP, emphasized spectrum peaks are set to 1 and all other points are set to 0 as in Fig. 2(b). We call this type of MTSP an LSP pattern.

In Fig. 2(b), each row is represented in 16-bit binary format and digits having a value of 1 correspond to local spectrum peaks. Assuming each row is a 16-dimensional vector, speech signals can be represented as a series of vectors. By considering a vector to be a state, one vector constitutes a network as shown in the diagram in Fig. 3(a) and is referred to as a state network. When the same speech is uttered repeatedly, similar state networks will be obtained. But they do not completely coincide. By adding new information to the old network, the state network for a specific speech pattern grows as in Fig. 3(b).

LSP patterns frequently reproduce the same peak pattern at parts of the speech signal which are steadily repeated. In such cases, the repeated part is replaced with a single state so that the state network can be compressed and memory can be conserved.

In the recognition phase, an unknown MTSP is input into all state networks, and a search is made for state networks in close agreement with the MTSP. The procedure is as follows.

If unknown speech X is the first letter of the alphabet, 'A', the peaks of the MTSP of X should coincide in many places with positions having the value of 1 in the state network of 'A', SNETA. Since the MTSP of X, MTSPX, is also considered to be a 16-dimensional vector series like SNETA, an inner-product of MTSPA and SNETA (innerXA) can be calculated as follows:

$$innerXA(i) = (MTSPX_i, NETA_i) = \sum_{j=1}^{16} X_{i,j} \cdot A_{i,j} \quad (1)$$

where,

$$MTSPX = (X_{i,1}, X_{i,2}, \ldots, X_{i,16})$$

$$NETA = (A_{i,1}, A_{i,2}, \ldots, A_{i,16})$$

and i is a time index.

Inner$XA(i)$ and inner$XA(i+1)$ are evaluated; if inner$XA(i+1)$ is greater than inner$XA(i)$, then the same evaluation is performed between the next time indexes, $i+1$ and $i+2$, and so on. If this procedure is carried out to the final state of SNETA, it is considered that X is most likely the letter 'A'[3][4][5][6].

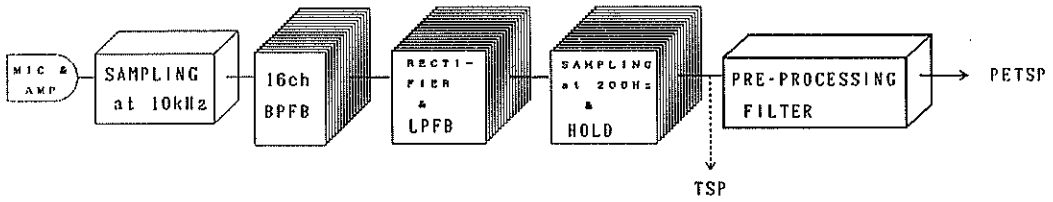


Fig. 1; A pre-processing diagram.

	0	78													
8	1	-10	-16	-9	-12	4	8	10	7	-6	0	0	3	6	7
10	2	-11	-17	-11	-15	5	9	11	9	-7	-1	1	3	8	9
11	3	-12	-19	-14	-19	5	10	13	10	-8	-1	1	4	9	10
14	4	-13	-20	-17	-23	6	12	15	12	-9	-2	1	5	10	12
16	5	-14	-22	-20	-28	6	14	18	14	-10	-3	1	6	12	14
18	7	-16	-23	-24	-33	6	16	20	16	-11	-3	2	7	14	16
21	8	-17	-25	-28	-39	6	18	23	18	-13	-4	2	8	16	18
22	10	-18	-26	-32	-45	6	20	25	21	-14	-5	3	9	18	20
25	11	-20	-28	-37	-51	6	22	27	23	-15	-5	4	11	20	23
27	13	-21	-28	-41	-57	5	24	30	26	-17	-5	6	12	22	25
29	14	-22	-30	-46	-64	6	27	32	28	-18	-5	8	14	24	28
32	17	-24	-31	-51	-70	6	29	36	31	-19	-5	7	14	26	32
36	19	-26	-33	-57	-77	6	32	39	33	-20	-5	7	14	29	38
39	20	-29	-35	-62	-85	6	34	42	36	-21	-5	6	16	33	42
40	23	-27	-31	-59	-92	6	34	42	37	-24	-6	6	17	35	44
38	28	-26	-9	-57	-93	0	30	38	34	-26	-10	1	16	35	43
31	34	-14	28	-50	-88	-12	20	29	30	-29	-16	-4	13	31	36
20	40	6	73	-40	-82	-28	6	17	22	-25	-22	-11	3	22	24
12	43	23	107	-42	-67	-41	-4	5	12	-9	-24	-13	-8	11	10
6	41	48	120	-47	-63	-47	-12	-5	4	15	-22	-15	-15	0	-4
4	39	61	115	-54	-70	-47	-15	-9	-5	45	-15	-14	-18	-6	-12
4	31	62	92	-59	-73	-38	-11	-12	-12	64	-5	-5	-14	-12	-17
1	24	61	82	-68	-82	-32	-9	-16	-16	81	5	1	-5	-15	-22
-2	19	51	78	-61	-84	-32	-10	-17	-22	81	14	14	3	-19	-24
-5	9	40	65	-53	-62	-30	-8	-25	-22	67	15	31	11	-23	-29
-11	3	34	62	-44	-34	-28	-11	-29	-22	53	9	40	15	-27	-32
-16	-4	27	56	-11	4	-33	-17	-33	-29	24	4	44	16	-28	-31
-25	-19	14	33	26	73	-29	-18	-43	-32	-3	2	37	12	-26	-34
[illegible]	-32	4	19	65	128	-23	-25	-50	-46	-23	19	22	2	[illegible]	[illegible]
			[illegible]	106	186	-22	-36	-59	-65	-31	[illegible]				

		0	78												
1	1	0	0	0	0	1	1	1	1	0	0	0	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1
1	1	0	1	0	0	0	1	1	1	0	0	0	1	1	1
1	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1
1	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1
1	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0
0	1	1	1	0	0	0	0	0	0	1	1	1	1	0	0
0	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0
0	1	1	1	0	0	0	0	0	0	1	1	1	1	0	0
0	0	1	1	0	1	0	0	0	0	1	1	1	1	0	0
0	0	[illegible]													

Fig. 2; Data of (a)modified time spectrum pattern(MTSP), and (b)local spectrum peak(LSP) pattern.

3. A NEW PRE-PROCESSING FILTER

The upper portion of Fig. 4 shows the original pre-processing filter, with LPF(a) and HPF(a). Figure 5(a) is a block diagram of LPF(a), which operates as a smoother and has a delay time(D) of 4. Figure 5(b) is a block diagram of HPF(a), where D is 16. HPF(a) operates as a lifter and so emphasizes LSP frequency readings. Figure 6(b) shows a waveform obtained by this filter.

Compared with the TSP waveform in Fig. 6(a), the original filter emphasizes LSP. Although these filters are simple, with good qualities, they are not sensitive enough for feature extraction of small fluctuations in TSP, such as for monosyllabic words in a state network, so we have proposed a new pre-processing filter (Fig. 4). Lifters for vowels and consonants were devised. Also, to avoid over-smoothing, the smoother was modified by insertion of an additional Cascaded Integrator Comb(CIC) filter.

First, let us consider the left side of Fig. 4. Whether TSP is a consonant or a vowel is automatically determined by the condition:

$$\sum_{j=1}^{5} X_{i,j} - \sum_{j=12}^{16} X_{i,j} < TH \qquad (2)$$

where i is a time index, j is a TSP channel number, and TH is threshold level, about 60. TH was determined experimentally from a number of utterances. When input TSP enters the pre-processing filter, for example, sum for the lower and upper 5 channels are calculated and compared following Eq.(2). If input TSP satisfies Eq.(2), it is considered a consonant and SW is moved downward so that input TSP enters HPF(b), where D is 8. (cf. Fig. 5(b)) However, if input TSP does not satisfy Eq(2), the input will be regarded as a vowel and SW is moved upward to HPF(a) where D is 16.

Next we consider the right side of Fig. 4. The output of HPF(a) or HPF(b) then enters in parallel with LPF(a) (D=4) and LPF(b) (D=32). Therefore the output of LPF(a) is equivalent to the output of the original pre-processing filter and the output of LPF(b) is smoothed excessively. The combined output, modified peak emphasized TSP (MPETSP), is obtained by calculating the difference between LPF(a) and LPF(b). A MPETSP waveform is shown in Fig. 5(c). From a comparison of the three waveforms in Fig. 5, the new pre-processing filter was found to be the most effective for feature extraction of speech signals.

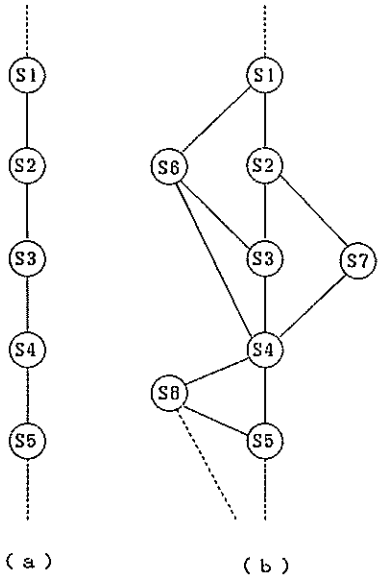


Fig. 3; (a)A state network(SNET) unit, and (b)a state network.

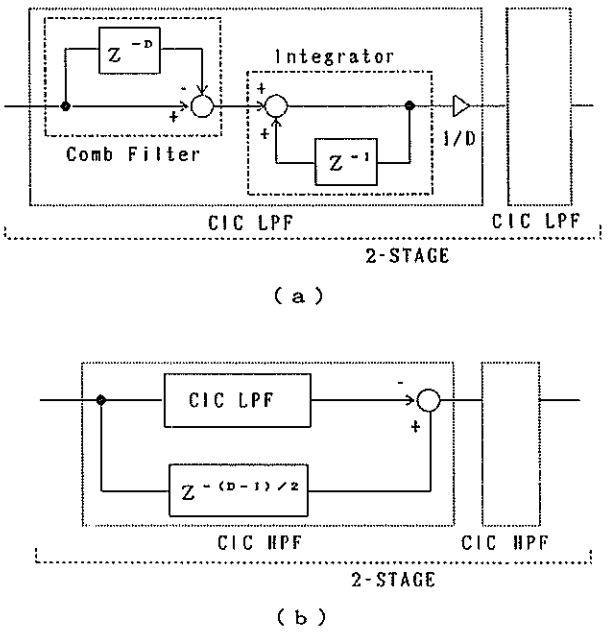


Fig. 5; Block diagrams of LPF(a), and HPF(b).

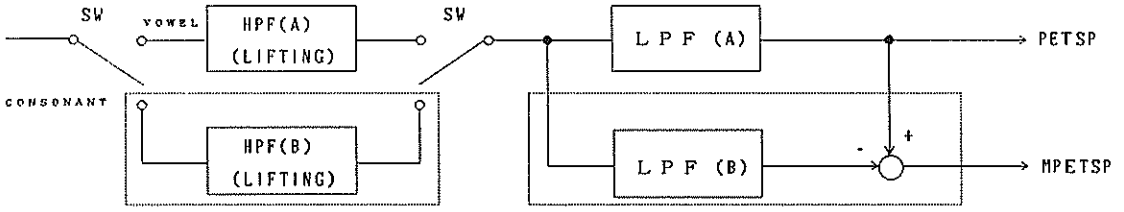


Fig. 4; Block diagram of the pre-processing filter.

4. SIMULATION RESULTS

To evaluate performance of the different feature extraction methods, a computer simulation of speech recognition was carried out. Ten sounds were uttered by a male speaker in Japanese. These utterances were used thirty-two times for organizing a state network and eighteen times for recognition. The resulting recognition rate for all ten figures is shown in Table 1. Using the new feature extraction method the recognition rate was 98.3%, while with the original method it was 80.0%. When the original pre-processing filter was used with only HPF(b) parallel to HPF(a), the recognition rate was 93.3%. These results show that the new pre-processing filter is extremely useful for speech recognition.

5. CONCLUSIONS

We have presented feature extraction using a new pre-processing filter for speech recognition. We demonstrated the efficiency and sensitivity of this filter. Also a state network architecture and a recognition algorithm were briefly described.

Computer simulations for the 26 letters of the alphabet are now being carried out. Our anticipated result is an overall recognition rate of greater than 90%.

Table I Recognition results.

	Original filter	Modified filter	
		HPF(b)	HPF(b)+LPF(b)
number of errors per 180 utterances.	36	12	3
recognition rate.	80.0%	93.3%	98.3%

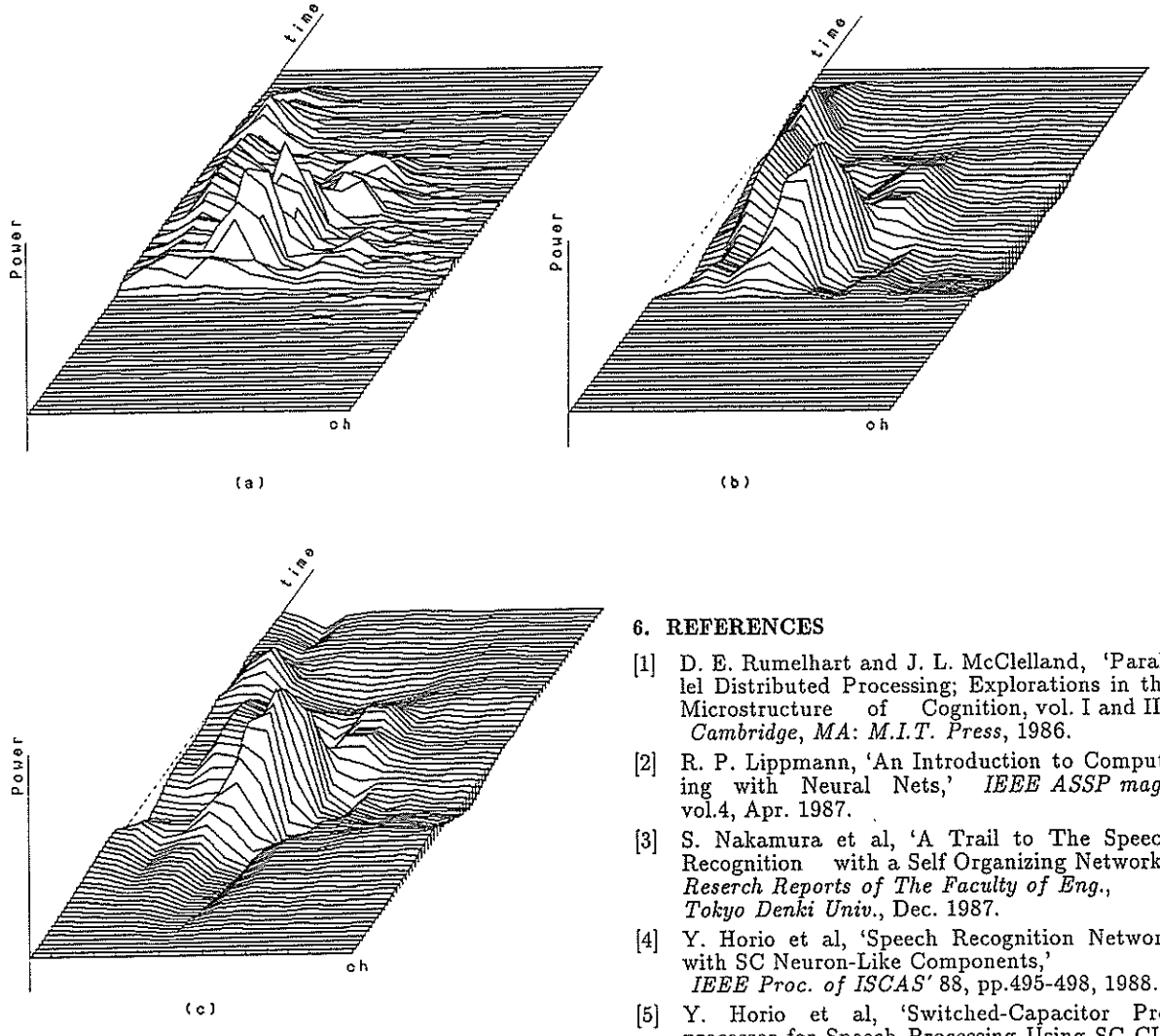


Fig. 6; Waveforms of TSP(a), PETSP(b), and MPETSP(c).

6. REFERENCES

[1] D. E. Rumelhart and J. L. McClelland, 'Parallel Distributed Processing; Explorations in the Microstructure of Cognition, vol. I and II,' *Cambridge, MA: M.I.T. Press*, 1986.

[2] R. P. Lippmann, 'An Introduction to Computing with Neural Nets,' *IEEE ASSP mag.*, vol.4, Apr. 1987.

[3] S. Nakamura et al, 'A Trail to The Speech Recognition with a Self Organizing Network,' *Reserch Reports of The Faculty of Eng., Tokyo Denki Univ.*, Dec. 1987.

[4] Y. Horio et al, 'Speech Recognition Network with SC Neuron-Like Components,' *IEEE Proc. of ISCAS'* 88, pp.495-498, 1988.

[5] Y. Horio et al, 'Switched-Capacitor Preprocessor for Speech Processing Using SC CIC Filter,' *IEEE Proc. of ISCAS'* 89, pp.1311-1314, 1989.

[6] K. Yoshizawa and H. Ishida et al, 'The Speech Recognition Using a Self Organizing Network,' Inst. Elec. Commun. Eng. Japan, Spring Conf., Mar. 1989 (in Japanese).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Eisevier Science Publishers B.V., 1990

PRINCIPAL AND DISCRIMINANT COMPONENT ANALYSIS FOR FEATURE SELECTION IN ISOLATED WORD RECOGNITION

E. Lleida, C. Nadeu

Dpto. of Signal Theory and Communications, ETSIT-UPC.
P.O. Box 30.002, 08080 Barcelona, Spain

ABSTRACT

In this paper we propose the use of the Principal Component Analysis and Discriminant Analysis as feature selection step in a word recognition system. The classic pattern-matching approach used in isolated word recognition assumes that the adjacent feature vectors are uncorrelated and that the variability of speech can be accounted for the same distance measure for all words. However, these assumptions are not true, and a feature selection process is needed to deal with these problems. This work proposes the use of two steps as feature selection process. One is related with the best representation for temporal selection and the other one is related with the discriminant properties for frequency selection.

I. INTRODUCTION

The first step in any speech recognition system is the signal feature measurement. Typically, the speech signal is modeled by a sequence of feature vectors called 'Template' in the IWR environment. Generally, feature measurement methods are block processing models giving N vectors of P features. In this work, the LPC technique has been chosen as a feature measurement method.

The speech signal has stationary parts which are represented by several feature vectors, having a great redundancy [1,2]. Therefore, we can look for a new model in which the correlation among feature vectors is removed. For this purpose we assume that there is an underlying set of "real" uncorrelated features, and the features we are working on are "impure" in the sense that is a linear combination of those "real" features. Then, the objective is to find a transformation which recovers the "real" features [3]. Basically, the problem is to represent the sequence of spectra by a superposition of the members of any orthogonal family of functions where the input template is represented with less coefficients. If y(n) is the nth LPC vector, the transformation obeys the following formulation

$$y(n) = \sum_{m=1}^{M} \alpha_m \phi_m(n) \qquad (1)$$

where ϕ_m is the mth transformation function and α_m is the new mth feature vector.

This work was supported by the PRONTIC grant 105/88.

A typical family of functions which performs this transformation is obtained by means for the Principal Component Analysis (PCA). It minimizes the mean square error between a vector and its estimation by a linear combination. In this work, the Principal Component Analysis is used to remove the temporal correlation in order to obtain M uncorrelated vectors, being M<<N. Thus, a new template is obtained with M uncorrelated vectors which are arranged in variance, not being required time-alignment to compare two templates. The temporal information is retained in the transformation functions (ϕ_m). These functions are found from a training set.

In order to increase the separability among words, a new transformation is proposed in the frequency dimension. In this case, taking the M vectors obtained in the temporal selection, a transformation matrix associated with each vector of a word is sought. Two approachs are proposed; one maximizes the projected variance with the constraint that these functions must be orthogonal to a set of given functions which represent to the others words or a set of words of the vocabulary.The purpose of this step is to obtain a new set of functions which projects a word over a new axis which are orthogonal to the axis of maximum variance of the other words. The second approach maximizes the distance among this vector and the corresponding vectors of the others words. Thus, a new vector δ_m, called discriminant feature vector, will be obtained by transforming the vector α_m as follows

$$\delta_m(q) = \sum_{i=1}^{P} \alpha_m(i)\, \varphi_{m,q}(i) \qquad 1 \le q \le Q \qquad (2)$$

where $\varphi_{m,q}$ is the qth discriminant function of the mth vector. After these processes, a template of MxQ dimension is obtained where M<<P and Q<<P, and the classification is done in this new space by comparing the discriminant vectors by means of the Euclidean distance and without time alignment.

In section 2 a description of the feature selection process is presented. Section 3 explains the training process and the test data base. The recognition experiments are reported in section 4.

II. FEATURE SELECTION

The feature selection process is performed in two steps called Temporal Selection and Frequency Selection.

TEMPORAL SELECTION

Temporal selection is the first step in our feature selection process. Its purpose is to obtain a time compression by removing the correlation of the temporal evolution of the spectrum. Given a NxP matrix Y of spectral parameters $\{y_i(n)\}$ representing N frames of P features, a finite family of orthogonal functions can be found in accordance with (1) by means of the Principal Component Analysis (PCA), reducing a large set of correlated features into a smaller number of uncorrelated features.

If the covariance matrix of a template Y corresponding to a word 'w' is defined as

$$C_{yy}^{w} = \frac{1}{P-1}\sum_{i=1}^{P} (\underline{y}_i - \bar{\underline{y}})(\underline{y}_i - \bar{\underline{y}})^t \qquad (3)$$

where

$$\bar{\underline{y}} = \frac{1}{P}\sum_{i=1}^{P} \underline{y}_i \qquad (4)$$

$$\underline{y}_i = \{y_i(1), y_i(2), \ldots\ldots, y_i(N)\}$$

then the orthogonal functions are obtained in the training step from the eigensystem

$$C_{yy}\,\phi_m = \lambda_m\,\phi_m \qquad (5)$$

where C_{yy} can be equal to an average of the C_{yy}^{w} of each word or an average of all the covariance matrices of all vocabulary words. From this eigensystem, N eigenvalues and their corresponding eigenvectors are obtained. However, only the M eigenvector with the largest eigenvalues are retained. Thus, the transformation matrix is composed by the M eigenvectors with the M largest eigenvalues, ranking them from the largest to the smallest one. Then, the new coefficients α_m have information about the interdependency among the feature vectors. It must be noticed that each orthogonal function is computed using the P features of each frame, thus, these functions carry information of the correlation of the P features. The first eigenvector represents the temporal trajectory of the spectrum with the largest variance, the second one represents the best temporal trajectory which can be obtained if the first eigenvector information is removed from the covariance matrix. As the eigenvalue decreases, the eigenvector asociated carries information of the small variation of the temporal trajectory of the spectrum. The new feature vectors are obtained projecting the initial template with the transformation matrix resulting a sequence of uncorrelated feature vectors. This sequence is the best representation of the initial template in the mean square error sense with the least number of frames. The new feature vectors are ranked from the largest to the smallest variance so the new template needs no time-alignment to be compared with another template.

FREQUENCY SELECTION

The second step of the feature selection process is to compute a transformation matrix for each new feature vector obtained in the temporal selection in order to select the frequency parameters. In the previous step a representation criterion was used in order to obtain a subset of M uncorrelated vector which retain as much information as possible of the initial template. However, this transformation does not take into account the discriminant properties of the feature vectors. Thus, after the temporal selection, a frequency selection step is proposed to obtain a set of discriminant features.

Two methods are studied. The first one is the classical two classes linear discrimination analysis. In this step, a set of Q discriminant functions $\varphi_{m,q}$ is asociated to each vector α_m of a word which increases the separability of this vector from the mth vectors of the other words. The new feature vector is obtained by means of eq. (2) In order to find the discriminant functions, two classes of vectors are defined. For a word 'w', the mth feature vector of any utterance of it forms the correct class and the mth feature vector of the other words forms the incorrect class. Thus, the problem is to maximize the mean of the between-class distance minimizing at the same time the mean of the within-class distance.

Defining the within-class mean distance matrix as

$$W = E\{(\alpha_c - \alpha_c^{p})\,(\alpha_c^{p} - \alpha_c^{p})^t\} \qquad (6)$$

and the between-class mean distance matrix as

$$B = E\{(\alpha_i - \alpha_c^{p})\,(\alpha_i - \alpha_c^{p})^t\} \qquad (7)$$

where α_c is a realization of the correct class, α_c^{p} is the reference prototype of the correct class and α_i is a realization of the incorrect class, the criterion function to be maximized is defined as [3,5]

$$J = tr(F_c B F_c^t) - \lambda(tr(F_c W F_c^t) - 1) \qquad (8)$$

where F_c is the discriminant matrix of the correct class vector, $F_c^t = [\varphi_{m,1}, \varphi_{m,2}, \ldots, \varphi_{m,Q}]$.

The solution of this optimization problem is the eigensystem $(W^{-1}B)\varphi_{m,k} = \lambda_k \varphi_{m,k}$. Therefore, the discriminant matrix is formed by the Q eigenvectors with the Q largest eigenvalue of $W^{-1}B$, whenever their eigenvalues were greater than 1. If an eigenvalue is smaller than 1 the within-class mean distance is greater

than the between-class mean distance. Thus, only those eigenvectors whose eigenvalues are greater than 1 can be used as discriminant functions. As in [4], this process can be seen as a method for finding an specific-frame distance, rotating the frequency dimension in order to better characterize each uncorrelated feature vector of each word.

The second method tries to maximize the projected variance with the constraint that this projection must be orthogonal to the others words or a set of them. So, it is a problem of PCA

$$\text{maximize } \varphi_{m,q}{}^{t} M_{\alpha m} \varphi_{m,q} \text{ with } \varphi_{m,q}{}^{t} \varphi_{m,q}=1 \quad (9)$$

where $M_{\alpha m}$ is the covarianze matrix of the m-th vector with the constraint that these functions $\varphi_{m,q}$ must be orthogonal to a set of given functions $U= \{u_1, u_2, ..., u_N\}$, that is

$$\varphi_{m,q}\, u_j = 0 \quad 1 \le j \le N \quad (10)$$

with this constraint, the function $\varphi_{m,q}$ is obtained by means of the Lagrange multipliers maximazing the function

$$J=\varphi_{m,q}{}^{t} M_{\alpha m} \varphi_{m,q} - \lambda(\varphi_{m,q}{}^{t} \varphi_{m,q}-1) - \sum_{n=1}^{N} \mu_n \varphi_{m,q}{}^{t} u_n \quad (11)$$

The solution are the eigenvectors of the greatest eigenvalues of the matrix

$$(I - U^{t} (U^{t}U)^{-1} U)\, M_{\alpha m} \quad (12)$$

Thus, the function $\varphi_{m,q}$ is the best function in the least square error sense that is orthogonal to a set of given functions. Therefore, the discriminant matrix is formed by the Q eigenvectors with the Q largest eigenvaules.

III. TRAINING PROCESS

Test data base

A data base consists of ten repetitions of the Catalan digits {u, dos, tres, kuatra, sink, sis, set, vuit, nou, zeru}uttered by six male and three female speakers (900 words) and recorded in a quiet room.

Feature measurement

The speech signal was sampled at 8 KHz, pre-emphasized ($H(z)=1-0.95z^{-1}$) and 8 Log-Area ratios were computed each 15 ms for the digit data base using the LPC analysis of 30 ms of the speech signal. A typical Hamming smoothing window was applied to the data. The beginning and end of every utterance were automatically detected by mean of an algorithm based on the signal energy. After the LPC analysis, templates were normalized to a fixed number N of frames, being N equal to 30 for all the words. The Log-Area ratios were chosen as feature because of their stability properties since any kind of transformation gives an stable system.

Feature selection training

In this work, we use a transformation matrix T for all the words of the vocabulary in the temporal selection step. In this case, the covariance matrix C_{yy} is obtained averaging the covariance matrix of each training word. The results of this process are quite similar to the Discrete Cosine transform[6].

The output of the temporal selection are templates of M feature vectors being M equal for all templates. The frequency selection step by linear discriminant analysis computes a discriminant matrix for each feature vector. For this purpose, a mean vector of the mth feature vector is computed and used later as reference. This mean vector is the reference prototype for the mth vector and it is used to compute the within-class and between-class mean distance matrix. In order to take the best discriminant functions, the number Q can be adapted to each word or can be fixed and equal to each word. The discrimination information are in the eigenvalues of $W^{-1}B$. A big eigenvalue indicates a good discrimination property for this feature. Table 1 shows the eigenvalues when M equal to 3 for the word /dos/. It can be seen how the first three eigenvectors have good discrimination properties. Therefore the word /dos/ can be represented by three vectors of three features.

The frequency selection step by PCA requires to define the orthogonalization vectors u_j. For this purpose, we define two methods. The first method solves a problem of two classes and the vectors u_j are the mean vector of the incorrect class. The second method solves a problem of multiple classes and the vectors u_j are the mean vectors of a set of confusable words of the incorrect class.

Q \ M	1	2	3
1	18.63	115.11	18.62
2	9.7	8.01	11.41
3	7.23	7.1	3.45
4	4.81	3.9	2.86
5	2.38	2.5	1.37
6	1.97	1.63	1.02
7	0.7	0.8	0.87
8	0.6	0.5	0.55

Table 1. Eigenvalues of the matrix $W^{-1}B$ for the first three uncorrelated vectors of the word /dos/.

IV. RECOGNITION EXPERIMENTS

A classical pattern recognition system which compares an input template with a set of reference templates by means of the Euclidean distance between frames was used. The system makes use of a linear frame to frame comparison. The references, obtained in the training process, are constituted by the new feature vector obtained in the feature select process and two transformation matrices. One of them is used to select the temporal feature vectors and it is the same for all the

words. The other one is used to select the frequency features and it is specific for each frame obtained in the temporal selection step.

The experiments were made with a speaker independent approach. In this case, the training set was made up by ten repetitions of six speakers and three speakers were used as test. In each recognition experiment, an evidence measure was computed as $Ev=(D2-D1)100/D1$; $0 \leq Ev \leq 100$; where D2 is the distance to the second candidate and D1 is the distance to the first candidate.

Four experiments were performed using: a) classical speaker independent system as in [6], b) only temporal selection, c) Discriminant Analysis and d) Principal Component Analysis with constraints. Table 2 shows a resume of the results of the four experiments. In the classical system, the best results where obtained using two candidates per word. The second experiment was made using only the temporal selection step with M=3. A mean vector was used as reference for each word. In the discriminant analysis [3], each word has only one candidate in the reference set,i.e. the mean vector of the training set. The number of temporal features M were equal to 3 and the frequency features Q were selected for each word in orden to minimize the error rate (the mean value of Q was 3). With these conditions, the error rate is 1,66 % with a mean evidence of 77 %. It can be noted the high evidence mean obtained in this approach. Finally, the fourth experiment was made using the PCA with orthogonality constraints. The first step in this experiment was to find the confusion word set for each digit by means of some experiments with and without frequency selection. This confusion word set was used to form the set of orthogonalization vectors u_i. Thus, the transformation matrix for each vector α_i was obtained using vectors of its confusion words. The best results was obtained using all the frecuency features, that is, Q=8. With this conditions, this method gives the best performance, decreasing the recongnition error to 0.66 % with a mean evidence of 50 %.

		% error	evidence
a)	Clustering System	2.00	45 %
b)	Temporal Selection	3.00	54 %
c)	Discrimin. Analysis	1.66	77 %
d)	PCA with constraint	0.66	50 %

Table2. Results for the speaker independent experiments.

With regard to the computational load, the number of multiplications needed for recognizing a word is very low. Using templates of NxP dimension with a transformation matrix T with M vectors, Q frequency features and V vocabulary words, the number of multiplications is (NxPxM) for the temporal selection step, VxMxPxQ for the frequency selection step and VxMxQ for the comparison step. Thus, in our experiments with the digit data base where N=30,P=8,M=3 and V=10 the number of multiplications is 960 for only the temporal selection, 1680 for the Discriminant Analysis with Q=3, 2880 for the PCA method with Q=8 and in the classical system with dynamic time warping and two templates per word reference is $(N^2/3)xVxPx2 = 48000$.

V. CONCLUSION

A two step feature selection process is introduced for isolated word recognition. The first step takes into account the correlation among the N frames of a template giving a new subset of uncorrelated frames. The second step takes into account the discrimination properties of the P features of each uncorrelated frame by means of two methods: Discriminant Analysis and PCA with constraints. Both methods increase the recognition performance of an independent speaker recognition system. The PCA with constraints gives the best results but its training process is more elaborated that the training process of the Discriminat Analysis. Further experiments must be made to test the feature selection process in a more difficult vocabulary.

REFERENCES

[1]E. Lleida, C. Nadeu, J.B. Mariño, "Speech parametrization and recognition using block and recursive linear prediction with data compression", European Conference on Speech Technology, pp. 300-303, Edinburgh- 1987.

[2]R. Pieraccini, R. Billi, "Experimental comparison among data compression techniques in IWR", ICASSP-83, Boston.

[3]E. Lleida, C. Nadeu, J.B. Mariño, "Statistical Features Selection for Isolated Word Recognition", ICASSP-90, Albuquerque, 1990.

[4]E.L Bocchieri, G.R. Doddington, "Frame-specific statistical features for speaker independent speech recognition". IEEE trans. on ASSP, Vol 34, Ag. 1986.

[5]K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, 1972.

[6]L.R. Rabiner et al. "Speaker-Independent recognition of isolated words using clustering techniques", Trans. on ASSP-27, Ag. 1979.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

SIGNAL SEGMENTATION INTO SPECTRAL HOMOGENEOUS UNITS

Jose C. Segura-Luna, Juan M. Lopez-Soler,
Antonio Peinado-Herreros, Victoria Sanchez-Calle,
Antonio J. Rubio-Ayuso

Dept. de Electronica y Tecnologia de Computadores
Univ. de Granada. 18071 Granada, Spain.

This paper discusses a preliminary study for using spectral distance measures instead of statistical tests of gaussian processes discrimination in detecting spectral changes in signals.
This spectral distortion measure allows to design algorithms for speech signal segmentation, which is usually necessary in recognition systems using subwords units.
The proposed spectral distance behaves as well as others previous reported measures but it is less noisy.

1. INTRODUCTION

Nowadays, it is possible to classify the recognition systems into two big groups. The first one includes those systems that use words as basic units for analysis and recognition, unlike the second one that uses subword units (phonemes, diphonemes, subphonemes, etc.) for analysis and decoding of speech, avoiding constriction imposed by large size codebooks (like with the first group). The second group systems need a preprocessor to perform the speech signal segmentation and then to obtain the analysis units. It is also possible to do the segmentation and the featuring at the same time. Therefore, the development of speech signal segmentation algorithms that provide homogeneous analysis units for acoustic-phonetic decoding is an interesting task.

The problem of detection of changes in spectral features of signals has been previously researched [1] for random signals modified by linear systems with unknown parameters, using statistical methods of gaussian processes discrimination. Recently [2], these methods have been successfully applied to speech signal segmentation (subphonemic unit extraction).

This paper discusses a preliminary study for using spectral distance measures instead of statistical tests of gaussian processes discrimination for speech signal segmentation.

1.1. Setting the problem

The problem is as follows: we have a signal with time-dependent spectral features, and we wish to know when a temporal variation occurs. The resolution of the following points is required in order to solve the changes detection problem:

1) Spectral features identification of the speech signal.

2) Definition of a parameter for spectral feature variation quantification.

3) Choice of an algorithm for determining changes and estimating the time in which the change occurs (all of this is performed from the values of the above parameter).

2. THE SIGNAL MODEL

A suitable model of the speech signals spectral features, and widely assumed, as the LP model [3], is used in this work.

In order to detect changes of the signal spectral features, two models of signal (evaluated in different times) are necessary. Like in [1] and [2], the first model is a global model of the signal segment that is candidate to be segmented as a homogeneous unit. This model is called M_0, and is evaluated over a window W_0 that is increased with the time. The second model, that it is called M_1, is evaluated over a short segment of signal, represented by a window W_1, that is moved along the time axe. In Fig. 1, we can see the position of the windows for model evaluation. We use a dotted line for the window position at the current time, and the continuous line is the position some time later. These models are characterized by the autocorrelation coefficients of the signal, from which it is possible to obtain the LPC parameters, used for the distance measure computation. This distance will be used as the parameter to discriminate models.

The model parameters are the following:

$$M_0 \ : \ \{ r_i^0 \}_{i=1}^{p} \qquad (2.1.a)$$

$$M_1 \ : \ \{ r_i^1 \}_{i=1}^{p} \qquad (2.1.b)$$

and their spectral models:

$$M_0 \ : \ H_0(z) = \frac{\sigma^0}{1 + \sum_{k=1}^{p} a_k^0 z^{-k}} \qquad (2.2.a)$$

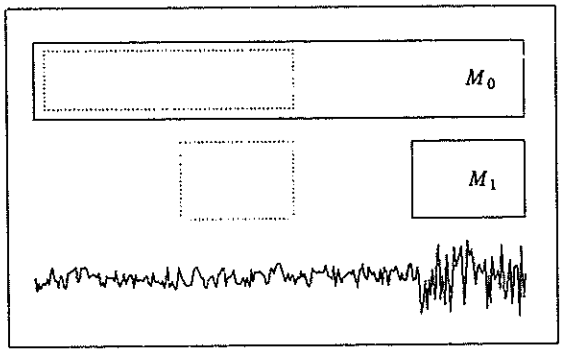


Figure 1

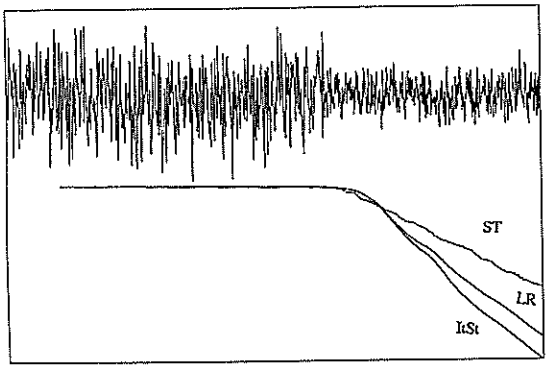


Figure 2

$$M_1 \;:\; H_1(z) = \frac{\sigma^1}{1 + \sum_{k=1}^{p} a_k^1 z^{-k}} \qquad (2.2.b)$$

where $\{ a_k^i \}_{k=1}^p$ are the LPC coefficients of model i, and σ^i is the filter gain of model i.

2.1. Change detection

The change detection is performed monitorizing a spectral distance measure $d(M_0, M_1)$ between the two models. The distance measures tried throughout this work are *Itakura-Saito* (ItSt) and *Maximum Likelihood Ratio* (LR). The computation of distances between M_0 and M_1 can be found in [4].

The simple computation of the distance measure is unsuitable for change detection purposes due to its noisy behavior, so we define a accumulated distance measure D:

$$D_n = - \sum_{i=1}^{n} d_i(M_0, M_1) \qquad (2.3)$$

Its performance is shown in Fig. 2 for a random signal with the following LPC parameters:

after the change:

$$a_1 = 1.67 \quad a_2 = -1.01 \quad a_3 = 0.06$$

before the change:

$$a_1 = 0.85 \quad a_2 = -0.25 \quad a_3 = 0.20$$

filter gain σ was constant.

In Fig. 2 it is also shown the behavior of the statistical discrimination test used in [1]. In this figure, it is clear that the spectral accumulated distances have a very similar behavior to that in the statistical test, and it is less *noisy* than this. (ST stands for statistical cummulative sum test).

For the change detection of the signal features, it is enough to determine a level in function D. However, the detection of the exact time in which the change occurs is not a easy with this function. For this purpose, we modify the spectral accumulated distance definition, adding a light *offset* that increases the distance in the area where there is no change, and lets the change of slope when this occurs.

$$\tilde{D}_n = \sum_{i=1}^{n} (\delta - d_i(M_0, M_1)) \qquad (2.4)$$

The change is detected when a significant change of the $\tilde{D}_n$ slope appears, which is obtained fixing a detection level, λ. Fig. 3 illustrates this procedure. The change is detected at point n, and the time in which the absolute maximum of $\tilde{D}_n$ occurs, is taken as estimation of the change time. These method, called Hinkley stop rule, is also used by other

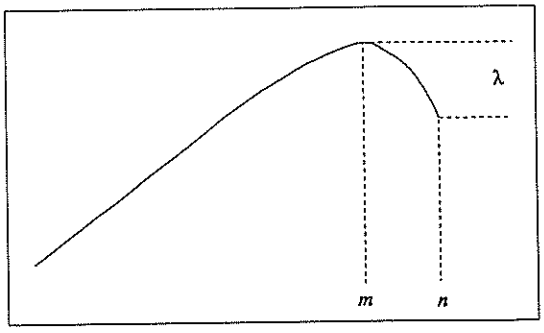


Figure 3

authors [3] in similar contexts.

3. WORD SEGMENTATION

For word segmentation, this method starts from the begining of the word up to detect a change. At this point, the procedure places the segment limit, and restarts from this point, repeating the process up to the end of the word.

Features of test signals

The signals used for the real cases trials were isolated words, sampled at 8 KHz, prefiltered at 4 KHz and preemphatized with a factor μ =0.95.

Parameter tunning

After several trials, $\delta = 0.1$ and $\lambda = 10$ were used as algorithm parameters, for optimal performance.

3.1. Experimental results

At Fig. 4.a, the limits obtained by this algorithm are depicted, for the word MUÑECA, with the *Itakura-Saito* distortion measure. It can be observed that the limits are not as we expected, and even some of them were not detected. The results for the *Maximum Likelihood Ratio* distance measure are shown at Fig. 4.b. Both figures show the signal and accumulated spectral distance evolution, with the limits detected by the algorithm. The limit detection deficiencys are probably due to the smooth signal parameter variation, so that the limits are not detected, or detected with some delay, and due to initialization sensibility of the method. In order to avoid these problems, the algorithm was modified, so that when a limit is detected, the current segment is processed (up to the previous limit) in inverse sense to detect a possible omitted limit [3]. With this modification, the obtained limits are more accurate, so that the number of omitted limits is smaller.

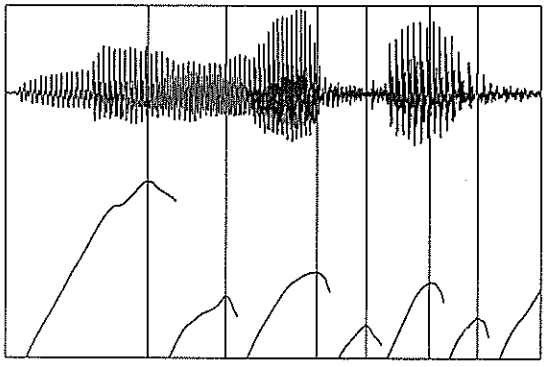


Figure 4.b

Figures 5.a and 5.b show the limits obtained with the *Itakura-Saito* distance measure for words DEDOS and MUÑECA with the previous modification. The upper levels of the figures show the evolution of the LPC spectrum peaks, the middle level is the speech signal, and the lower level is the evolution of functions $\tilde{D}_n$ (with backward evaluation). In these figures, the limit positions are related to areas with similar behavior of the LPC spectrum peaks, corresponding to three types of segments:

a) Stationary segments, like the case of the segment corresponding to the initial murmur in word DEDOS (Fig 5.a), or to sounds /e/ or /s/ in the same word.

b) Transitional segments, like those at the beginning or the end of the vowels, like in MUÑECA (Fig 5.b) at

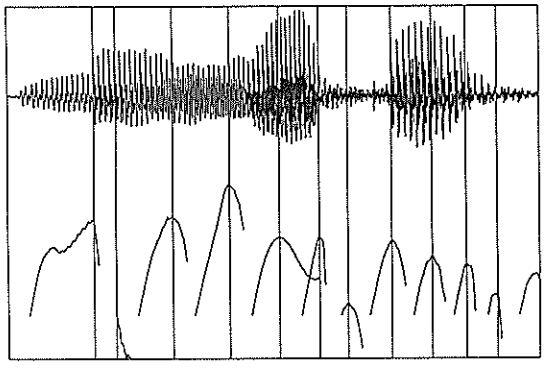


Figure 4.a

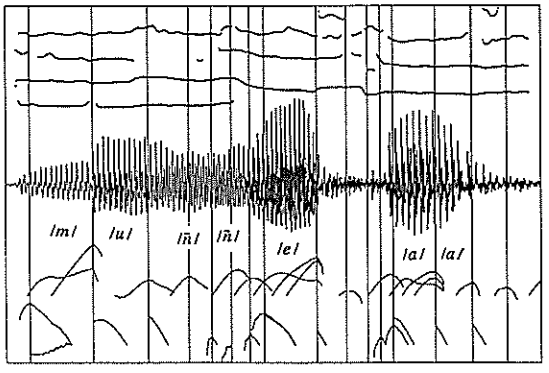


Figure 5.a

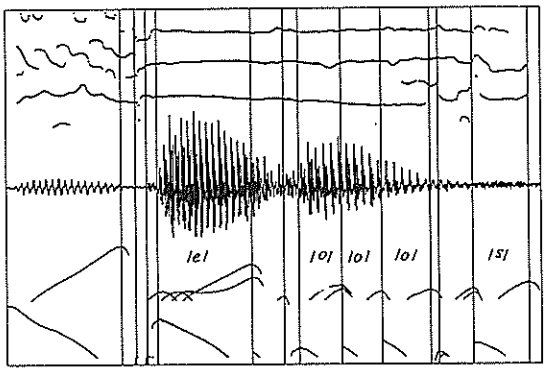


Figure 5.b

the end of /a/ or the beginning of /e/.

c) Short segments, like those in DEDOS (Fig. 5.a) after the sound /o/.

To compare, at Fig. 6 we find the evolution and limits obtained by the algorithm used in [2] for the word MUÑECA.

4. SUMMARY AND DISCUSSION

This work proposes a new parameter to detect signal spectral changes, based on spectral distance measures instead of statistical tests of gaussian process discrimination (that assumes random input for the LPC model).

This method can be applied with different spectral distance measures, with the possibility of incorporating the same distance measure as that one used by the recognition/processing system in which the method is applied.

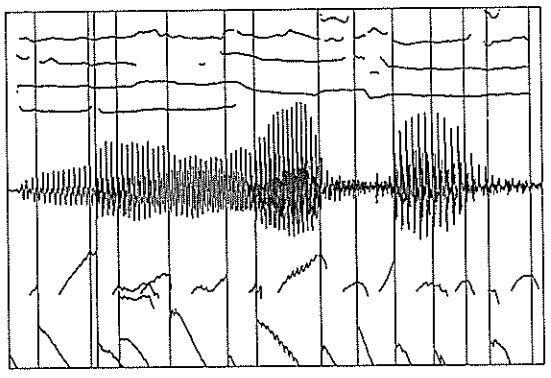


Figure 6

The obtained results are similar with the different distance measures tried in this work, and also similar to those obtained in [2], so we consider that the use of distance measures for segmentation purposes is profitable.

The *Itakura-Saito* distance measure show the best behavior. However, the choice of the distance will depend on the system in which the method will be used.

Extensive studies and forward results evaluation is ongoing in our research lab.

REFERENCES

[1] Michele Baseville and Albert Benveniste. "Sequential detection of abrupt changes in spectral characteristics of digital signals". *IEEE Trans. Information Theory*, vol. IT-29, pp. 709-724, September 1983.

[2] Regine Andre-Obretch. "A new statistical method for the automatic segmentation of continuous speech signals". *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 29-40, January 1988.

[3] J.D. Markel and A.H. Gray, Jr. "Linear Prediction of Speech". *Springer-Verlag Berlin. Hildelberg New York 1976.*

[4] Robert M. Gray, Andres Buzo, Augustine H. Gray and Yasuo Matsuyama. "Distortion measures for speech processing". *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 367-376, August 1980. H. Gray, Jr. and J.D. Markel.

[5] Toshifumi Matsuoka and Tad J. Ulrich. "Information theory measures with application to model identification". *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 511-517, June 1986.

[6] A.H. Gray, Jr. and J.D. Markel. "Distance measures for Speech processing". *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 380-391, Oct 1976.

This work has beeen supported by the CICYT (Comision Interministerial de ciencia y Tecnologia) under project CIT 88-0774.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

AN EMPIRICAL EVALUATION OF FEATURE MAPS AND OTHER CLUSTERING TECHNIQUES FOR FRAME LABELING OF SPEECH.(+)

Gabriela Andreu(*), Enrique Vidal(**), Francisco Casacuberta(**).

(*) Dep. de Ingenieria de Sistemas Computadores y Automática.

(**) Dep. de Sistemas Informáticos y Computación.

Universidad Politécnica de Valencia (Spain).

Abstract.

Obtaining compact and effective frame labeling or "Vector Quantization" of speech is an important issue, since the size of the alphabet or "codebook" strongly impacts on the accuracy and complexity of most procedures which have to be carried out with the resulting speech representation. Furthermore, this alphabet size is limited by the estimation reliability that can be achieved with the available training speech data. In this paper, we compare standard Vector Quantization techniques with other less standard procedures such as Phonotopical Mapping with respect to their ability to obtain such compact and effective frame labeling. We also introduce a two-stage procedure which seems to outperform the capabilities of the above procedures.

(+) Work supported in part by the spanish CICYT under grant TIC-0448/89

1.- INTRODUCTION.

For many situations in speech processing and recognition, it is necessary or convenient to convert vector-string representations of speech objects into strings of labels from a certain discrete set, which is often called "alphabet" or "code-book". Several schemes have been proposed so far to carry out this task, which is usually referred to as "Vector Quantization". Typical alphabet sizes which have been adopted in many practical situations usually run from 64 to 256. However, the complexity and/or accuracy of the procedures to be carried out with the resulting speech representation tend to be strongly dependent on this alphabet size. While larger sizes tend, in principle, to produce greater accuracy, they do require larger computational resources and often fail to lead to robust estimation. This can result, in fact, in important performance degradations.

This applies directly to discrete HMM modeling, where the number of parameters to be estimated grows linearly with the alphabet size. With finite training data, a convenient tradeoff must be arrived at, such that the overall number of parameters to be estimated be small enough to allow a sufficiently robust estimation from the avaieable data. Many other situations also exist in which having the speech signals represented by strings composed of a small number of symbols is even more important.

One of these situations is the Kohonen's interesting algorithm known as "Dynamically Expanding Context" /1/. This algorithm is used as a postprocessor to "correct" pseudophoneme strings so as to produce actual strings of phonemes. For this algorithm to be properly applicable, the pseudophoneme strings must be composed of symbols that roughly correspond to actual phonemic categories. In order to obtain such pseudophonetic strings, Kohonen uses his "phonotopical maps" which are obtained through a partially supervised clustering technique that is based on connectionist principles /4/.

Another situation demanding small alphabet sizes arises with the use of our "Error Correcting Grammatical Inference Algorithm" (ECGI) /2/. This algorithm obtains accurate finite-state structural models of speech objects from samples of these objects, which are assumed to be represented as strings of frame-labels. Although greater recognition accuracy can be obtained with increasingly finer frame labeling (increasingly lager alphabet) this improved accuracy is achieved at the expense of a corresponding increase in the size of the finite-state network produced by the ECGI. Consequently, the trading of accuracy for speed often requires alphabets that, while being sufficiently small, have enough acoustic-phonetic descriptive power for permitting the necessary accuracy. We have found a satisfactory solution through a two-stage C-means procedure followed by a K-Nearest Neighbour (K-NN) classification /2/.

Following the above discussion, this paper compares the relative merits of three speech frame labeling schemes; namely, the standard C-means clustering procedure followed by Minimum Distance classification /3/, usually adopted in discrete HMM modeling; the Kohonen's Phonotopical Mapping /4/, and the two-stage C-means and K-NN procedures which is currently in use as the front-end of our ECGI algorithm /2/.

The above labeling schemes were originally developed for their corresponding specific recognition methods. Obviously, if these schemes are to be compared, a common recognition framework is necessary. The comparative study consists of the evaluation of the three clustering and labeling techniques with respect to their performance in a conventional HMM (estimation and) recognition task involving speaker-independent, isolated words. The results show that reducing the size of the alphabet through the use of Phonotopical Maps, can produce a corresponding (important) reduction in computational requirements, but this produces significantly worse results than those obtained with "standard" code-books and Minimum Distance classification. On the other hand, a similar reduction in computational requirements can be achieved with the two-stage C-means procedure proposed here, with similar, or even better recognition results than those obtained with the standard procedures.

Other empirical comparisons between Vector Quantization techniques have already been carried out /5/. However, these comparisons dealt with the relative computational merits of the different approaches and with their accuracy as measured by the quantization error. The results of this paper complement these previous results in that, here, we deal with the tradeoff between alphabet-size and accuracy as measured by the contribution of each resulting speech representation to the overall recognition rate, using a standard speech recognition method.

2.- CLUSTERING TECHNIQUES.

The conventional clustering techniques to be compared will be briefly described, in this section, along with the newly introduced two-step approach.

2.1.- C-MEANS VECTOR QUANTIZATION.

Let $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ be a set samples consisting of frames of speech signals from a known number C of classes into which they are to be clustered.

The following procedure clusters $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ into C clusters and finds the cluster means $\mu_1,\ldots,\mu_c$ such that the sum of distances from the samples in each cluster to their corresponding mean is a local minimum /3/:

1. Choose some initial values for the means $\mu_1,\ldots,\mu_c$.

Loop: 2. Classify the n samples by assigning them to the class of the closest mean.

3. Recompute the means as the average of the samples in their class.

4. If any mean has changed value, go to loop; otherwise, stop.

Once the means $\mu_1,\ldots,\mu_c$ ("codebook") are obtained, they are assigned convenient labels and the Minimum Distance rule /3/ is used to convert each (new) vector $\mathbf{x}$ (frame) of a speech signal into a corresponding "codeword" or "microphonetic label".

2.2.- TWO-STAGE C-MEANS.

A C-means clustering algorithm is used to obtain a set of M clusters, and their associated means. These mean vectors or prototypes are then clustered again with the same algorithm into a set of N clusters (N << M) and each original prototype is labeled with the identifier of the final cluster it belongs to. The number of final clusters, N (and their identifiers) have to be chosen so as to more or less tune them to the acoustic-phonetic categories that appear in the speech data involved. Once such a set of M labeled prototypes is available, the K-NN classification rule /3/ is applied to convert (new) parametric speech representations into "microphonetic" symbol strings; that is, each input parameter vector or frame is assigned the (microphonetic) label which is in majority among the set of labels associated to the prototypes, which are the K Nearest-Neighbours of this input vector.

2.3.- PHONOTOPICAL MAPS.

Kohonen's Phonotopical Mapping algorithm creates a vector quantizer by adjusting "weights" from P common "input nodes" to M "output nodes" arranged in a two dimensional grid /4/. Input P-dimensional vectors are presented sequentially in time without specifying the desired output. After enough input vectors have been presented, the weights (which are organized as a set of M vectors) will specify cluster vector centers that sample the input space in a way such that the point density function of vector centers tends to approximate the probability density function of the input vectors. In addition, the weights are organized so that topologically close nodes are sensitive to inputs that are physically similar. Output nodes will thus be ordered in a "natural" manner /4/.

The algorithm that was used in this work is described in /4/. The weights of the net are initialized to some small random values. For each new training sample, the distance between the input data and each vector weight is computed. The output node j^* with minimum distance is selected, and this and all nodes in their topological neighborhood are updated as follows:

$$W_{ij}(t+1)=W_{ij}(t) + \alpha(t)(x_i(t)-W_{ij}(t))$$

Where W_{ij} $1\leq i\leq P$, $j \in NE_{j^*}(t)$, is the weight from input node i to output node j at time t, $x_i(t)$ is the input value to node i at time t, $\alpha(t)$ is a gain term, that decreases with time, and $NE_{j^*}(t)$ is the neighbour region of node j^* whose size decreases with time. This process is repeated until no weight changes significatively from one iteration to another or, until the neighbour region is empty.

3.- EXPERIMENTAL FRAMEWORK.

Applications of the above techniques to a particular recognition task are described in this section. The experiments were made with a corpus that is composed of ten repetitions of the Spanish digit vocabulary which was uttered by ten different speakers (5 males and 5 females).

The acquisition and endpoint detection procedures were rather standard, and the signal was sampled at approximately 8.5 Khz with a 12 bit A/D converter. The acquired speech data were then converted into strings of 11-dimensional parameter vectors of Cepstrum Coefficients, which were obtained from a 16 channel Mel frequency scale filter bank at the rate of 66.66 vectors per second.

3.1.- C-MEANS VECTOR QUANTIZATION.

Two code-books of different sizes were obtained by using the C-means Vector Quantization technique that is described in Section 2.1. One of them had 255 code-words, and the second 15 code-words. In both cases, the training data consisted of five repetitions of three speakers (1 female and 2 males), for a total of 4171 11-dimensional vectors. Two sets of labeled speech samples were obtained from the whole corpus by using the corresponding code-books and the Minimum Distance classification rule.

3.2.- TWO-STAGE C-MEANS VECTOR QUANTIZATION.

A code-book of 255 code-words clustered into 15 groups was computed by using the two-stage C-means Vector Quantization technique that is presented in Section 2.2, with M=255, N=15 and the same training data as above. A new set of 15-label labeled speech samples from the whole corpus was obtained by using the new code-book and the K-Nearest-Neighbour classification rule whith K=7.

3.3.- PHONOTOPICAL MAPS.

The training-set was the same as above. For this experiment, the net comprised 96 output nodes arranged in a two dimensional grid (8*12) and 11 input nodes (9216 weights). The training-set was supplied to the net input twenty two times. For the first four times, the neighborhood radius decreased in size with time. Afterwards the training-set was presented eighteen times with a constant neighborhood radius of one. The total number of effectively presented frames was 91762.

The weights were updated by computing the gain term $\alpha(t)$ for every new input depending on the type of neighborhood. Once the above process was completed, the weights were already fixed. Then the nodes were labeled /4/. For that labeling, we used a subset of the training-set whose frames were hand-labeled into a set of 13 "microphonetic" labels, each of which representing one possible, typical sound of the Spanish digits.

3.4.- HIDDEN MARKOV MODELS (HMM).

Hidden Markov Modeling was used to investigate the comparative performance of the methods described above.

An HMM of 8 states without skips was assigned to each word of the dictionary. The Baum-Welch algorithm was used to estimate the probability distribution functions of the models /6/. All the labeled utterances of 6 speakers were used as a training set for a total of 600 word samples. The labeled utterances of the rest of the speakers were used as a test set for a total of 400 word samples.

4.- RESULTS.

Experimental results are presented in this section. Each experiment consisted of model training and recognition by using each set of labeled samples that was obtained by applying each of the methods described above. These experiments are:

- C-means Vector Quantization with a codebook of 255 entries: **VQ255.**
- C-means Vector Quantization with a codebook of 15 entries: **VQ15.**
- Two-stage C-means Vector Quantization with a codebook of 255 entries and 15 labels: **2VQ15.**
- Phonotopical maps with 96 nodes and 13 labels: **PM.**

The results of the different experiments are

shown in Table 1.

	number of codewords	number of symbols	recognition rate %
2VQ15	255	15	98
VQ255	255	255	97.75
VQ15	15	15	97.75
PM	96	13	95.75

Table 1.- HMM results for different labeling techniques. Training set: 6 speakers (600 utterances); Test set: 4 different speakers (400 utterances).

The best result corresponds to the 2VQ15 technique that is proposed here. No differences exist between VQ255 and VQ15; while the former provides more accurate labeling, the models seem to be better trained in the latter. Finally, the worst result is obtained using PM.

5.- DISCUSSION AND CONCLUSIONS.

While increasing the number of VQ symbols would, in principle, permit more precise modeling, the limited amount of training data does not allow for a robust estimation of all the parameters involved, resulting ,in fact, in the same recognition accuracy as that obtained by using a much smaller alphabet which permits a more robust estimation of the fewer parameters. Furthermore, by performing the labeling procedure more carefully (2VQ15), better results have in fact been obtained with the smaller alphabet. The results obtained with the phonotonical-mapping-based labeling technique clearly seem to be the worst.

While these results are not at all conclusive, they seem to indicate that improving current techniques of frame labeling of speech certainly deserves further research.

References:

/1/ T.Kohonen "Dynamically Expanding Context, with Application to the Correction of Symbol Strings in the Recognition Speech". IEEE Proc. ICPR, Paris october 1986.

/2/ H. Rulot, N. Prieto and E. Vidal. "Learning Accurate Finite-state Structural Models of Words Through the ECGI Algorithm". Proc. ICASSP89 pp. 643-646.

/3/ R. O. Duda and P.E. Hart. "Pattern Classification and Scene Analysis". ed. John Wiley and Sons 1973.

/4/ T. Kohonen, K. Makisara and T. Saramaki. "Phonotopic Maps, Insighful Representation of Phonological Features for Speech Recognition". IEEE 7 Int. Conf. on Patt. Rec., Proc. Canada 1984.

/5/ Frank H. Wu and Kaylan Ganesan. "Comparative Study of Algorithms for Design VQ Using Conventional and Neural-Net Based Approaches". ICASSP89 pp. 751-754.

/6/ L.R. Rabiner. "Mathematical Foundations of Hidden Markov Models". In "Recent Advances in Speech Understanding and Dialog Systems". H. Niemann et al. (eds.). Springer-Verlag. 1988.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

REALIZATION OF AN EFFICIENT ALGORITHM IN SPEECH RECOGNITION SYSTEMS

JIN LIU

Heinrich-Hertz-Institut für Nachrichtentechnik GmbH, Einsteinufer 37, 1000 Berlin 10

This paper describes a hardware realization of a low-cost speech recognition system for isolated words. Many factors influence the performance of a recognizer. This paper tries to reveal the connections between these factors and the system performance and to give a system configuration with the minimal system cost for the recognizer, at which the high recognition rate will not be sacrificed.

I. Introduction

To realize a speech recognition system for isolated words, several problems must be solved. These include (1) extraction of speech parameters, that can best describe the speech characteristics with the smallest computation cost; (2) compression of the parameter vectors so that storage demands and the computation time for the pattern comparison process could be minimized; (3) implementation of the algorithm using a favourable hardware configuration.

The choice of the appropriate speech parameters relies on two considerations: They must best represent the speech signal in the sense of recognition rate, and they must be easy to compute. In the presented paper, the speech signals are modelled as an AR process and the cepstrum coefficients are chosen as the speech parameters. Cepstrum coefficients can be obtained easily from linear predictor coefficients, which in turn can be computed with the Durbin-Levinson algorithm. In this paper, the newly developed split Levinson algorithm is applied to calculate the predictor coefficients. This algorithm needs only about half of the multiplications the conventional Durbin-Levinson algorithm needs. For an algorithm implementation on a fixed point signal processor, it must be investigated at first, how the computer representation precision used in split Levinson algorithm influences the algorithm stability. A comparison of the numerical stability of this algorithm with that of the conventional Durbin-Levinson algorithm and the value distributions are made in Section II.

The computation cost of parameter extraction depends on many factors, such as the coefficient orders, the window size, the window form and the overlap grade of signal intervals etc. For example, the multiplication with a window function (except with a rectangular window) as well as the overlap of signal intervals are all computationally expensive. A 50% overlap, for instance, requires a parallel analysis of two successive intervals. This does not only double the computational time but also makes memory accesses more difficult. All these influences of the factors on the recognition rate are discussed in Section III.

The extracted parameters are further compressed in the processing chain with a method called "Trace Segmentation". This idea is based on the fact, that the adjacent parameter vectors extracted from the speech signal with the same analysis interval length change usually very few. Obviously these parameter vectors can be irregularly resampled or interpolated if necessary. In this way a great number of vectors, which carry practically no new information about the speech signal, can be compressed. The influences of different segment lengths on recognition rate are also given in Section III.

Another interesting result described in Section III shows that the spectral aliasing distortion of speech signals has little influence on the recognition rate. This means that one can implement the recognition system without anti-aliasing filter, while the recognition rate will not be affected.

The proposed algorithm is completely implemented on a TMS32020 signal processor. The results are given in Section IV.

II. Split Levinson Algorithm and Its Numerical Analysis

When the speech signals are modelled as an AR model, they can be described with the following linear equations.

$$A_M(z) = \sum_{i=0}^{M} a_i z^{-i} \quad , \quad a_0 = 1 \qquad (1)$$

$$\mathbf{R}_M \mathbf{a} = [\sigma_M, 0, \ldots\ldots, 0]^T \qquad (2)$$

Where $\mathbf{R}_M = [r_{i-j}, 0 <= i, j <= M]$ is the Toeplitz matrix, $\mathbf{a} = [1, a_1, \ldots, a_M]^T$ the linear predictor coefficient vector, and σ_M the predictor error.

To solve the equations Delsarte and Genin [Del '86] introduced an algorithm, which is called the split Levinson algorithm. In contrast to the conventional Durbin-Levinson algorithm, which uses the reflection coefficients, the

new algorithm uses the singular predictor coefficients. The following equations illustrate this algorithm (eqs: 3 - 6; for k = 1,...,M).

$$\tau_k = \sum_{i=0}^{k} r_i p_{ki} = \begin{cases} \sum_{i=0}^{t-1} (r_i + r_{k-i}) * p_{ki} & , \text{ für } k = 2t-1 \\ \sum_{i=0}^{t-1} (r_i + r_{k-i}) * p_{ki} + r_t p_{kt} & , \text{ für } k = 2t \end{cases} \quad (3)$$

$$\alpha_k = \tau_k / \tau_{k-1} \quad (4)$$

$$p_{k+1,i} = p_{ki} + p_{k,i-1} - \alpha_k p_{k-1,i-1} \quad (5)$$

$$a_k = a_{k-1} + p_{M+1,k} - \lambda_M p_{M,k-1},$$

$$\text{with } \lambda_M = \sum_{k=0}^{M+1} p_{M+1,k} \Big/ \sum_{k=0}^{M} p_{M,k} \quad (6)$$

Where $\tau_k = \sigma_k / \lambda_k$ is modified predictor error, r_i the autocorrelation functions, α_k the recursive parameter, and p_{ki} the singular predictor coefficients.

The cepstrum coefficients can then be directly computed from the predictor coefficients with the following equation (for m=1,...M):

$$c_1 = a_1 \qquad c_m = a_m - \frac{1}{m} \sum_{k=1}^{m-1} k c_k a_{m-k} \quad (7)$$

Compared with the Durbin-Levinson algorithm the new algorithm only needs about half of the multiplications and nearly the same additions.

The numerical stability of the split Levinson algorithm is similar to that of the Durbin-Levinson algorithm. The values of the recursive parameters α_k in the new algorithm lie between 0 and 4, this guarantees the stability.

The computer representation precision of different parameters as well as their quantization precision in Durbin-Levinson algorithm have been thoroughly studied in the literature [e.g.: Mar '76]. So the analysis of the split Levinson algorithm can be simply carried out through the comparison of the numerical characteristics of the two algorithms. Figure 1 shows the histograms of singular predictor coefficients for a given speech signal. A piece of speech signal with length of 19.08 s was taken in this analysis. The window has a length of 30 ms (480 samples) and moves in a 15 ms frame rate (50% overlap). Because of the symmetrical characteristic of singular predictor coefficients only the coefficients p_1 to p_7 are illustrated in the Figure.

Histograms of the predictor coefficients and of the cepstrum coefficients calculated with the same speech signal and the same bound conditions are shown in Figure 2 and Figure 3, respectively. They provide the comparison basis because the predictor coefficients and the cepstrum coefficients possess the same histograms in the two algorithms.

It can be seen from Figure 1 to Figure 3, that singular predictor coefficients possess nearly the identical histogram distribution as the predictor coefficients. This means that the split Levinson algorithm can be carried out with the same computation precision.

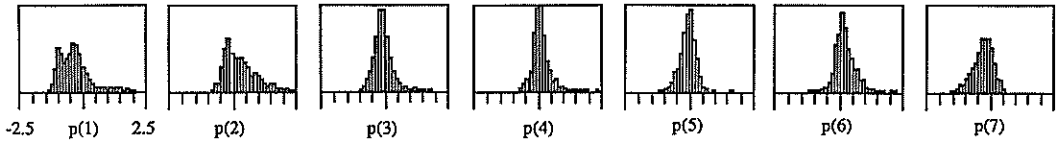


Figure 1: Histograms of the singular predictor coefficients

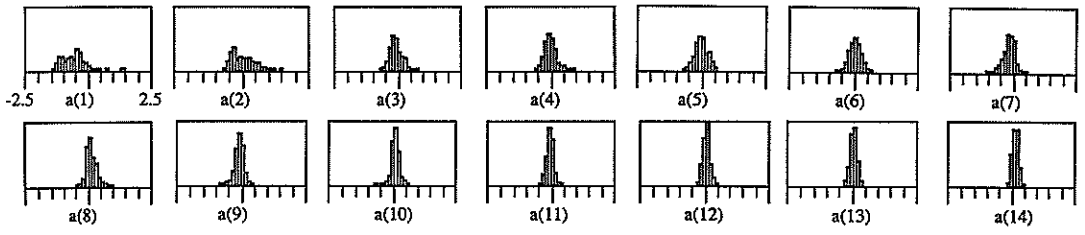


Figure 2: Histograms of the predictor coefficients

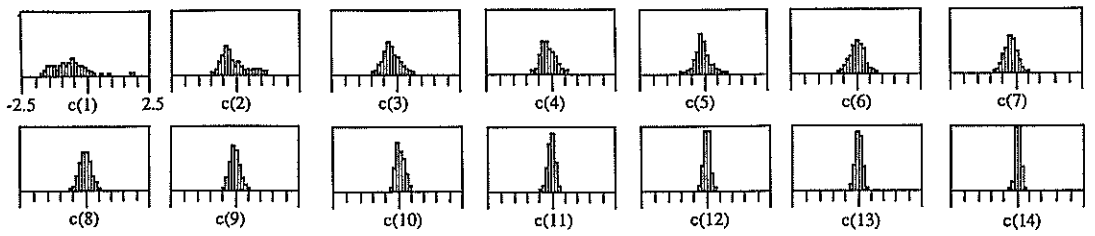


Figure 3: Histograms of the cepstrum coefficients

III. Influences of Different Factors on the Recognition Rate

At first the analog signal bandwidth of the recognizer has to be fixed. A broader analog bandwidth enables an exacter and better representation of the speech phonetics. On the other hand the computation cost of the recognizer increases with growing sampling frequency. Consequently the sampling rate should be hold as low as possible, while the recognition rate is not dramatically affected.

Figure 4 shows the influences of the sampling rate and the antialiasing filter on the average recognition rate and its standard variance. At point b and point c the speech signals have an 6,9 KHz analog bandwidth and are sampled with two sampling rates (16 KHz and 8 KHz), respectively. It shows that the higher sampling rate will not yield a significantly better recognition rate. Also interesting is that the recognition rate of the system without antialiasing filter is better than that with such one (compare point a and point b). This means that the analog bandwidth does not have to be changed with the decreased sample rate.

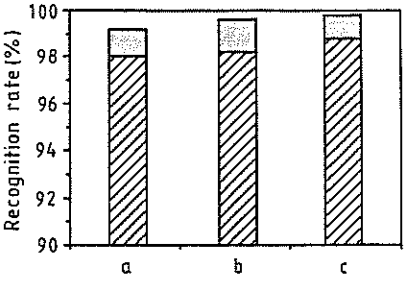


Figure 4: Average recognition rate;
Standard variance

Figure 5 shows that the conventional Hamming window and greater overlap grade (>= 50%) for speech signal do not contribute significantly to the recognition rate. It can be replaced by rectangular window with smaller overlap grade or without it.

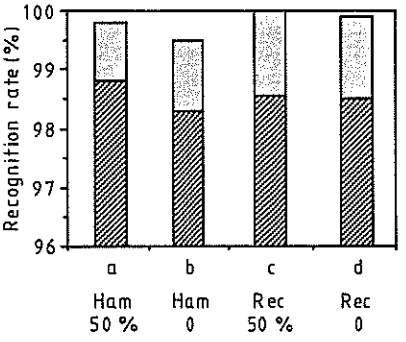


Figure 5: Average recognition rate
Standard variance

The optimal segment number for isolated word recognition is experimentally investigated. Figure 6 represents the influences of different segment number

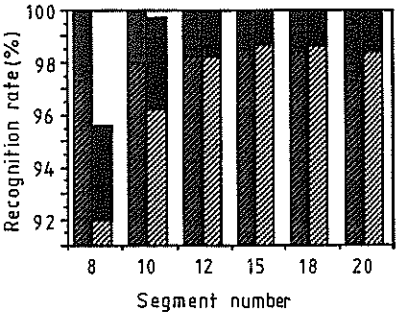


Figure 6: Average recognition rate
: shorter words
: longer words and
standard variance

compressed from different pattern lengths on recognition rate (average pattern length by shorter words: 32 intervals by 30 ms window size; average pattern length by longer words: 47 intervals). It shows that a unified segment number for different pattern length falls between 15 and 18. A compression rate from 1/2 to 1/3 can be achieved.

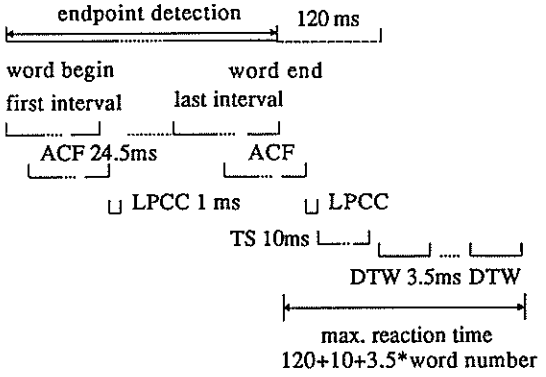
A complete analysis of all the other factors involved in the system implementation can be seen in [Liu '89]. The results showed the recognition rate will not be strongly influenced by the fluctuation of the factors - provided the parameters vary in tolerant ranges. So the design of the recognizer can be focused on achieving a configuration with a favourable computation expense.

IV. System Implementation

The complete algorithm was first simulated with a fixed point arithmetic (32 bit intermediate results and 16 bit end results) on VAX 11/730 and then implemented on a TMS 32020 signal processor. The intervals are discarded, when either overflow or instability of the predictor filter appears. In experiments the discarded intervals lie under 0.05%.

The algorithm includes the following steps: (1) Detection of the beginning and end of words; (2) extraction of speech parameters; (3) compression of the parameter vectors using Trace Segmentation and (4) pattern comparison using dynamic time warping. The time course of the algorithm on the TMS 32020 is shown in Figure 7.

The ACF calculation needs ca. 80% of the whole computational time. To detect the speechless intervals within a word, a time delay of 120 ms is considered after the word end. Ca. 130 patterns can be directly matched within 500 ms with the recognizer.



ACF: Autocorrelation functions; LPCC: Cepstrum coefficients;
TS: Trace segmentation; DT: Dynamic time warping

Figure 7: The time course of the algorithm

The 120 ms waiting time can be used to do the subsequent calculations. If no pause is later detected, the results are valid, otherwise they will be recomputed. This reduces further the reaction time of the recognizer.

In the presented recognition system ca 220*16 bit memory are required per reference pattern.

V. Acknowledgment

The work presented here is a part of the author's doctor thesis at the Technische Universität Berlin, Institut für Fernmeldetechnik. The author would like to thank her supervisor, Prof. Dr.-Ing. K. Fellbaum for instructive discussions and kindly supports.

VI. References

[Del '86] Delsarte, P.; Genin, Y.V.: The Split Levinson Algorithm. IEEE vol. ASSP-34, No. 3, June 1986, p. 470-478.

[Liu '89] Liu, J.: Zur Untersuchung und Optimierung von Spracherkennungssystemen für isoliert gesprochene Wörter. Dissertation, Technische Universität Berlin, VDI Verlag: Reihe 10, Nr. 111. 1989.

[Mar '76] Markel, J. D.; Gray, A. H. Jr.: Linear Prediction of Speech. Springer-Verlag Berlin Heidelberg New York 1976.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

FAST AND ACCURATE SPEAKER INDEPENDENT SPEECH RECOGNITION USING STRUCTURAL MODELS LEARNT BY THE ECGI ALGORITHM.†

Francesc Torró *††, Enrique Vidal*, Héctor Rulot**.

* D.S.I.C., Universidad Politécnica de Valencia (Spain).
** Centro Informática, Universitat de Valencia (Spain).

While the Error Correcting Grammatical Inference procedure has proved to be a very precise structural modelling technique for speech objects, its effectiveness in practical applications was often limited by the rather large number of states of the learnt models. We present in this paper some improvements on the Viterbi-based recognition procedure of ECGI, which takes advantage of the very sparse nature of the learnt models to dramatically cut down the number of actually visited states; this puts the ECGI technique in favorable comparison with respect to other more conventional techniques of Automatic Speech Recognition.

1.-Introduction.

The ECGI (Error Correcting Grammatical Inference) algorithm is a recently introduced automatic learning method [1] which was designed to achieve an "abstraction ability" that aimed at capturing all the relevant variability which is exhibited by the concatenation of local substructures of the considered patterns, as well as by the lengths (durations) of these substructures. The ECGIA was designed to automatically obtain stochastic structural finite-state models of speech objects (phonemes, syllables, words, etc.) from samples of these objects. In previous papers [1-3], the capabilities of this algorithm were demonstrated through successive series of experiments in which its different features were successively implemented and applied to increasingly difficult speech recognition problems. The results have shown a good performance of ECGI in obtaining structural models of isolated words which enable a very high speaker-independent recognition rate to be achieved. In this paper we present new developments of ECGI, which are mainly oriented to improve the computational speed of the recognition procedure. Also, we report on the results of the experiments carried out to test these improvements.

The most reliable results on the accuracy that is achievable by ECGI are presented in [3] where a recognition rate of 99,80% was obtained in a speaker-independent recognition task, corresponding to a 800 word training set (Spanish digits uttered by 10 male/female speakers) and a effective test set of 1000 utterances from different speakers. The average net size of the word models was 196.5 states in this experiment. The recognition performance of ECGI has been also verified in other more difficult tasks, including speaker-independent recognition of the Spanish alpha-set with a (not previously reported) recognition rate of 73.8%.

† Supported in part by Spanish CICYT, under grant TIC89/0448.
†† Supported by a Spanish MEC postgraduate grant.

These results clearly indicate that the recognition accuracy of ECGI is more than sufficient for practical applications. However, the computational cost of the basic Viterbi algorithm that is associated with ECGI recognition can be somewhat higher than other competitive approaches, such as (multitemplate) Dynamic Time Warping (DTW), or certain approaches which are based on Hidden Markov Modeling (HMM). This has led us to search for different techniques to reduce the computational cost.

The first consists of reducing the complexity of the learnt models by pruning out those states and/or transitions which are the "least important" in some specific sense; e.g. those which have rarely been used in the parsing of the training set. Though this technique is currently under investigation, it seems to indicate that it is possible to prune out more than half of the states, without sacrificing the recognition performance. This approach has also been exploited as a method for automatically learning the topology of Hidden Markov Models from training speech samples [4].

The other approach, which is the aim of this paper, capitalizes on the very "sparse" structure of the finite-state networks supplied by the ECGIA. This suggests that the Viterbi-based ECGI recognition procedure could be conveniently modified so as to reduce the overall effective number of states to be searched through the trellis.

2.-The ECGI algorithm.

The ECGI algorithm automatically learns a Regular Grammar or its corresponding Finite-State Acceptor (FSA) through an incremental procedure which considers the training strings one after the other, with no repetitions.

Initially, a trivial FSA is built from the first training string. For each new training string, the algorithm finds the best alignment between this string and the closest string in the language which is recognized by the current acceptor. To determine such alignment, a non-stochastic *error-correcting parsing* is performed by means of a Viterbi-like (Dynamic Programming) procedure [5]. The sequence of error transitions in the optimum alignement path lead, in the *construction phase* of ECGI, to the addition of a minimum number of new states and/or transitions to the current acceptor in order to accept the new training string with a minimal modification of the acceptor. Simultaneously, a frequency-of-use count is updated in each non-error state and transition of the optimum alignment, in order to obtain the probability of each state and transition at the end of the learning process. A similar frequency count is also performed for the substitution error transitions, that have been used for this alignment, to obtain an estimation of their probabilities. These probabilities are utilized in the recognition phase. A more detailed description of ECGI can be found in [1-3].

To perform recognition, the ECGI uses a similar Viterbi procedure as that used in the construction phase; but in this case a *stochastic parsing* is done, using the estimated probabilities. Recognition of a given test-string then consists of maximum likelihood deciding the class for which the accepting probability is the greatest.

3.-Improving the complexity of the recognition task.

While ECGI has already been shown to achieve high recognition accuracy, the large size of the learnt models was often considered a drawback in the recognition phase.

Let $A_{sf}=(Q,\Sigma,q_0,F,\delta,P_\delta)$ a (nondeterministic) stochastic FSA in which Q,Σ, and F are finite sets of states, input symbols, and final states, respectively; q_0 is the starting state; and δ is the set of state transitions to which P_δ assigns a set of probabilities. Given a test sample x, the Viterbi Algorithm (VA) builds the associated trellis and determines, at every stage $j=1..|x|$ the current best path that reaches each state; that is

$$\underline{\text{for}}\ j=1..|x|\ \forall q\in Q\ P_q=\max_{\forall q'\in Q}(P'_{q'}\ p(q',x_j,q)) \quad (1)$$

where $|x|$ is the size of the test sample, P'_q and P_q are the probabilities of reaching the state q at the previous and current stages, respectively, and $p(q',x_j,q)$ the probability of going from q' to q with the symbol x_j which takes into account the substitution error probabilities, i.e. $\forall (q',c,q) \in \delta$, $p(q',x_j,q) = P_\delta(q',c,q)\cdot p_{subs}(c,x_j)$ where $P_\delta(q',c,q)$ is assumed to be 0 if $(q',c,q)\notin\delta$); and $p_{subs}(c,x_j)$ is the probability of substituting c for x_j.

With this direct approach, a quadratic time complexity arises with the number of states $|Q|$. The practical impact of this complexity is especially high with the rather large ECGI models (about 200 states per model in Spanish digits task). Moreover, this is a very wastful procedure, particularly for the very sparse structure of ECGI-learnt models. Fortunately, this point can be easily improved as will be explained below.

First of all, we take advantage of the linked structure of models supplied by ECGI in order to reduce the "$\forall q'\in Q$" loop of (1) that is required for maximization, to "$\forall q' \in \{ q\in Q \mid (q',a,q)\in\delta\ a\in\Sigma \}$", i.e. to only those states which can be predecessors of q. It must be noted that all the subsequent references to complexity reduction (visited state percentages) will be compared with the computational cost of this approach, which is proportional to $|Q|\cdot B$, where B is the average branching factor of A_{sf}.

Secondly, the Viterbi multistage search is organized so that, in each stage, only those states which have been actually "reached" (with non-zero probability) from the previous stage need be considered. This is achieved by actually performing a *foreward search* and structuring those states which are reached at every stage into an appropriate set data structure. The adopted structure consists of a list that is embedded onto an array. This allows both the "for every active state" loop to be performed without needlessly visiting non-reached states, and to know in a single time step whether a state is already included in the current list by just accessing it through the corresponding vector index.

This strategy is very significant with the ECGI models whose branching-factor is usually very low, 1.86 in average for the same task as above. For such a task this leads to an overall 82% saving in the number of actually visited states.

Not only this approach results by itself in quite profitable savings, but, what is more important, the well known *beam-search* technique can now be easily applied to achieve further reductions without sacrificing the recognition accuracy. This technique [6] consists of only searching a few "best" paths in parallel, i.e. those paths whose probability is higher than a threshold defined as a function of the highest probability at each stage. More precisely, these states (the "beam") are $\{ q\in Q \mid P_q > P_{max}/\alpha \}$ where P_{max} is the highest probability of a state at the considered stage, and α is a constant which defines the beam width. This constant is determined empirically to be large enough to avoid losing the best global path at any time. (In fact, we use log probabilities so that the beam is actually defined as $\{ q\in Q : \log P_q > \log P_{max} - \alpha \}$ with $\alpha > 0$).

There is, however, one small technical problem with this heuristic; namely P_{max} is not known until all the candidate states have been considered. This is solved, as in [6], with a double linked list whose first position is

always reserved for the highest current state probability. The behavior of this approach has also been checked empirically, obtaining a negligible number of extraneous states placed in these pseudo-best lists, as compared with those placed by using the original exact criterion.

Finally, a new improvement in recognition speed can be achieved if we take advantage of the beam-search heuristic and search through only a single multistage graph ("hypertrellis"), rather than searching sequentially through each class trellis. Thereby, when a given utterance of a word is analyzed, the beam tends to be quickly centered in the slice of the trellis corresponding to the class it belongs to. The bound effect of the beam is, thus, very much beneficial, because those classes (words) which are very different from that of the uttered word tend to be discarded very soon.

Figure 1 represents the behavior of the beam during the recognition of a test utterance belonging to the Spanish word /seis/ ("six") when the beam width is defined by α=30. The "hypertrellis" is built bottom up, i.e. the first stage is at the bottom and the trellis advances upwards. The states at every stage are grouped by class and arranged in increasing order (first the word /cero/ ("zero"), then /uno/ ("one"), and so on). Due to the properties of the models that are learnt by ECGI, the states that are reached at every stage always have higher identifiers than those at the previous one. On the other hand, the height of a point in a line (or stage) represents the log probability (normalized by the stage number) of reaching the associated state.

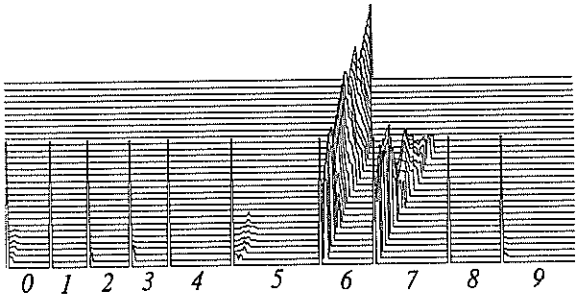


FIG 1. "Hypertrellis" evolution during the recognition of the Spanish word /seis/ ("six") when α=30.

4.-Experiments and results.

The successive speed improvements which are discussed above have been tested in a Spanish Speaker-Independent Digits Recognition task. The corpus comprised a total of 1000 utterances obtained from 10 speakers (5 male and 5 female) with 10 repetitions per speaker and digit. These speech data were preprocessed as in [3] and partitioned in a training set of 600 utterances from 6 speakers and a test-set of 400 utterances from the remaining speakers. The computation reductions that were achieved by the introduced techniques are summarized in Table 1.

Technique	states/model	percentage
Original.	197	100 %
Set of reached states.	35	18 %
Sequential beam-search.	28	14 %
Hypertrellis beam-search.	18	9 %

TABLE 1. Average number of states that are visited in each stage for the same error rate (0.25%).

Figure 2 summarizes the computation reduction effect of the beam depending on its width, which is defined by the threshold α (horizontal axis) that determines the beam tolerance. The behavior of the error rate (E) is exponential, dropping quickly for the first values of α and more softly later. On the other hand, the increase of the rate of visited states (T) is nearly lineal and, in consequence, the joint evolution is the desired one.

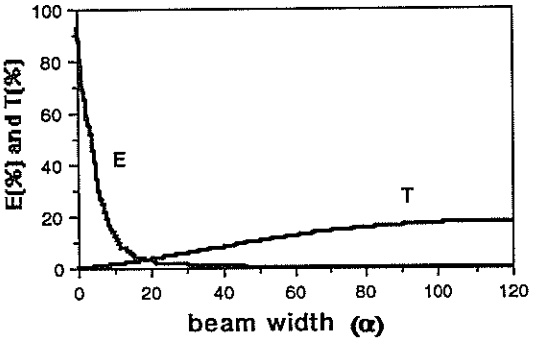


FIG 2. Error Rate (E) and Rate of Visit (T) as a function of beam width. Note that the original error rate (E=0.25) is maintained until T=9%, which corresponds to 18 states that are visited on the average.

5.-Final remarks.

The results presented in the previous section clearly indicate that important computation reductions can be achieved from the basic Viterbi recognition procedure of ECGI. However, the finite-state models learnt by ECGI usually have a rather large number of states and it should be convenient to put the computation costs of ECGI recognition, as reported in this paper, in a comparison with other more standard techniques. In this respect it is worth noting that the 35 states which are visited on the average by applying the basic "set of reached states" search technique is equivalent to the computational effort which is demanded by the Hidden Markov Model (HMM) approach if 3 or 4 models per word are used. Some results on the accuracy of single-model-per-word HMMs in the Spanish Digits Speaker Independent recognition task have been reported in [4]. These results are worse than those achieved by the ECGI in the same task, and we argue that, in fact, more than 3 HMM per word (appropriately trained through a convenient Clustering procedure, e.g. [7]), whould possibly be needed to attain the ECGI accuracy.

Therefore, with the basic set-of-reached-states technique, the ECGI seems to compare quite well with HMM, with the advantage of making the use of any clustering procedure, with its associated drawbacks of computational cost and need of experimental tuning to determine the appropriate number of models, unnecessary.

Moreover, this is not all, and the basic set-of-reached-states approach quite naturally invites to implement the beam-search technique in order to further increase the recognition speed. Finally, the use of the beam-search technique has suggested us a new improvement which we call "*anticipated recognition*" [8]. This technique consists of prematurely aborting the Viterbi beam-search whenever a stage is arrived at in which all the active states belong to the same word model. Using this technique, preliminary results seem to indicate that only 6 states (3%) need to be effectively visited on average to keep the recognition accuracy at the same level than in the original full search procedure.

All the results clearly indicate that both the recognition accuracy and the speed of the ECGI technique present significant advantages over other more conventional techniques usually adopted for discrete utterance Speech Recognition.

References.

[1] H.Rulot, E.Vidal. "*Modelling (Sub)string-Length Based Constraints through a Grammatical Inference Method*". In 'Pattern Recognition Theory and Applications', P.A.Devijer and J.Kittler eds, Springer-Verlag, pp.451-459, 1987.

[2] H.Rulot, E.Vidal. "*An efficient algorithm for the inference of circuit-free automata*". NATO ASI Syntactic and Structural Pattern Recognition., G.Ferrate et al eds, Springer, 1988.

[3] H.Rulot, E.Vidal, N.Prieto. "*Learning accurate finite-state Structural Models of words through the ECGI Algorithm*". ICASSP-89, pp.317-321, 1989.

[4] F.Casacuberta, E.Vidal, B.Mas, H.Rulot. "*Learning the structure of HMM's through Grammatical Inference techniques*". To be published in ICASSP-90.

[5] G.D.Forney. "*The Viterbi Algorithm*". IEEE Proc., vol.61, no.3, pp.268-278, 1973.

[6] B.T.Lowerre."*The Harpy speech recognition system*".Ph.D.Thesis, Carnegie Mellon Univ., 1976.

[7] L. R. Rabiner, C. H. Lee, B. H. Juang, J. G. Wilpon. "*HMM Clustering for Connected Word Recognition*". ICASSP-89, pp. 405-408.

[8] F.Torró. "*Estudio de alternativas en la reducción de la complejidad del Algoritmo de Reconocimiento basado en el método ECGI*". Proyecto fin de carrera, Fac. de Informática, Univ. Pol. de Valencia (Spain), 1989.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Evaluating a Grammar as a Language Model for Speech

R.A.Sharman

IBM(UK) Science Centre, Athelstan House, Winchester, SO23 9DR, England

Language Models for use in Speech Recognition are commonly based on finite-state machines which predict likely word sequences. Such *n-gram* models for Natural Languages are effectively restricted to no more than 3 word sequences, by both computational limitations, and by the ability to make accurate estimates of the likelihood of word sequences from observation of real data. It is often conjectured that a *grammar* of a Natural Language should be superior to 3-gram models, since it could capture more of the constraints of a language, and thus might achieve a lower entropy, closer to the true entropy of the language. This idea is tested with reference to a *context-free grammar* of English (derived from a large annotated corpus), which has a wide coverage. It is found that the entropy of the grammar derived is about equivalent to the entropy of a 3-gram model derived from the same data. The question of how to develop a grammar of English which has significantly lower entropy than a 3-gram model is considered.

A *language model* is a device which can perform certain linguistic tasks, such as the ability to predict the next word, given the preceding words. This capability is required for accurate speech recognition in order to limit the search among the many possible candidate words suggested by the given acoustic evidence [4]. *n-gram* models, which use the probability of sequences of *n* words, have been found very effective as language models for speech recognition in English and other languages. They are readily derivable from data [10]; have desirable computational characteristics [5]; can be easily compared by determining their entropy [12]; and can be evaluated using simple algorithms [8] which belong to a class of problems (the EM-algorithms) for which optimal solutions are known [2].

The problem with *n*-gram modelling for natural languages is that as either *n* increases, or the number of vocabulary items increases, the number of parameters of the model also increases rather sharply. This results in a model which is either impracticably large, or whose parameters cannot be accurately estimated. For example, a vocabulary of 5000 words (itself not a realistically large vocabulary of words for a language such as English) would require 5000 probabilities for a 1-gram model, 25 million probabilities for a 2-gram model, and 125 billion probabilities for a 3-gram model. Not only would this require more computer storage than is available on any reasonably priced current or future hardware, but the reliable estimation of all these parameters would require unrealistically large amounts of training data to be observed.

It is often conjectured that a language model based on a grammar, such as a *context-free grammar*, should have better predictive properties for a Natural Language since the grammatical phrases effectively accumulate arbitrary length histories of the text. This should enable a grammar to capture more of the constraints of the language than a simple finite-state automaton could. In order to test this conjecture it is necessary to find some context-free grammar which can parse the whole language, and calculate its entropy. It can then be objectively compared with *n*-gram models, whose entropy is already known. However, this is difficult in practice since no effective context-free grammar of unrestricted English is available [9]. However, a new technique (described below) enables an approximate grammar to be obtained, and thus an indication of the relative merits of grammars and *n*-gram models to be ascertained.

The entropy of language models

A Natural Language can be modelled as an ergodic Markov source [11] which produces symbols in sequences forming sentences of the language. Let x be a symbol in some symbol set. The symbols can be letters, groups of letters, words, or phrases. The probability that x will occur is $Pr(x)$. It can be shown from Information Theory [6] that the proper measure for the information contained in a symbol is the *self-information* of the symbol, which is $-\log_2 Pr(x)$. Logarithms to base 2 are convenient for interpreting information in terms of bits, and will be assumed throughout.

The average amount of self-information in symbols emitted from a source is the *entropy* of the source. If the symbols are taken from a symbol set of size m, denoted by $x_1, x_2, \ldots, x_m$, and the symbols are independent of each other, the entropy of the source emitting the symbols, H, is defined by:

$$H = -\sum_{i=1}^{m} Pr(x_i) \cdot \log Pr(x_i)$$

The entropy of English considered as a source outputting single letters independently, according to the relative frequency of letters as they are known to occur, is the *first-order entropy*, H_1. It is found by setting $m = 27$ (there are 26 letters and one blank symbol) and using the single letter probabilities.

However, knowing the dependencies between letters is an example of introducing an extra constraint into the model. The model of English as a source which outputs letter pairs is the *second-order entropy*, H_2,

$$H_2 = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} Pr(x_i, x_j) \cdot log\ Pr(x_i, x_j)$$

In the limit, as the block of letters which comprises a symbol increases, this function will approach the true entropy of the source. Thus, the entropy of the language, H_∞, is found as the block length, n, increases to infinity, or

$$H_\infty = -\lim_{n\to\infty} \frac{1}{n} \sum Pr(x_1,...,x_n) \cdot log\ Pr(x_1,...,x_n)$$

where the sum is taken to be over all possible symbol sequences of length n. Thus the numbers $H_1, H_2,...,H_n,...,H_\infty$ form a sequence which is a scale of approximations tending in the limit to the true entropy of the language. Any other language model can be compared to this scale if its entropy is known. If, for example, a particular language model has an entropy equal to H_9 then it could be said that this language model knows exactly the same amount of information about the language as is obtained by a knowledge of all the 9-letter sequences in the language. In this way radically different kinds of language model can be tested and compared.

Perplexity as a measure of a language model

Assume that some source, S_1, has entropy, H, and has $|S_1|$ symbols which are not equi-probable. There must be another source, S_2, which has $|S_2|$ symbols, and has the same entropy, H, such that the $|S_2|$ symbols are equi-probable.

But the number of symbols of a source of entropy, H, is 2^H if the symbols are equi-probable. So the number of symbols of the equivalent equi-probable source can be used as a normalised measure for any source with non-equi-probable symbols of a given entropy, for the difficulty of predicting the source. The size of the corresponding equi-probable source, S_2, is the *perplexity*, P, of the source of entropy, H, or $P = 2^H$.

Thus an entropy of H implies that H bits are needed to code each letter, on average, whereas the perplexity is the chance of guessing the next letter by rolling a P-sided die.

Computing higher order entropies

The value of the entropy of a natural language is in practice unknowable, since it requires knowledge of an infinite number of word sequences. However, it will be possible find an estimate of the value by successive approximation using larger and larger groups of words. In order to do this, it is necessary to observe a long sequence of words actually occuring in the language, and to count the frequencies with which all the various letter combinations appear. From these frequencies it is possible to calculate the probability of each symbol, and thus evaluate the formula given above for the entropy of the source, by the direct evaluation of the formula above.

Because of the large number of possible sequences of symbols the direct computation of entropy in this way has usually been stopped at H_3 [1,10,11], although some attempts to reach H_5 have also been made [12]. Each successive model is expected to produce a better approximation to the true entropy of the language, since more and more of the constraints of the language are being incorporated into the model.

In a natural language there are larger blocks of letters which form readily identifiable symbols, such as words and phrases. If the probability of words is known, the entropy of the single word model, H_w, can be calculated from an observation of word frequencies. Zipf's empirical relationship for word frequencies, $Pr(x_i) = 0.1/i$, has often been used [1,11], in the absence of extensive knowledge of actual word frequencies. Mandelbrot's modified formula has also been used [1].

The entropy of English as a source outputting words, H_w, can be compared to the values obtained for English as a source outputting letters, as before, by assuming that for some k,

$$H_k = H_W/\overline{w}$$

where $\overline{w}$ denotes the average length of a word. Because the letters appearing in words are not indpendent this is at best only an approximation. Since it is observed from large corpora of natural text [9] that there are 5.3 letters per word, on average, this means that it is expected that $H_W \simeq H_{5.3}$. This fact is attested by a comparison of the results of [12] and [1,11]. Similarly, if the frequencies of two word groups, and three word groups are known [9] it is possible to obtain approximations to $H_{10.6}$ and $H_{15.9}$, respectively.

Further, the entropy of English as a source outputting phrases, H_P, can be compared to the $H_1, H_2,...$ scale by assuming that, for some k,

$$H_k = H_P/\overline{p}$$

where $\overline{p}$ denotes the average length of a phrase. Because there are dependencies between the words appearing in phrases this is again only an approximation. What is the length of a phrase? From observation [9] again, it is found that there are 1.6 words per phrase, on average. This means that it would be expected that $H_P \simeq H_{8.48}$. This hypothesis is tested, below.

Testing a grammar Language Model

In order to obtain a value for the entropy of English, based on the likelihood of phrases a context-free grammar was constructed in the following way. One million words of the Associated Press (AP) newswire was analysed for grammatical phrases, using a natural phrase encoding system, based on native speaker intuitions [7]. The parsing was done manually, thus using the best parser currently available, namely, native speakers of English.

A systematic parsing scheme was employed, using a system of 286 word tags (classifying each word according to its part-of-speech). These pre-terminal tags are parsed into phrases labelled with a non-terminal symbol taken from a set of 64 grammatical classes selected according to traditional usage in English grammar, making a total of 350 non-terminal symbols. The non-terminal symbols are used to form parse trees over the real words, which come from a vocabulary of about 50,000 words, and form the terminal symbols of the grammar.

The observed frequency of phrases of each unique pattern can be used to calculate the probability distribution of phrases which is shown in Figure 1. The most common type of phrase is a *prepositional-phrase*, formed from a *preposition*, followed by a *noun-phrase*. Altogether, there are about as many unique types of phrase in the text as there are unique words in the text [9]).

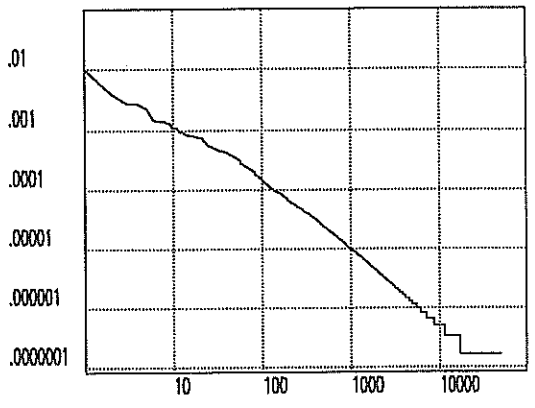


Figure 1. The frequency distribution of phrases in English: The probability of the phrase is plotted on the vertical axis, and the rank order of the phrase is plotted on the horizontal axis (logarithmic scales).

The relationship is similar to the well-known Zipf law observed for words. In fact, similar relationships are obtained for every choice of symbol granularity from single letters up to three word groups [9]. Phrases are evidently analysable as another type of block encoding of the language source.

The resulting phrases can be used to construct a context-free grammar of English by taking (1) the label on each phrase as the left hand side of a CFG rule, and the constituents of the phrase as the right hand side of that rule, and (2) a part-of-speech tag as the left hand side of a rule, and the word itself as the right hand side of that rule. This procedure derives a 100,000 rule grammar, which has 350 non-terminal symbols, and 50,000 terminal symbols (the vocabulary of the corpus). This grammar has the merit that it definitely parses at least one million words of randomly selected naturally occuring text (a good achievement for any grammar). In trials on unseen texts of the same sort it has been found that it can parse about 94% of sentences and thus has genuinely captured a significant amount of knowledge of English Grammar.

The value of the entropy can be calculated for each available symbol type (letter, word, and phrase) and from this the average amount of information per symbol can be estimated and compared. If the grammar derived above is a good language model then it should have a low value of information per word. In any case, a direct comparison can be made with known results for *n*-gram models. Combining previous results [1,11] with the analysis of this data, and computing all results by direct evaluation of the entropy equation, gives:

letters	symbol	Bits per letter	Bits per word	Bits per phrase
1	letter	4.03	21.36	34.17
2	two letters	3.32	17.59	28.15
3	three letters	3.10	16.43	26.28
5.3	word	1.95	10.35	16.56
8.48	phrase	1.16	6.17	9.88
10.6	two words	1.46	7.77	12.43
15.9	three words	1.20	6.36	10.17

It can be seen from this table, that the amount of information per letter decreases as the size of the symbol increases. This is due to the fact that the larger symbols embody more constraints in the use of language, and thus require less information to be supplied by the symbol itself. The following results are obtained for perplexity (of the larger sized symbol systems, only):

letters	symbol	letter	word	phrase
5.3	word	3.86	1305	96617
8.48	phrase	2.23	72	942
10.6	two words	2.75	218	5518
15.9	three words	2.29	82	1152

In other words, the 3-gram model is more than an order of magnitude better than the simple single word frequency model, in terms of the difficulty of choosing the next word. This demonstrates very graphically why 3-gram models are so effective in tasks such as Speech Recognition.

The information in a grammar based on phrases

The figure for the entropy of the phrase model needs some explanation, since it does not fit neatly in the progression of values obtained from the letter and word *n*-gram models, based solely on the average number of letters contained in a phrase. In particular $H_P < H_{10.6}$, and therefore $H_P \neq H_{8.48}$ which was the hypothesis proposed above.

It was stated above that a grammar-based model ought to incorporate more information about the dependencies in language, and therefore have a lower entropy, than a simple *n*-gram model, since it contains knowledge of higher level structures. In the case of the particular grammar used here the size of phrases varies between 1 and 30 words, and this evidently has an influence on the model, which is clearly better than the model based on blocks of letters of the same size as the average number of letters per phrase. Why is the

phrase model not also significantly better than the 3-gram model based on the same data?

The answer is not obtainable from this data, but appears to lie in the choice of symbols which make up the blocks of the coding system. In the grammatical coding system which was used here, a very general coding scheme was used, for speed and efficiency of annotation. It is evidently not the most efficient means of encapsulating the information content of the corpus in terms of a grammar. This scheme, as it is, constitutes a model of language which captures about as much information as a 3-gram model of the same data. Thus, if a parser uses this data as its grammar, it can expect results roughly comparable with that which would be obtained from using 3-grams. The present work indicates that if a grammar can be constructed which is more efficient that the present one, it should be able to exceed the capability of 3-gram models, though perhaps not by a great deal.

The question is: can any grammar based model of language be significantly better than simple n-gram models? Perhaps there are other grammars whose choice of non-terminal symbols constitutes a better coding scheme, which more closely approximates the true entropy of the language source. It is a conjecture, as yet unproven, that the feature-based grammar models of linguistic theories like GPSG [3] are attempting to build just such a coding scheme. The required task for grammar writers is to construct a complete grammar, with very wide coverage of the language, (similar to that of that of the grammar reported here) and then to calculate the entropy and perplexity of the new grammar. In this way an exact test of the power of grammar-based language models can be made.

References

1. Brillouin, L., *Science and Information Theory* (Academic Press, New York, 1956)
2. Dempster, A.P., N.M.Laird and D.B.Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Royal Statistical Society B, (1977) vol. 39 p1-88.
3. Gazdar, G., E.Klein, G.Pullum, and I.Sag, *Generalised Phrase Structure Grammar* (Blackwell, Oxford, 1985).
4. Jelinek, F., *The Development of an Experimental Discrete Dictation Recognizer* Proc. IEEE vol. 73, No. 11, Nov 1985.
5. Jelinek, F., *Self-organised Language Modelling for Speech Recognition*, (IBM Research Report, NY, 1985).
6. Jones, D.S., *Elementary Information Theory* (Clarendon Press, Oxford, 1979).
7. Leech, G., and R.G.Garside, *Running a Grammar Factory: the production of syntactically analysed or "treebanks"*, (Mouton de Gruyter, forthcoming, 1991).
8. Rabiner, L.R., and B.H.Juang, *An Introduction to Hidden Markov Models*, IEEE ASSP Journal vol. 2, no. 4, 1986.
9. Sharman, R.A., *Observational Evidence for a Statistical Model of Language*, UKSC Tech. Rep. 205, 1989.
10. Suen, C.Y., *n-Gram Statistics for Natural Language Understanding and Text Processing*, IEEE Trans. Patt. Anal. & Mach. Intell. (1979) vol 1, No.2.
11. Welsh, D., *Codes and Cryptography*, (Clarendon Press, Oxford, 1988).
12. Yannakoudakis, E.J., *An Insight into the Entropy and Redundancy of the English Dictionary*, IEEE Trans. Patt. Anal. & Mach. Intell. (1988) vol 10, No.6.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A TOP-DOWN DISCOURSE ANALYSIS IN A SPEECH DIALOGUE SYSTEM

Yasuhisa NIIMI and Yutaka KOBAYASHI

Department of Electronics and Information Science, Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto 606, Japan

This paper describes a method for the discourse analysis performed in the speech dialogue system we are developing. The purpose of the analysis is to provide the system with top-down discourse hypotheses. These are translated into linguistic constraints and used to reduce a search space in recognizing an utterance to come next. This effect of the discourse analysis has been proved by a simulation of the dialogue system using typed sentences.

1. Introduction

The recent advance of speech science and related technology has made it possible to build continuous speech recognition systems working in real time. Using such systems as an interface, we can construct man-machine dialogue systems [1],[2]. In the speech dialogue system, the discourse analysis, that is, the analysis of structures of dialogues plays important roles in interpreting utterances.

This paper describes a method for the discourse analysis performed in the speech dialogue system we are developing [3]. The purpose of the discourse analysis is to provide the system with top-down hypotheses on words likely to appear in utterances of the partner of a dialogue.

The task performed through the man-machine dialogue is to make plans, for example, plans for seeing the sights of a city. The system is supposed to have a relational database about the sights of the city. A user (speaker) can access to the system by voice and collect information necessary to make plans.

The speech dialogue system consists of three components; a speech interface, a dialogue controller and a planner. The speech interface recognizes utterances from users and passes their semantic interpretations to the dialogue controller. The dialogue controller analyzes structures of conversations and returns to the speech interface top-down expectations of what would be said next. The planner gives to the controller a guide line on how it should carry out a conversation with users as well as helps users to make plans by supplying information requested and suggesting some tour plans.

The discourse analysis is performed in terms of topics and discourse goals based on [4]. The transition of topics through a conversation is represented by an AND-OR tree. In this formulation the prediction of topics likely to be mentioned in the next utterance is equivalent to the determination of which node of the AND-OR tree is to be expanded next.

Each utterance of a dialogue has its own purpose (speaker's intention). The purpose of an utterance is called discourse goal. Since successive utterances are grouped to achieve a larger goal, discourse goals of a conversation can be described by a context free grammar, which we call discourse grammar. The top-down analysis of a sequence of utterances (a partial dialogue) by this grammar predicts discourse goals the next utterances could express.

A simulation of the dialogue system using typed inputs has proved that these top-down hypotheses can reduce a search space in recognizing utterances by about 60%.

U101 I'd like to do a few day sights in Kyoto.
S101 Which are you interested in, temples, gardens, or traditional artifact?
U102 I've heard Kyoto is famous for gardens.
S102 I see. I introduce two day tour for visiting gardens. The gardens in what eras would you like to visit?
U103 Muromachi, Momoyama and Edo.
S103 O.K. Famous gardens in Muromachi era are the stone garden of Ryoanji temple, Kinkakuji temple and Ginkakuji temple.
U104 What is the stone garden?
S104 It is made of a few large stones and white sand only. It is believed that the simplicity of the garden is related to "Zen".
U105 That's great! Who built it?
S105 I'm sorry, it is not known.
U106 Who built Ryoanji?
S106 It's believed that Katsumoto Hosokawa did.

Fig. 1 An example of the dialogue.

2. Representation of the context

Fig. 1 illustrates an example of dialogues which

the system would have with users. This example shows the conversation is developing in two different modes; system-initiative and user-initiative modes. In the system-initiative mode, the system, repeating questions to the user, elucidates the specification of his sightseeing tour, that is, a period of the tour, hotels to stay, places to visit and so forth, and then offers some candidates for these items. In the user-initiative mode, the user, inquiring detail information on these items, decides what are worth to involving in his plan. The questions issued by the user in this mode are interpreted as retrieval commands to the relational database. Responses of the system are generated based on the retrieved information.

2.1 Topic transition tree

It has been known [5] that topics in a goal-oriented dialogue move according to a task-dependent tree structure. In fact the topics in the illustrated example are specialized along the structure as shown in Fig. 2, which we call the topic tree. Nodes of the topic tree are entities related to the database such as names of relational tables, items included in the tables, and values of items. These correspond to topics the system can understand.

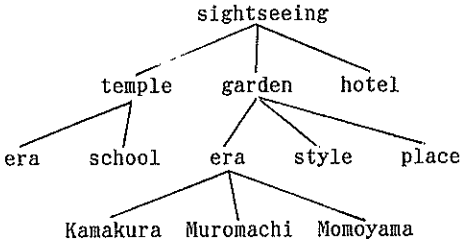


Fig. 2 A part of topic tree.

As we have reported in [3],however, an AND-OR tree is more suited for representing movements of topics than a simple tree. In the AND-OR tree, which we call a topic transition tree, AND-nodes represent topics introduced by the user, and OR-nodes represent topics introduced by the system. If the user inquires about two or more sights (each assumed to be a topic), the system must offer information on all of them. On the other hand, even if the system proposes two or more candidates for a visit, the user is not interesting in all of them, but might move to the other topics. An AND-OR tree is suited to reflect this difference. The topic transition tree is considered a trace of a subtree of the topic tree.

2.2 Discourse goals

Each utterance in a conversation has its own purpose(speaker's intention). In the dialogue illustrated in Fig. 1, an utterance U101 requires to make plans for sightseeing and presents information on a period of the tour. The following four utterances S102 to U103 have a larger purpose that the system tries to know what places the user wants to visit. The next utterance S104 proposes candidate sights to visit. In the utterances following S104, the user asks questions on those sights to judge which of them are worth to visiting, and the system provides some information on them. Thus discourse goals in a dialogue form a hierarchical structure like a tree.

This hierarchical discourse structure can be described by the context free grammar in which terminal symbols are discourse goals corresponding to a single utterance and nonterminal symbols are larger discourse goals corresponding to a group of utterances. Fig. 3 shows an example of the grammar for the discourse analysis. Underlined strings indicate terminal symbols. For example, a terminal symbol 'prst-alt' (present alternatives) represents a discourse goal of an utterance, like S102 in Fig. 1, used to present multiple choices, and a terminal symbol 'slct-alt' (select alternatives) represents a discourse goal to select one of alternatives presented. 'rqst-spec' (request for a part of specification) is a discourse goal of wh-questions issued by the system, and 'ans-spec' (answer a part of specification) works as an answer to 'rqst-spec'.

An utterance could have different discourse goals in different contexts, and a discourse goal could be expressed by various forms of utterances. Thus the relation between utterances and discourse goals is many to many correspondence.

```
(1) mk-plan     --> exm-spec,exm-plan
(2) exm-spec    --> exm-spec,exm-spec ¦
                    dcd-spec,exm-spec ¦ dcd-spec
(3) dcd-spec    --> prst-alt ¦
                    prst-alt,slct-alt ¦
                    prst-alt,chng-spec ¦
                    rqst-spec,ans-spec ¦
                    rqst-spec,chng-spec
(4) exm-plan    --> exm-plan,exm-plan ¦
                    {rqst-cand,}
                    prst-cand,exm-plan-1
(5) exm-plan-1  --> dcd-cand ¦
                    exm-plan-1,exm-plan-1
(6) dcd-cand    --> (aqr-knwldg)*{,response}
(7) aqr-knwldg  --> rqst-knwldg,ans-knwldg ¦
                    recommend
(8) response    --> accept ¦ reject
```

Fig. 3 A subset of rules for the discourse analysis.
{ } indicates an optional term, and (x)* indicates a null string or the repetition of x as many times as necessary. Rule (2-3), for example, refers to the second rule with the third alternative of the right hand side.

3. Analysis of discourse structure

Fig. 4 shows the flow of the discourse analysis. It involves bottom-up and top-down analyses. The bottom-up analysis performs the semantic analysis of an input utterance and then makes the bottom-up hypotheses, that is, candidates for topics and discourse goals of the utterance. The top-down analysis predicts topics and discourse goals likely to appear in the current utterance referring to the context so far restored which is represented by the AND-OR tree of topics and the parsing history of discourse goals. The bottom-up hypotheses are matched against the top-down predictions. The best match gives the interpretation of the utterance, which is preserved as contextual information.

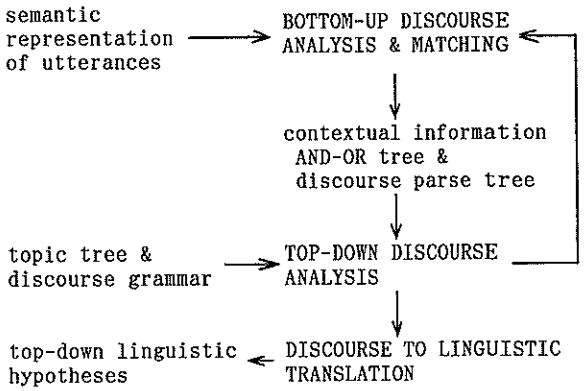


Fig. 4 Flow of the discourse analysis.

3.1 Bottom-up Analysis

The first stage of the bottom-up analysis is the semantic analysis of utterances, while the syntactic analysis of them is supposed to finish in the speech interface. It is performed based on the case grammar. Case frames associated with verbs are used to represent the meaning of sentences in the case grammar. They are described by a set of slots, each indicating one of relations between the verb and a noun phrase, like an agent, object and instrument. The semantic analysis assigns noun phrases included in an utterance to some slot of the case frame of the main verb based on semantic markers of the noun phrases. The semantic interpretation of an utterance is represented by a list of four terms, a main verb, a case frame with slots filled, aspect information and the style of a sentence.

The head nouns of the case slot fillers (noun phrases) are proposed as candidates for the topic of the utterance being analyzed. Those of cases of topic, object and purpose are given a higher priority than others.

A lexicon is prepared to make bottom-up hypotheses on discourse goals. It contains the relation among a verb, aspect, the style of a sentence including the verb, and a discourse goal which the verb could express. Bottom-up hypotheses on the discourse goal are built up by consulting this lexicon. It is generally difficult to to uniquely determine the topic and discourse goal of an utterance only by the bottom-up analysis.

3.2 Top-down Analysis

As mentioned in section 2.2, a grammar for the discourse analysis is formulated by the context free grammar. Thus an analysis of the conversation so far carried out results in a tree structure. Leaves of a discourse parse tree correspond to utterances. Fig. 5 shows an example of the top-down discourse analysis. It illustrates a discourse parse tree resulting from utterances U101 and S102 shown in Fig. 1 and discourse goals possible to be expressed by an utterance following S102. Rules used in the analysis of utterances U101 and S102 are rules (1), (2-1), (2-3), (3-1), (2,2), {(3-2) or (3-3)}. The last stage of the rule applications is ambiguous and incomplete. It is ambiguous in that there are two applicable rules (3-2) and (3-3), and incomplete in that all terminal symbols of the right hand sides of the applied rules are not consumed. Thus the next discourse goal would be 'slct-alt' if the rule (3-2) is applied and 'chng-spec' if the rule (3-3) is applied.

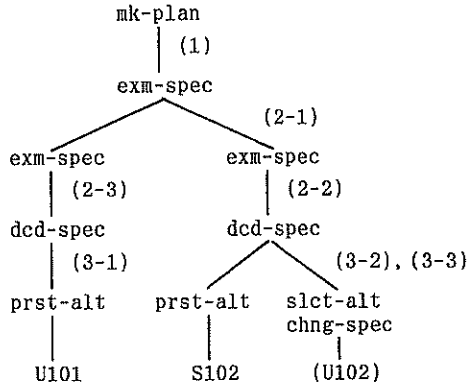


Fig. 5 An analysis of a discourse structure and a prediction of discourse goals.

The movement of topics through a conversation is stored in the topic transition tree. Thus for the top-down prediction of topics it is necessary to determine which node of this tree is to be selected or expanded next. New nodes which will result from the expansion can be known by referring to the topic tree because the topic transition tree is considered a trace of a sub-

tree of the topic tree. In this connection, the concept of a focus is very important. The focus means the topic currently focused. We define it as the last of topics which either of the system or the user has uniquely mentioned.

Which node is to be selected or expanded depends on the discourse goal of the next utterance. Fig. 6 shows the relation between discourse goals and topics in the top-down prediction. In order to shift the focus to nodes at higher levels than the currently focused node, it is necessary that all the topics under the new focus have terminated. The conditions for a topic to terminate can be stated as follows.

(1) A node with AND successors can terminates only when all the successors have terminated.
(2) A node with OR successors can terminate when one of the successors has terminated.

disc. goals	node to select or expand
slct-alt	select one of successors of the focus.
ans-spec	expand the focus.
chng-spec	expand a node at a higher level or select and expand one of successors of the focus.
rqst-cand	move the focus to a node at a higher level.
rqst-knwldg	expand the focus.

Fig. 6 The Relation between discourse goals and topics in the top-down prediction.

3.3 Translation of discourse hypotheses into linguistic hypotheses

A topic can be express by several words. For example, a topic 'garden' is expressed by two Japanese words 'niwa' (Japanese origin) and 'teien' (Chinese origin). So we have a table by which real words can be looked up from conceptual topic words. By using this table top-down hypotheses on the topic are translated into words likely to appear in utterances.

As mentioned in section 3.1, we have the lexicon describing the relation between a verb and discourse goals which the verb can express. This lexicon is also used to translate top-down hypotheses on the discourse goal into linguistic ones. First a set of verbs capable of expressing a hypothesized discourse goal is found by consulting this lexicon, and then semantic categories of those nouns which can occur together with these verbs are obtained by looking up case frames of these verbs.

4. Effect of the discourse analysis

The speech dialogue system reported here has not been complete. So we simulated the dialogue system using typed sentences in order to measure an effect of the discourse analysis described in the previous sections. The dialogue controller accepted a dialogue consisting of typed sentences and generated top-down discourse hypotheses every time a sentence was input, and then the linguistic processor of the speech interface analyzed the sentence following the one just input in the dialogue, using the linguistic constraints translated from the discourse hypotheses and predicted words possible to follow each word of the analyzed sentence.

The average number of the predicted words, a kind of branching factor, was computed as a measure of the effect of the discourse analysis on the speech recognition. Assumed the vocabulary consist of about 600 words of which the nouns are about 360, a conversation composed of 60 sentences was analyzed. The average number of the predicted words was 240. This means that the discourse analysis has reduced the vocabulary size by 60 %.

5. Conclusion

The method for the discourse analysis performed in the speech dialogue system we are developing has been reported. The contextual information is analyzed in terms of topics and dialogue structures.

The discourse analysis involves the bottom-up and top-down analyses. The bottom-up analysis proposes candidates for topics and discourse goal based on the semantic representation of an utterance being analyzed. The top-down analysis makes hypotheses on topics and discourse goals referring to the contextual information. The bottom-up hypotheses are matched against the top-down hypotheses. The best match gives the interpretation of an utterance.

The top-down hypotheses are translated into the hypotheses at the linguistic levels, which are given to the speech recognition system. A simulation of the dialogue system using typed sentences has proved that these hypotheses can reduce a search space in recognizing utterances.

REFERENCES

[1] Young, S.J. and Proctor, C.E., Computer Speech & Language, 3 (1989) pp. 329-353.
[2] Young, S.R. et al., Com. ACM, 32 (1989) pp. 183-194.
[3] Niimi, Y. and Kobayashi, Y., Preprints of the Second Symposium on Advanced Man-Machine Interface Through Spoken Language,(1988) pp. 33.1-33.8.
[4] Grosz, J.B. and Sidner, C., Computational Linguistics, 12 (1986) pp. 175-204.
[5] Grosz, J.B., Discourse knowledge, in: Walker D.E. (eds.) Understanding Spoken Language (North-Holland, New York, 1978) pp. 229-337.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

USE OF PROCEDURAL NETWORKS FOR TASK ORIENTED DIALOGUE MODELLING IN MOBILE ROBOT-OPERATOR VOICE COMMUNICATION.

Bianca ANGELINI*, Giuliano ANTONIOL*, Massimo DAL ZOTTO*, Renato DE MORI**, Diego GIULIANI*, Roberto GRETTER*, Gianni LAZZARI*

*I.R.S.T. Istituto per la Ricerca Scientifica e Tecnologica - 38050 Pantè di Povo - Trento - Italy - tel. -39 461 810105
**Mc Gill University - School of Computer Science - 3480 University Street - St. W. Montreal - Quebec - Canada

This paper describes the first version of a system that aims to become both a demo prototype of a robot-operator dialogue component and a starting version of a probabilistic language model acquisition environment. A simulation has been carried out, in which a mobile robot, guided or teleguided by an operator, is able to orientate itself all along the corridors of our Institute and to carry out simple navigation tasks. The most important functions of the dialogue, in this simulation, are the definition of goals and the possibility of avoiding obstacles or passing through critical points.
The spoken dialogue system is composed of three modules: an automatic speech recognizer, a text-to-speech synthesizer and a component implementing the spoken dialogue model.
Two components constituting this system will be described and discussed, that is the recognizer and the dialogue model. Preliminary recognition performance of the system is also presented.

1. INTRODUCTION

The state of the art of automatic speech recognition and voice synthesis, together with the progress in the development of Natural Language understanding and Artificial Intelligence (AI), allow the construction of prototype systems for voice communication between a user and an "intelligent machine" [1]. Most of these systems are oriented to data base inquiry operations in fields where knowledge is quite well defined [2] [3] [4].

In the industrial robotics field voice recognition systems, even with limited vocabulary, are especially useful in "hands-busy, eyes-busy" situations where operators can speak to computers without having to pause and write or keypunch. These systems improve data input accuracy and free operator hands at the same time; the languages adopted in these cases are usually very simple [5].

Applications in which spoken dialogue is used to guide or tele-guide a mobile robot have not yet been fully investigated. An important aspect of the IRST AI project is to provide intelligence for a robot that operates in the IRST environment and interacts with a human being. For example, it moves on the basis of some objective, changes the state of the environment, interacts with the user supplying information, influencing and collaborating with him in the decision process.

In general, robot behaviour can be characterized by three functions: perception, reasoning and activity. Although all these functions may take place in parallel, some of them are constrained to be sequential. For instance, mission planning has to follow goal understanding, which may involve a dialogue with the robot. When robot behaviour is characterized by sequences of actions, these sequences can be well described by Augmented Transition Graphs (ATGs) or Procedural Networks (PNs). In other cases when, for example, the behaviour involves complex interactions of actions, other computational schemes may be useful. These schemes may invoke, inherit or compose methods which may use PNs. Procedural Networks formalism [6] has been chosen in order to model mobile robot-operator dialogue in the first implementation of the system.

A first key point in our project concerns modelling both the dialogue between an operator and a mobile robot and the activities of the robot. In particular, it is important to consider the way in which the robot behaviour in the current situation and in the past, together with the knowledge of the physical environment, can be used to improve the performance of a probabilistic language model for speech recognition by the robot. For instance, when the robot sensor system observes a physical obstacle blocking its path, the associated words and concepts should be assigned higher probabilities. This relates to earlier work indicating that current probabilistic language models can be improved by incorporating a component that tracks recently mentioned words [7].

A second important issue is related to the construction of a probabilistic language model. Instead of relying on language models for speech recognition that are determined a priori, we would like to estimate probabilities for words and phrases which respond as dynamically as possible to the recent history of the robot and its physical situation [8].

2. THE SYSTEM

2.1. Overview

This section presents the whole dialogue system. Figure 1 shows the main components, i.e.:
- world knowledge, divided into static (topology of IRST, people working in it, objects and relations between them) and dynamic (actual position of the robot, presence of obstacles);
- language model, represented by a set of regular grammars;
- modules activated by the PN: speech recognizer, semantic interpreter, speech synthesizer [9], planner, effector.

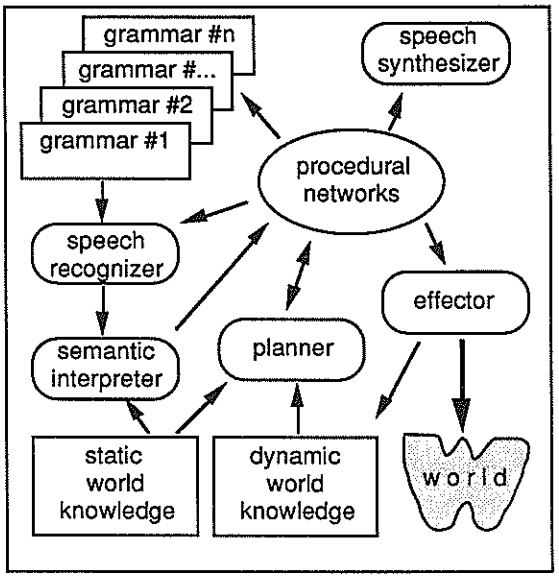


Fig. 1: overview of the whole system.

When a session begins, the PNs drive the dialogue between the user and the computer, sending the synthesizer the request for some goal to reach. After that, a grammar is selected that allows goal definition and the speech recognizer is activated on the incoming utterance. The best scored phrase is then passed to the semantic interpreter, which checks if the goal is well-defined and returns to the PNs either a request for more information or the identified goal. In the first case the message is synthesized, the appropriate grammar is selected and the recognizer is activated; otherwise the PNs call the planner in order to find a path which leads to the target place; finally, if such path exists, the effector must both perform the action and update the dynamic world knowledge with the new robot position.

2.2. The speech recognizer

2.2.1. Signal processing

The input signal is sampled at 16 kHz rate, pre-emphasized by a first order digital network (with a transfer function $H(z) = 1 - 0.95\ z^{-1}$) and then blocked into 20 ms frames shifted every 10 ms. Hamming window is used on each frame. 8 Mel-scaled cepstral coefficients [10], 8 Mel-scaled cepstral derivatives (i. e. delta Mel) [11] and energy together with energy derivative are computed. These three sets of features are vector quantized in three different codebooks having 128, 128, 32 prototype vectors respectively. Using multiple codebooks results in multiple symbols per frame. A variant of the Linde-Buzo-Gray algorithm [12] with Euclidean distance is used for the codebook projects.

2.2.2. Hidden Markov Models (HMMs)

The recognizer is based on HMM technology. To allow the production of multiple output symbols for each time frame, multidimensional discrete symbol distributions HMMs are used. One way to combine the multiple output observations is to assume that they are independent: under this assumption the probability of emitting multiple symbols is simply given by the product of the probabilities of producing each symbol [13].

Each word is modeled with a left-to-right HMM that does not allow state skipping transitions; a different number of states is used according to word lengths. The models are trained by 4 iterations of the Baum-Welch [14] algorithm.

2.2.3. Search strategy

The speech recognizer works with medium size vocabulary (about 120 words) and isolated word pronunciation. The user is requested to make a short pause between words (at least 0.1 second): by so doing the acoustic preprocessor, which has to divide the incoming signal into silence and speech segments looking at its energy and energy ratio, cannot distinguish pauses from plosive sounds. The search for the optimum path will take into account this fact simply by allowing each word to span either one or more segments, according to the number of plosive sounds contained in it. During the search the recognizer uses a regular grammar in a predictive way, i.e. given a partial sentence hypothesis that does not span the whole phrase, it tries to recognize only the words allowed by the grammar; every word of this subset is scored and the resulting new sentence hypothesis is pushed onto a stack structure, in accordance with a global score; in this way the right hypothesis has a good probability of being analyzed first, despite local errors. A pruning strategy is also performed in order to limit the search space.

2.3. Language model

The language model is composed of a set of Finite State Automata (FSAs), each one selected by the PNs in a particular situation of the dialogue. For example, an automata allows the definition of a goal, another permits a better specification of an object when a previously given description does not identify it uniquely, another one takes into account the possibility of stopping the robot when it is moving, and so on.

During the start-up phase, all FSAs are

loaded into memory and a global dictionary is built up containing the words appearing in the various FSAs. Then, for every word, the corresponding Markov model is loaded, that will be used during recognition.

The language generated by the FSAs is not trivial; for example, the definition of a goal may follow this scheme (translated into English):
Robot: I am waiting for a command.
User: go near to the printer.
Robot: which printer ? *(there are several printers)*
User: the one in the secretary office near John's office.
Robot: Ok; I am going near to the printer *printername*.

Of course, within a grammar, language flexibility is very limited: these sentences are considered typical interactions for the given phrases, and therefore understood even if they appear complex.

The lexicon used in our simulation may be divided in various classes:
- verbs (go, come back, stop, ...)
- functional words (articles, prepositions, ...)
- person names (proper nouns, director, secretary, ...)
- groups - areas (speech recognition, vision, ...)
- objects (printers, coffee machine, ...)
- rooms (offices, library, ...)

Our IRST environment description allows the use of ownership relations (... John's office ...), inclusion relations (... printer *in* the secretary office ...), and proximity relations (... *near* the John's office ...). Every object defined is uniquely identified by its name, but the user may refer to it by using the above defined relations, together with a class hierarchy (for instance, the object *printername* belongs to the class *printer*).

2.4. Dialogue model

The PNs formalism is similar to that used for an Augmented Transition Network Grammar. It is made up of states and arcs connecting the states, which stand for the robot behaviour. Three entities can be associated with an arc: a condition, an operation and an action. The condition must be verified for the transition associated with the arc to be carried out. An operation will be of the form PUSH, POP or JUMP. In our model for a mobile robot, an action involves the robot memory, sensors or effectors: i.e. the robot stores a piece of information, carries out speech recognition or speech synthesis, moves from place to place. Let us consider now a PN (Figure 2) that can be used for communicating to the robot the mission it has to perform.

Let Sq be a robot state describing its situation of waiting. If a message is received that calls its attention, it reaches the state S1 and sends a request for a mission to the operator. The robot is now in the state S2 and starts a dialogue Dm, whose task is that of defining the mission, for example navigation. Dm is another PN and, when finished, has to return the control to S3 with a message frame containing the type of task and some attributes, in this case the coordinates of the goal. If the message frame cannot be completely fulfilled, the robot reaches the state S4 and can either come back to S2 with the action *new request* or quit to state Sq sending a message report. If the frame message is fulfilled, the control returns to the calling PN, where the planner will be activated. When a sentence is pronounced a semantic interpreter of Dm generates frame instantiations and control slot filling. Each possible destination (goal) is represented by a frame prototype. After matching an input sentence with the associated language model, a termination condition decides when the content of the slot is sufficient for filling the frame.

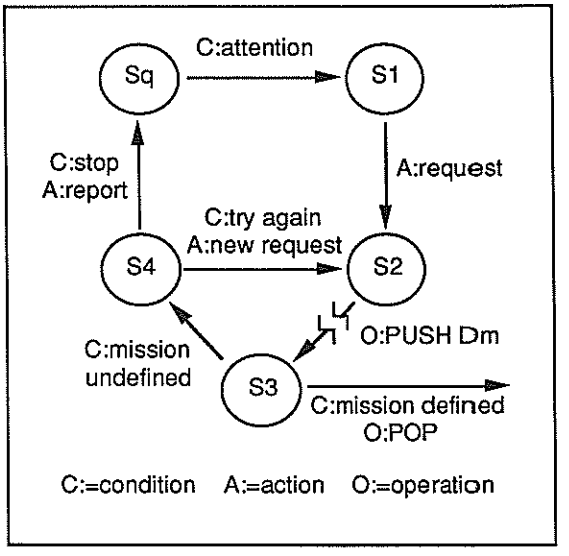


Fig. 2: a procedural network for mission communication.

3. SIMULATION OF THE ROBOT ACTIVITY

In order to evaluate and test the system, a simulation of the robot activity (i.e. navigation) in the IRST environment (i.e. topology, organization, personnel and objects) has been implemented. A graphic picture of the 2D IRST map is printed on the screen when the system starts up. After that the robot is ready to execute navigation tasks from point to point. The destination points are defined in term of persons, objects and locations that a user can find in the IRST environment, defined in a static knowledge representation. Personnel, organization and objects are related to a default topology. For instance, *a given researcher works in an office that is located in the area Speech Recognition, a laser-printer is located in front of the office of personnel secretary*, and so on. The operator can define goals either by specifying directly destinations that correspond to given coordinates in the default topology, or by using the above mentioned relations of ownership, inclusion and proximity.

If the semantic interpreter does not fill a frame because of ambiguity, i.e. there are more than one candidate entity matching the sentence, the dialogue manager generates a question asking for more details, and this process is repeated until the goal is defined. The planner computes the best path

according to the topology of the building and to the status of path (stable or momentary obstacles). Once the path is computed, the robot starts moving on the screen, simulating the real path. The robot now can be stopped either by the operator or by the sensor subsystem. In this case the dialogue manager generates a question asking "What shall I do?". The operator may decide either to quit the task and define a new goal, or change the path because of an obstacle, or stand up for a certain period of time and then go on with the task completion.
The programs implementing the recognizer and the Procedural Networks are written in C and Lisp language and the software runs on a workstation Sun4/330.

4. RESULTS AND CONCLUSIONS

In order to test the recognizer, we collected at our Institute a speech data base of 25 speakers, each one pronouncing from 2 to 5 repetitions of a 200-word vocabulary and 5 sentences.

With this data base we first trained our HMM word models. Then two experiments were performed on 61 speech sentences: the first one using the FSAs described in section 2.2.3; the second one without any language model, i.e. with perplexity equal to the number of words used by the various FSAs. Results are shown in table 1, where *phrase errors* stands for the number of phrases with one or more word errors, and *semantic errors* takes into account only the phrases whose meaning has been changed because of some word error.

	with FSAs		without FSAs	
	number	perc.	number	perc.
phrase number	61		61	
phrase errors	10	16.4	21	34.4
semantic errors	1	1.6	15	25.0
word number	329		329	
word substitutions	10	3.04	27	8.20
word insertions	0		1	0.30
word deletions	0		0	

Table 1: recognition errors

It is worth noting that most of the word substitutions involve functional words like prepositions (i.e. *del* (*of the*, male singular) is often confused with *della* (*of the*, female singular)) and do not alter the phrase meaning.

In conclusion, we have described the first version of our spoken dialogue system. We think this is a starting point for setting up a system for modelling both the dialogue between an operator and a mobile robot and the activities of the robot. The future work will focus on building a probabilistic language model on which the probabilities are not determined a priori but estimated dynamically. We believe that, as the use of voice-activated devices spreads to non expert users, the kind of problems we dealt with and the type of solutions we have described will become very important.

ACKNOWLEDGEMENTS

The authors would like to thank O.Stock for fruitful discussions and careful reading of the manuscript and also E.Merlo and Y.Normandin for their contribution in the development of our HMM package.

REFERENCES

[1] Stringa L., 'An Artificial Intelligence Approach to Speech Recognition and Understanding', Pattern Recognition Letters 8 (1988), 1, 39-45.
[2] Young S.R., Hauptmann A.G., Ward W.H., Smith E.T. and Werner P., 'High level knowledge sources in usable speech recognition systems', Communications of the ACM, February 1989, vol. 32, n. 2.
[3] Levinson S.E. and Rabiner L.R., 'A task-oriented conversational mode speech understanding system', Bibliotheca Phonetica 12 (1985), pagg. 149-196.
[4] Young S.J. and Proctor C.E., 'The design and implementation of dialogue control in voice operated database inquiry systems', Computer Speech and Language 3 , (1989), 329-353
[5] Hackenberg R.G., GE/RCA, 'Intelligent voice control of cameras and robots', Avios Proceedings, Alexandria, Virginia, October 1987, pagg. 325-334.
[6] De Mori R., Merlo E., Palakal M., J Rouat, 'Use of procedural knowledge for automatic speech recognition', Proceedings of IJCAI 87, vol. II, pagg. 840-843.
[7] De Mori R., Kuhn R. 'Cache-based Stochastic Language Modelling', accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence.
[8] De Mori R., Gretter R., Khun R., Lazzari G., Stringa L., 'Modelling Operator - Robot Oral Dialogue for Applications in Telerobotics', accepted at 10th ICPR, Atlantic City, June 1990.
[9] Gretter R., Omologo M., 'The Use of Line Spectrum Representation in Speech Synthesis', Alta Frequenza 58, (1989), 3, pagg. 293-300.
[10] Davis S. B., Mermelstein P., 'Comparison of Parametric Representation of Monosyllabic Word Recognition in Continuously Spoken Sentences' in IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. ASSP-28 No. 4, August, 1980
[11] Furui S., 'Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum' in IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. ASSP-34 No. 1, February, 1986
[12] Linde Y., Buzo A., Gray R. M. 'An Algorithm for Vector Quantizer Design' in IEEE Transactions on Communication, Vol. COM-28 No. 1, January, 1980
[13] Lee K. F. 'Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System' Ph. D. Thesis, Carnegie Mellon University, Pittsburgh, April 1988
[14] Baum L. E. 'An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes.' Inequalities 3:1-8,1972

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ISOLATED–UTTERANCE SPEECH RECOGNITION USING HIDDEN MARKOV MODELS WITH BOUNDED STATE DURATIONS

Hung–yan Gu *, *Chiu–yu Tseng*# *and Lin–shan Lee* *

* Department of CSIE, National Taiwan University,
# Institute of History and Philology, Academic Sinica,
Taipei, Taiwan, Republic of China

In this paper, Hidden Markov Models (HMM's) with Bounded State Durations (HMM/BSD) are proposed to explicitly model the state durations of HMM's and more accurately consider the temporal structures existing in speech signals in a simple, direct but effective way. The state durations of HMM/BSD are simply lower and upper bounded by two bounding parameters for each state in the recognition phase, as compared to the approaches of using Poisson, gamma or other distributions proposed previously to model state durations. These bounding parameters, on the other hand, can be estimated during the training phase.

A series of experiments have been conducted for speaker dependent applications using all the 408 highly confusing first–tone Mandarin syllables as the example vocabulary. It was found that in the discrete case the recognition rate of HMM/BSD (78.5%) is 9.0%, 6.3% and 1.9% higher than the conventional HMM's and HMM's with Poisson and gamma distributed state durations, respectively. In the continuous case (partitioned Gaussian mixture modeling) the recognition rates of HMM/BSD (88.3% with 1 mixture, 88.8% with 3 mixtures and 89.4% with 5 mixtures) are 6.3%, 5.0% and 5.5% higher than that of the conventional HMM's, and 5.9%(with 1 mixture), 3.9%(with 3 mixtures) and 3.1%(with 1 mixture), 1.8%(with 3 mixtures) higher than HMM's with Poisson and gamma distributed state durations, respectively. As to computation complexity and recognition speed, it turns out that the computation complexity required by the new modeling method proposed here is much less than that for HMM's with Poisson or gamma distributed state durations.

1. Introduction

In recent years, the techniques of Dynamic Time Warping (DTW) [1,2], Vector Quantization (VQ) [3,4], and Hidden Markov Models (HMM's) [5,6,7] have been successfully applied to various speech recognition problems. One feature common to all these techniques is that they all have tried to model the temporal structures of speech signals implicitly or explicitly. By temporal structures we mean the relative ordering and time durations of acoustic events occurring in speech signals. In the approaches of HMM's, it has been noted by many researchers that HMM's explicitly modeling the state durations can achieve higher recognition rates than conventional HMM's, because such modeling methods can more accurately describe the temporal structures of speech signals. For example, either nonparametric [5,6] or parametric [8,9] modeling techniques for state durations have been shown to be able to significantly improve the recognition rates. In conventional HMM's the distribution for state durations is geometric because the probability of remaining in state i for duration d is $p_i(d) = (a_{ii})^{d-1} * (1-a_{ii})$, where a_{ii} is the state transition probability from state i to itself and $(1-a_{ii})$ from state i to other states. In nonparametric modeling of state durations, the probability density function (PDF) $p_i(d)$ must be estimated for all possible combinations of i and d, requiring an enormous amount of training utterances. On the other hand, in parametric modeling only one or two parameters have to be estimated for each i and the values of $p_i(d)$ can be evaluated from the corresponding PDF formula, e.g. Poisson or gamma PDF. However, more CPU time will be needed to evaluate the probability of specific state durations, and the assumed PDF forms are not necessarily adequate for all states of all candidates in the vocabulary.

2. Hidden Markov Models with Bounded State Durations

In the new approach of HMM/BSD proposed in this paper, in addition to the state transition and observation production parameters of the conventional HMM's, there are two more parameters for each state. These two parameters are intended to confine the durations of the states within the minimum and maximum allowable limits (time normalized) in the recognition phase. Therefore, they are called the lower and upper bounds or the bounding parameters in this paper. In the training phase, these bounding parameters have to be estimated, and in the recognition phase, these bounding parameters will slightly modify the definition of the maximum likelihood for conventional HMM's. Let

$O = o_1 o_2 \cdots o_T$ be a specific observation sequence,

l_i be the time normalized lower bound of the state duration for state i,

u_i be the time normalized upper bound of the state duration for state i,

$d(i)$ be the duration of state i,

$\hat{s}$ represent a possible state transition sequence obtained by successively concatenating $d(1)$ repetitions of state 1, $d(2)$ repetitions of state 2, ..., and $d(N)$ repetitions of state N, where N is the total number of states, i.e.,

$$\hat{s} = \underbrace{s_1 s_1 \cdots s_1}_{d(1)} \underbrace{s_2 s_2 \cdots s_2}_{d(2)} \cdots \cdots \underbrace{s_N s_N \cdots s_N}_{d(N)}$$

S be the set of all possible state transition sequences based on some given state transition model, for example, the three-transition model and the two-transition model in Fig.

1(a) and (b), respectively, $\mathcal{M}$ represent a HMM/BSD.

(a) three-transition model (b) two-transition model

Fig. 1 State transition models.

Then, in HMM/BSD, the maximum likelihood, $P(O|\mathcal{M})$, for $\mathcal{M}$ to produce the observation sequence O is

$$P(O|\mathcal{M}) = \max \{ P(O \text{ and } \hat{s} \mid \mathcal{M}) \} \quad (1)$$

$$\sum_{i=1}^{N} d(i) = T$$

$$d(i)=0 \text{ or } l_i \le d(i)/T \le u_i$$

$$\hat{s} \in S$$

In equation (1), the value of $d(i)$ can be zero if the model in Fig. 1(a) is used, because in that model some states may be skipped. The definition of equation (1) is adopted only in the recognition phase but not in the training phase, because the bounding parameters of state durations must be learned in the training phase, and it is thus unreasonable to confine the state durations of the training utterances.

To estimate the parameters of HMM/BSD, the Viterbi algorithm is used in the training phase. The parameters a_{ij} and $b_j(o_t)$ are obtained by exactly the same algorithm as used in the conventional HMM's, and the training of the bounding parameters l_i and u_i is based on the maximum likelihood state transition sequences from which the parameters for a_{ij} and $b_j(o_t)$ are estimated. For illustration, let d_{ik} be the duration of state i in the maximum likelihood state transition sequence of the k-th training utterance, T_k be the time length of the k-th utterance, and K be the total number of training utterances. Then l_i and u_i are estimated as

$$l_i = \min_{k=1}^{K} \{ \max[d_{ik}/T_k, 1/T_k] \} \quad (2)$$

$$u_i = \max_{k=1}^{K} \{ \max[d_{ik}/T_k, 1/T_k] \} \quad (3)$$

From the above description, it can be found that HMM/BSD is a parametric and explicit modeling method for state durations to model the temporal structures existing in speech signals in a relatively simple way. In HMM/BSD, the PDF's of state durations are geometric but with the lower and upper portions removed. The lowest curve of Fig. 2 is a typical example in logarithmic scale.

As compared to the other parametric modeling methods proposed previously, e.g. HMM's with Poisson [8] (HMM/Poisson) or gamma [9] (HMM/gamma) distributed state durations, due to the parametric nature, the three modeling methods (HMM/BSD, HMM/Poisson and HMM/gamma) can all operate under the condition in which a relatively small number of training utterances is available. However, there is a very significant difference between HMM/BSD and HMM/Poisson or HMM/gamma in the recognition phase, i.e., in HMM/Poisson or HMM/gamma it is very possible for a state to occupy many more

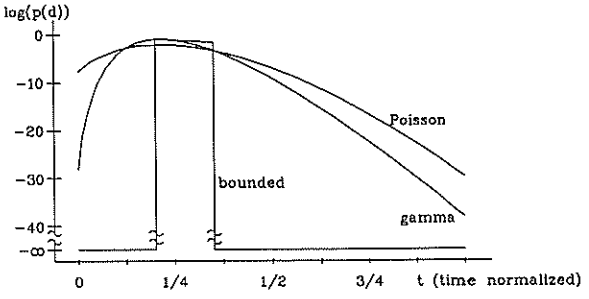


Fig. 2 The typical PDF's of the three state duration models.

or fewer frames than the expected mean at the price of only slightly decreased probability density (in logarithmic scale as shown in Fig. 2), but in HMM/BSD it is always impossible for a state to occupy more or fewer frames than is defined by the state's bounding parameters, as can be observed from the typical state duration PDF's shown in Fig. 2, in which for HMM/BSD the discontinuities exist in the boundaries of allowable state durations but for HMM/Poisson or HMM/gamma the PDF's are continuous and relatively smooth in the full duration of an utterance. HMM/BSD is, therefore, more effective than HMM/gamma or HMM/Poisson in inhibiting a state occupying more or fewer signal frames than is adequate, and can more accurately model the temporal structures in speech signals by better avoiding the problem of erroneous match between a testing utterance and a confusing speech model. As will be shown later in this paper, the above statements will be supported by experimental results.

In implementing the recognition algorithm for HMM's with explicit modeling of state durations, it can be found that some identical computations will have to be performed repeatedly if the conventional recognition algorithm is directly used. This is because the conventional algorithm treats the state duration index d as a dependent variable on the time index t and, therefore, simply performs the searching process on the two-dimensional (time t, state i) space. To reduce those duplicated computations, some restructuring of the dynamic programming principle usually used in DTW or connected speech recognition[2,10,11] is found to be very helpful in this circumstance. The basic idea here is to treat the state duration index d as the third independent variable, such that the dynamic programming searching process can be performed on a three-dimensional (t, i, d) space to avoid those duplicated computations. In this way, not only can the computation load be reduced, but the derived algorithm is found to be more suitable for parallel processing than the conventional one because of its inherent structure. The space complexity of the derived algorithm, on the other hand, can be shown to be identical to the conventional one because the temporal values for those combinations, such as (t-2, i, d), (t, i, d), (t+2, i, d), ..., can share the same memory location. Such an implementation was used in all the following experiments.

3. The Speech Database and Test Conditions for the Experiments

In the experiments, a total of 408 very confusing first-tone Mandarin syllables are taken as the example vocabulary [12]. Speaker dependent applications are considered. For such a vocabulary and speaker dependent requirements it is not very practical to use too many training utterances because the training speaker may become tired of producing a large number of training utterances. Thus, the speech data collected here included only 5 sets of utterances produced by a single male speaker. Each set of utterances consisted of one reading of all the 408 Mandarin

syllables. The speech data were sampled set by set at a sampling rate of 8 KHz into a VAX-11/730 computer via a DSC-200 audio data conversion system with an anti-aliasing filter in a sound proof room.

A program was designed to detect the end points of each utterance [13] after being pre-emphasized with the digital filter, 1-0.95/z [14]. The length of the utterances ranges from 270ms to 570ms with the average being on the order of 400ms. The speech samples between the end points were then Hamming windowed [14] into signal frames with a length of 20ms and overlap of 10ms to adjacent frames. For each signal frame, 16 LPC cepstrum coefficients were evaluated using the auto-correlation coefficients [15]. We then applied bandpass liftering [16] to the cepstrum coefficients evaluated for each frame and stored the resulting feature vectors into a file.

Before the speech database could be used for discrete HMM's, the feature vector of each signal frame had to be vector quantized [17]. A quantization method of full search through the codebook was used. The codebook here consists of 124 code words and was trained from the first two sets of utterances. After the parameters of HMM/BSD or HMM's were estimated, the parameters of observation production were then smoothed [18] because the training utterances were far from sufficient to accurately estimate those parameters.

In this paper, a series of experiments for various aspects of HMM/BSD had been conducted. The results obtained from each experiment are, in fact, the averages of 5 small experiments. Each small experiment uses one set of utterances as testing data and the other 4 as training. The number of states was set to be 6 for all the experiments. Because the training utterances are not sufficient to accurately estimate the covariances of Gaussian mixtures of certain states in the continuous case, the covariances of all Gaussian mixtures across different states and models will be assumed to be equal and set to be the identity matrices. Then, in partitioned Gaussian mixture modeling [6,7], the probability density to produce a feature vector x can be evaluated as

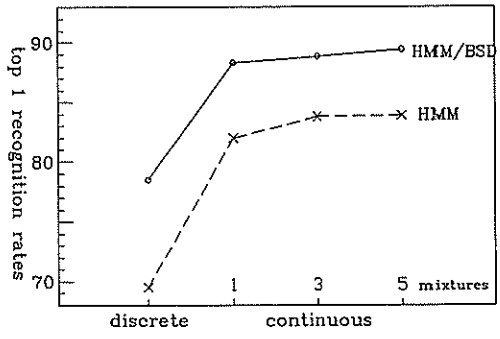
$$b_j(x) = \max_{m=1}^{M} b_{jm}(x)$$

$$b_{jm}(x) = C_1 * e^{C_2*[(x_1-u_1)^2 + \ldots + (x_n-u_n)^2]} \qquad (4)$$

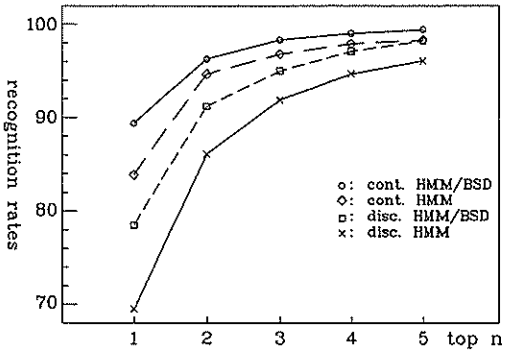
where C_1 and C_2 are constants (in this paper C_2 is set to be 1) and u_i is the estimated mean of the i-th component of x for the m-th mixture in a specific state. It is noted that the expression in the square bracket of equation (4) is the same as the distance measure used for vector quantization in the discrete case.

4. Experimental Results(I) - HMM/BSD Vs Conventional HMM's

From preliminary experiments conducted, it was found that for both cases (HMM/BSD and conventional HMM's) the two-transition model of Fig. 1(b) always has slightly higher recognition rates than the three-transition model of Fig. 1(a). Therefore, only the two-transition model will be adopted in the following experiments. To compare the performance of HMM/BSD with that of conventional HMM's, a series of experiments were conducted in both discrete and continuous (with various numbers of mixtures) cases. The resulting recognition rates were then used to draw Fig. 3, where Fig. 3(a) is



(a) Top 1 recognition rates vs. discrete and continuous mixtures



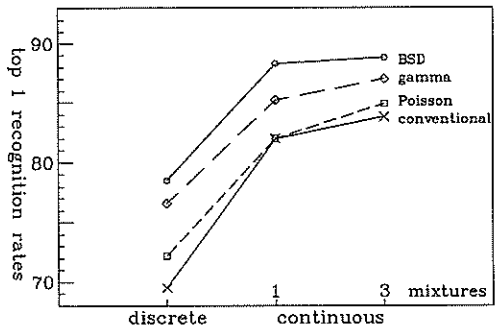
(b) Top n recognition rates vs. n

Fig. 3 Recognition rates of HMM/BSD vs. conventional HMM's

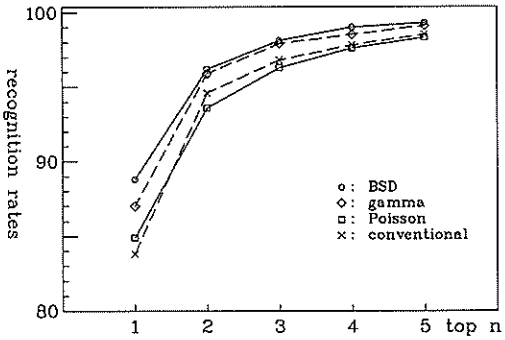
for top 1 recognition rates vs. both discrete and continuous (with different numbers of mixtures) cases, and Fig. 3(b) is for top n recognition rates (5 mixtures for the continuous case) vs. n. Here, the top n recognition rate is the rate that the correct syllable for a testing utterance is among the top n candidates. From Fig. 3, it can be seen that for top n rates HMM/BSD always performs better for any n, n=1,2,3,4,5. As for top 1 rates, HMM/BSD are 9.0%, 6.3%, 5.0% and 5.5% higher than the conventional HMM's for the discrete and continuous (with 1, 3 and 5 mixtures) cases, respectively. Also, it can be seen that the rates for 1, 3 and 5 mixtures are very close to each other in both HMM/BSD and HMM's cases; therefore, the recognition rate is not very sensitive to the number of Gaussian mixtures used.

5. Experimental Results(II) - HMM/BSD Vs HMM/Poisson and HMM/gamma

Since HMM/BSD proposed here is shown to provide improvement as compared to the conventional HMM's, we should then compare HMM/BSD with other parametric modeling methods previously proposed to improve the performance of HMM by explicitly modeling the state durations, e.g., HMM's with Poisson or gamma distributed state durations. A series of experiments for HMM/Poisson and HMM/gamma were then performed. The training algorithm used was the Viterbi instead of the Baum-Welch approach in order to make a correspondence with that used for HMM/BSD, and the estimated values for means and variances of state durations are based on the maximum likelihood state transition sequences and are time normalized. The resulting recognition rates of HMM/Poisson and HMM/gamma together with the rates for conventional HMM's are used to draw Fig. 4



(a) Top 1 recognition rate vs. discrete and continuous mixtures



(b) recognition rates (3 mixtures) vs. top n candidates

Fig. 4 Recognition rates of HMM/BSD vs. conventional HMM's, HMM/Poisson and HMM/gamma.

(Fig. 4(a) for discrete and continuous cases, Fig. 4(b) for top n rates). From Fig. 4, it can be found that the recognition rates of HMM/Poisson are always better than conventional HMM's, those of HMM/gamma are always higher than HMM/Poisson, and the recognition rates of HMM/BSD are always significantly higher than both HMM/Poisson and HMM/gamma in both discrete and continuous cases. One explanation for the above phenomenon is that HMM/BSD is more effective than the other two modeling methods in inhibiting one state occupying more or fewer signal frames than are adequate. This, thus, results in better modeling capability for temporal structures existing in speech signals, as was discussed in section 2. In addition to better performance in recognition rates, HMM/BSD is also preferred from the standpoint of computation time requirement, because the state duration PDF for HMM/BSD is much simpler and easier to compute than that of HMM/Poisson or HMM/gamma. In practice, it was found that the CPU time requirement for HMM/Poisson and HMM/gamma are, on average, 1.3 and 2.4 times that for HMM/BSD in the discrete case, and are 1.1 and 1.7 times that in the continuous (with 1 mixture) case, when those computations which can be performed in the training phase are replaced by table lookup in the recognition phase.

6. Conclusion

In this paper, slight modification is proposed for HMM's, i.e., the lower and upper bounds for state durations, to more accurately model the temporal structures existing in speech signals. The experiments show that this proposed modification has higher modeling capability and provides better recognition rates than the conventional HMM's and the other two parametric modeling methods, i.e., HMM's with Poisson or gamma distributed state durations. Also, due to its simple form for state duration PDF's, this modification requires much less computation time than HMM's with Poisson or gamma distributed state durations.

References

[1] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. Acoust., Speech, Signal Processing, Vol. 23(1), pp. 67-72, Feb. 1975.

[2] L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition - Theory and Selected Applications", IEEE Trans. Communications, Vol. 29(5), pp. 621-659, May 1981.

[3] D. K. Burton, J. E. Shore and J. T. Buck, "Isolated-word Speech Recognition Using Multisection Vector Quantization Codebooks", IEEE trans. ASSP, Vol. 33(4), pp. 837-849, 1985.

[4] G. E. Kopec and M. A. Bush, "Network-Based Isolated Digit Recognition Using Vector Quantization", IEEE trans. ASSP, Vol. 33(4), pp. 850-867, 1985.

[5] L. R. Rabiner, B. H. Juang, S. E. Levinson and M. M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities", AT&T Tech. J., Vol. 64(6), pp. 1211-34, July-Aug. 1985.

[6] B. H. Juang and L. R. Rabiner,"Mixture Autoregressive Hidden Markov Models for Speech Signals", IEEE trans. ASSP, Vol. 33(6), pp. 1404-13, 1985.

[7] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. IEEE, Vol. 77(2), pp. 257-286, 1989.

[8] M. J. Russell and R. K. Moore, "Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition", ICASSP, pp. 5-8, 1985.

[9] S. E. Levinson,"Continuously Variable Duration Hidden Markov Models for Speech Analysis", ICASSP, pp. 1241-1244, 1986.

[10] S. Nakagawa, "A connected Spoken Word Recognition Method by O(n) Dynamic Programming Pattern Matching Algorithm", ICASSP, pp. 296-299, 1983.

[11] H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition", IEEE trans. ASSP, Vol. 32(2), pp. 263-271, 1984.

[12] Lin-Shan Lee and Chiu-yu Tseng, "Mandarin Speech Input/Output Techniques for Chinese Computers - The State of the Art", Proc. of the National Science Council (Physical Science and Engineering) (Taipei, Taiwan), Vol. 11, No. 4, pp. 273-290, July 1987.

[13] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the End Points of Isolated Utterances", Bell Syst. Tech. J., Vol. 54(2), p. 297-315, Feb. 1975

[14] J. D. Markel and A. H. Gray, Jr., Linear Prediction of Speech, New York, NY: Springer-Verlag, 1976.

[15] A. H. Gray, Jr. and J. D. Markel, "Distance Measures for Speech Processing", IEEE Trans. ASSP, Vol. 24, pp. 380-391, Oct. 1976.

[16] B. H. Juang, L. R. Rabiner and J. G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition", IEEE Trans. ASSP, Vol. 35, pp. 947-953, Jul. 1987.

[17] R. M. Gray, "Vector Quantization", IEEE ASSP Magazine, Vol. 1, pp. 4-29, Apr. 1984.

[18] Kazuhide Sugawara, etc., "Isolated Word Recognition Using Hidden Markov Models", ICASSP, pp. 1-4, 1985.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

INCREASING THE DIFFERENCE BETWEEN THE SIGNIFICANT AND THE NON-SIGNIFICANT SINGULAR VALUES IN A MODEL OF LPC EXCITATION BASED ON THE SVD

Victoria E. Sanchez-Calle, Juan M. Lopez-Soler,
Jose C. Segura-Luna, Antonio M. Peinado-Herreros,
Antonio J. Rubio-Ayuso

Dept. de Electronica y Tecnologia de Computadores
Univ. de Granada. 18071 Granada, Spain.

This paper presents an attempt to reduce the amount of information we have to send in a model of LPC excitation based on the SVD by applying the SVD to the matrix H^tH instead of to the matrix H. For the matrix H^tH the difference between the significant singular values and the non-significant singular values is higher and we study if we can reduce, due to this fact, the information concerning the positions of the elements that we consider different from zero in both schemes. We also make an study of the performance of both schemes in different cases and we finally conclude that we can find practically no difference between both schemes save for a slightly better SNRseg in our scheme and a slightly better performance of the other scheme for high frequencies.

1. INTRODUCTION

One of the main problems we have when it comes to lowering the bit rate, maintaining a high quality synthesis, is to find a representation for the excitation of the LPC filter that, with few kbits/s, gives us a high voice quality. Up to now the representation that requires a lower bit rate is the traditional one based on the extraction of the pitch and the decision voiced/unvoiced/silence, but this model doesn't produce natural-sounding speech, even at higher bit rates.
In the last years two successful models have been developed for the excitation. These are the multipulse excitation model [1] and a stochastic excitation model (CELP) [2], which gets a bit rate of 3kb/s for the excitation but at a large computational cost.
Recently a new representation has been proposed [3], based on the Singular Value Decomposition (SVD). This is applied to the matrix $\mathbf{H}$, where $\mathbf{H}$ represents the impulse response of the LPC filter.
In this paper we also use the Singular Value Decomposition to represent the excitation, but we apply the SVD to the matrix $\mathbf{H}^t\mathbf{H}$. In this case the difference between the significant singular values and the non-significant singular is higher, and we study whether we can get any advantage from this fact. Besides we make an study of the performance of both schemes in different cases.

2. REPRESENTATION OF LPC FILTER

Let $h(n)$ be the impulse response of the LPC filter and we consider a frame of N samples over which the LPC filter parameters are assumed constant. Then we can express $s(n)$ as,

$$s(n)=\sum_{k=0}^{n}h(n-k)x(k)+s_0(n) \quad n=0,1,...,N-1 \tag{1}$$

where $x(n)$ is the input to the LPC filter, and $s_0(n)$ is the contribution to the present frame of the excitation in the previous frames.
Let's now consider the error,

$$e(n)=s(n)-\left[\sum_{k=0}^{n}h(n-k)x(k)+s_0(n)\right] \tag{2}$$

from this equation we can obtain the excitation $x(n)$ that minimizes the squared error E. Previously we are going to weight E with a weighting filter of the form

$$W(z)=\frac{1+P(z)}{1+P(z/\gamma)} \tag{3}$$

where

$$P(z)=\sum_{k=1}^{p}a_k z^{-k} \tag{4}$$

By filtering this way we can tolerate larger errors in the formant regions than in the in-between formant regions. The value of γ is between 0 and 1 and controls in what degree we de-emphasize the error. So the final expression for the weighted squared error will be

$$E_w=\sum_{n=0}^{N-1}\left[s_w(n)-s_{w0}(n)-\sum_{k=0}^{n}h_w(n-k)x(k)\right]^2 \tag{5}$$

and minimizing with respect to the excitation amplitudes we have in matrix form

$$\mathbf{c}=\mathbf{H}_w^t\mathbf{H}_w\mathbf{x} \tag{6}$$

where $\mathbf{c}=\mathbf{H}_w^t(\mathbf{s}_w-\mathbf{s}_{w0})$ is the cross-correlation between the

impulse response $h_w(n)$ and the signal $s_w(n)$-$s_{w0}(n)$.
In the same way weighting expression (1) and expressing it in matrix form,

$$\mathbf{s}_w-\mathbf{s}_{w0}=\mathbf{H}_w\mathbf{x} \tag{7}$$

where $\mathbf{H}_w$ is a matrix of the form,

$$\mathbf{H}_w=\begin{bmatrix} h_w(0) & 0 & . . & 0 \\ h_w(1) & h_w(0) & . . & 0 \\ . & . & . . & . \\ . & . & . . & . \\ h_w(N-1) & h_w(N-2) & . . & h_w(0) \end{bmatrix} \tag{8}$$

We can see that $\mathbf{H}_w$ is a lower triangular Toeplitz matrix, and as $h_w(0)=1$, then the determinant of $\mathbf{H}_w$ is $(h_w(0))^N=1$ and we can conclude that $\mathbf{H}_w$ is a matrix of rank N.
The elements of the $\mathbf{H}_w^t\mathbf{H}_w$ are

$$\left[\mathbf{H}_w^t\mathbf{H}_w\right]_{i,j}=\sum_{n=0}^{N-1}h_w(n-i)h_w(n-j) \tag{9}$$

as we can see that $\mathbf{H}_w^t\mathbf{H}_w$ is a symmetric matrix.

3. SINGULAR VALUE DECOMPOSITION

Lets now apply the Singular Value Decomposition (SVD) [4]. The singular value decomposition theorem states that if we have an arbitrary $m\times n$ complex-valued matrix $\mathbf{A}$ of rank k, there exist positive real numbers $\sigma_1\geq\sigma_2\geq..\geq\sigma_k>0$ (the so called singular values of A), an $m\times m$ unitary matrix $\mathbf{U}=(\mathbf{u}_1,\cdots,\mathbf{u}_m)$, and an $n\times n$ unitary matrix $\mathbf{V}=(\mathbf{v}_1,\cdots,\mathbf{v}_n)$ such that $\mathbf{A}$ can be expressed as

$$\mathbf{A}=\mathbf{U}\Sigma\mathbf{V}^h \tag{10}$$

where the $m\times n$ matrix Σ has the structure

$$\Sigma=\begin{bmatrix}\mathbf{D} & 0\\ 0 & 0\end{bmatrix} \tag{11}$$

and $\mathbf{D}$ is a $k\times k$ diagonal matrix. The elements of the diagonal are called the singular values of $\mathbf{A}$. $\mathbf{V}^h$ is the hermitian transpose of the matrix $\mathbf{V}$, as $\mathbf{U}$ and $\mathbf{V}$ will be in general complex matrices. In the case of $\mathbf{A}$ being a real matrix , $\mathbf{U}$ and $\mathbf{V}$ would be real as well and $\mathbf{V}^h$ would become $\mathbf{V}^t$ where t means the transpose of $\mathbf{V}$.
Lets apply the SVD to the matrix $\mathbf{H}_w^t\mathbf{H}_w$ of rank N, in doing so we have

$$\mathbf{H}_w^t\mathbf{H}_w=\mathbf{V}\mathbf{D}\mathbf{V}^t \tag{12}$$

This matrix is a special case as is a real symmetric matrix , in this case the singular values resulting from the SVD are the eigenvalues of $\mathbf{H}_w^t\mathbf{H}_w$, and the column vectors $\mathbf{v}_1,\cdots,\mathbf{v}_N$ are the eigenvectors of $\mathbf{H}_w^t\mathbf{H}_w$.

If we apply (12) to (6) then

$$\mathbf{c}=\mathbf{V}\mathbf{D}\mathbf{V}^t\mathbf{x} \tag{13}$$

multiplying both members of this equation by $\mathbf{V}^t$

$$\mathbf{V}^t\mathbf{c}=\mathbf{D}\mathbf{V}^t\mathbf{x} \tag{14}$$

as we can see we are applying the unitary transformation $\mathbf{V}^t$ to the cross-correlation $\mathbf{c}$.
In the same way applying the SVD to $\mathbf{H}_w$, as is done in [3], instead of applying it to $\mathbf{H}_w^t\mathbf{H}_w$, we get

$$\mathbf{H}_w=\mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^t \tag{15}$$

where the singular values of $\mathbf{H}_w$ are the square root of the singular values of $\mathbf{H}_w^t\mathbf{H}_w$.
Substituting (15) in (7) we have

$$\mathbf{s}_w-\mathbf{s}_{w0}=\mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^t\mathbf{x} \tag{16}$$

and

$$\mathbf{U}^t(\mathbf{s}_w-\mathbf{s}_{w0})=\mathbf{D}^{1/2}\mathbf{V}^t\mathbf{x} \tag{17}$$

as in [3], in this case we are applying the unitary transformation $\mathbf{U}^t$ to the weighted signal $\mathbf{s}_w-\mathbf{s}_{w0}$.
What we are going to do in both cases (14) and (17) is to consider equal to zero the smallest elements of $\mathbf{V}^t\mathbf{c}$ and $\mathbf{U}^t(\mathbf{s}_w-\mathbf{s}_{w0})$ respectively. We will see that the biggest values of (14) and (17) normally ,but not always, correspond to the biggest singular values, being this true more often for (14) than for (17) as the difference between the significant singular values and the non significant singular values is higher for the matrix $\mathbf{H}_w^t\mathbf{H}_w$ than for the matrix $\mathbf{H}_w$. To synthesize we will simply calculate $\mathbf{x}$ from (14) and (17) and substitute in (1).
We will study next how many elements we can consider 0 in both schemes, what is the statistical distribution of the positions of the elements that we finally consider different from 0 and if we can get any advantage from the fact that in (14) these positions correspond more often to the biggest singular values of $\mathbf{H}_w^t\mathbf{H}_w$.

4. RESULTS

4.1. Analysis and simulation conditions

For performance evaluation we used an about three second Spanish sentence pronounced by a female. The sample was low-pass filtered at 4 kHz cut-off frequency and digitized by a 12 bit A/D converter at 8 kHz sampling.
The length of the analysis frame was considered of 15 msec, 120 samples, and it was divided into 3 sub-blocks of 40 samples each. Synthesis filter order was established to be 10 and the value of $\gamma=0.8$. We used the Segmental SNR for objective evaluation.

4.2. Experiment design

We studied the performance of both methods, that is (14) and (17), in three different cases:

A- We consider different from zero 10 values of (14) and (17) in each sub-block.

B- We consider different from zero 15 values of (14) and (17) in each sub-block.

C- We consider different from zero 20 values of (14) and (17) in each sub-block.

In all the cases without quantization. In each of these cases we consider another three possibilities:

A1- Different from zero the biggest 10 elements .

A2- Different from zero the biggest 10 elements among the 20 elements corresponding to the 20 biggest singular values.

A3- Different from zero the 10 elements corresponding to the 10 biggest singular values, that is, the first 10 elements.

B1- Different from zero the biggest 15 elements .

B2- Different from zero the biggest 15 elements among the 20 elements corresponding to the 20 biggest singular values.

B3- Different from zero the 15 elements corresponding to the 15 biggest singular values, that is, the first 15 elements.

C1- Different from zero the biggest 20 elements.

C2- Different from zero the biggest 20 elements among the 30 elements corresponding to the 30 biggest singular values.

C3- Different from zero the 20 elements corresponding to the 20 biggest singular values, that is, the first 20 elements.

First of all we studied the statistical distribution of the positions, in a sub-block, of the elements different from zero in cases A1, B1, C1. Later we obtained the synthesized signal for the nine cases stated above, for both schemes without quantizing, and finally we compared the results obtained.

4.3. Evaluation

Let's analyze first the statistical distribution of the positions. Our purpose was to study whether we could reduce the information concerning the positions by considering that there is a small probability of them having certain values. In figures 1a, 1b, 1c we have the histograms of the positions for a sub-block corresponding to the cases A1, B1, C1. The continuous line corresponds to the results obtained applying the SVD to $\mathbf{H}_w^t\mathbf{H}_w$ and the dotted line to those obtained applying the SVD to $\mathbf{H}_w$. As we can see the positions of the biggest values of (14) and (17) in cases A1, B1, C1 concentrate mostly in the first 20 positions, being this fact more exaggerated for (14) than for (17).

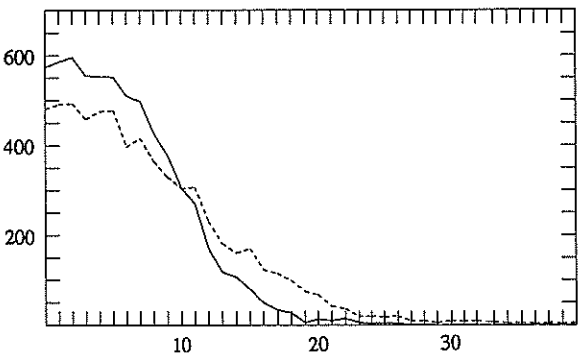


Figure 1a.- Positions histogram (case A1).

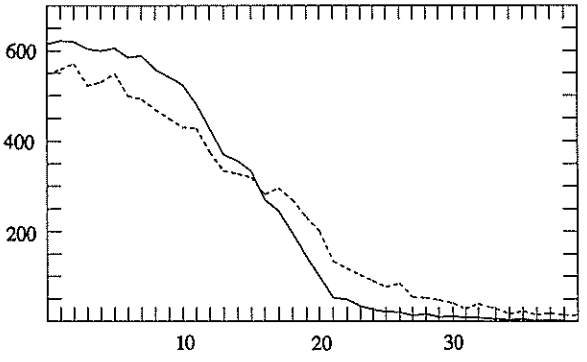


Figure 1b.- Positions histogram (case B1).

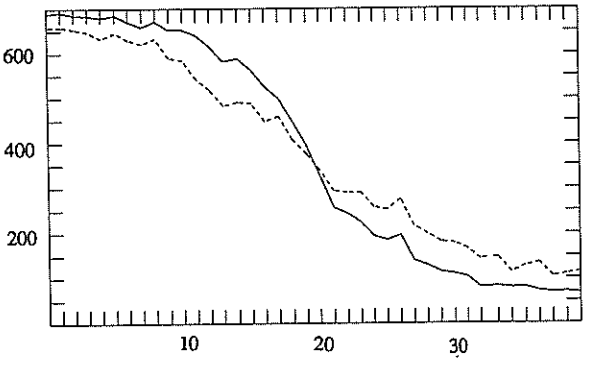


Figure 1c.- Positions histogram (case C1).

So we can conclude that in both cases the biggest values of (14) and (17) correspond to a great degree to the biggest singular values and, as we could deduct from (14) and (17), this is more often true for the first scheme. According to these results, first we are not going to consider the last 20 positions, cases A2 and B2, or the last 10 positions, case C2, and finally we will consider that the positions coincide with the positions of the biggest singular values.
Synthesizing in the nine cases described above we get the

following values for the SNRseg without quantizing, figures 2a, 2b, 2c. The results with an x correspond to the first scheme (14) and those with a + to the second (17). From the results obtained we can see that for cases A1, B1, C1 and A2, B2, C2 we get in general slightly better results for the first scheme (14) and, logically, exactly the same results for A3, B3, C3 in both schemes.

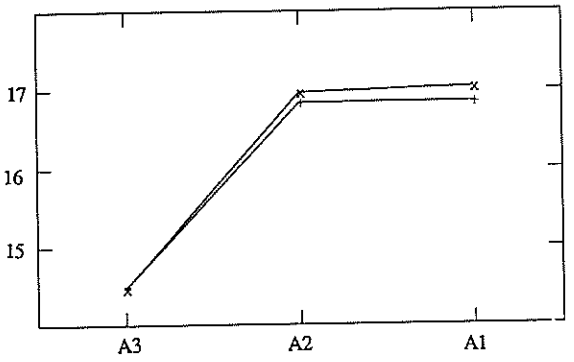


Figure 2a.- SNRseg(dB) for cases A3, A2, A1.

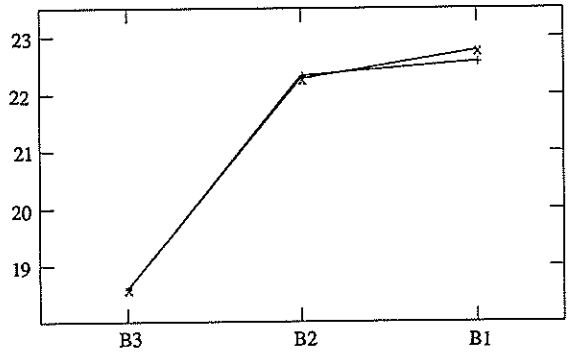


Figure 2b.- SNRseg(dB) for cases B3, B2, B1.

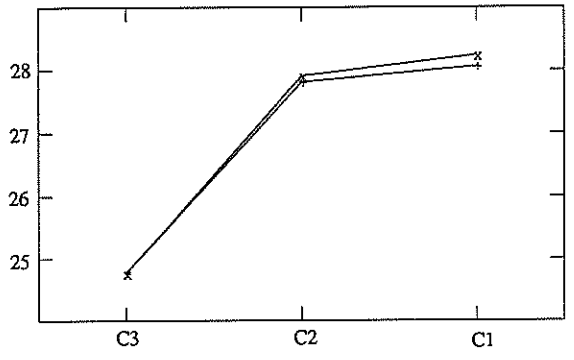


Figure 2c.- SNRseg (dB) for cases C3, C2, C1.

We can also see that the objective performance between cases A1 and A2, as well as between cases B1 and B2, and C1 and C2 is very similar, being worse for A3, B3, C3.
Perceptually we could not find any significant difference between both schemes, only a very slightly better performance of the scheme in [3] over our scheme for high frequencies. What it is noticeable for both schemes is that as we reduce the elements different from zero we loose high frequencies, sounding case A quite similar to B and C save for high frequencies.
In general we can affirm that we get speech practically indistinguishable from the original in cases C1 and C2, very high quality speech in cases C3, B1 and B2, quite a good quality speech in cases B3, A1, A2 and only acceptable quality in case A3.

5. CONCLUSIONS

As we can see we get practically no difference between both schemes, so we have not succeeded in our attempt to reduce the information concerning the positions as both schemes sound very similar. In general applying the SVD has a theoretical interest, but in practice if we quantized we would get a very high bit rate in comparison to multipulse and specially to CELP. As is suggested in [3] only by investigating the interdependence of the elements in (17) and the importance of the different singular values we could reduce the bit rate.

References

[1] B.S Atal, J.R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates", *Proc. ICASSP-82*, 1982.

[2] M.R. Schroeder, B.S. Atal, "Code-excited linear prediction (CELP): high quality speech at very low bit-rates", *Proc. ICASSP-85*, 1985.

[3] B.S. Atal, "A model of LPC excitation in terms of eigenvectors of the autocorrelation matrix of the impulse response of the LPC filter", *Proc. ICASSP-89*, 1989.

[4] S. L. Marple Jr., "Digital spectral analysis with applications", *Prentice-Hall*.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
Elsevier Science Publishers B.V., 1990

DISTANCE MEASURES PERFORMANCE IN VECTOR QUANTIZATION

Juan M. Lopez-Soler, Antonio Peinado-Herreros,
Jose C. Segura-Luna, Victoria Sanchez-Calle,
Antonio J. Rubio-Ayuso.

Dept. de Electronica y Tecnologia de Computadores
Univ. de Granada. 18071 Granada, Spain.

An attempt to find the best distortion measure for Vector Quantization is carried out by making both objective, as well as subjective studies over the same training sequence. Several 1024 codewords vector quantizers are made and compared using MSE on the short-term autocorrelations coefficients, MSE and MSE-weighted on the LPC-coefficients, and MSE-Weighted on the LPC Cepstral coefficients. Another vector quantizer is compared using the Spectral Likelihood Ratio Distortion Measure, so called Gain-Normalized Model Spectral Measure. In this paper spectral distortion is shown to be the best nearest-neighbor search rule for speech coding using VQ.

1. INTRODUCTION

On the lights of *Rate-Distortion Theory* [1] one important result is that Vector Quantization (VQ) can achieve better performance than scalar quantizers for low-rate speech coding [2].

In VQ several parameters are simultaneously coded into a single vector, unlike of scalar quantizers where the parameters are separately coded. A vector quantizer is defined by a partition of the space of all input vectors **x** (each one characterized by one template vector called the *codeword*) jointly a nearest-neighbor search rule (distortion measure) that assign each input vector with his template. The set of templates is called the *codebook*.

Well known locally optimum algorithms have been developed to built the codebook like are **LBG** or *generalized Lloyd's algorithm* [3] [4].

Some spectral distortion measures have been proposed to compare two spectral models (the input vector and all the codewords of the codebook) giving the best codeword that matches the input vector spectrum.

Other distortion measures have been developed using the Euclidean criterion called "*Mean Squared Error: MSE*", in which the nearest-neighbor is the one that minimized the distance between the input vector and all the codewords.

However there are not objective foundations to choose the distortion measure that carries out the best performance. In this paper, spectral distortion is shown to be objectively and subjectively the best nearest-neighbor rule for speech coding using **VQ**.

This work has been supported by CICYT (Comision Interministerial de Ciencia y Tecnologia) under project CIT 88-0774. Spain.

2. CODEBOOK GENERATION

As mentioned above, to design the codebook we partition the N-dimensional space of the vectors **x** into **L** clusters or cells $\{C_i, 1 < i < L\}$ and associate each cell $\mathbf{C_i}$ with a vector $\mathbf{y_i}$, called the *codeword* or *centroid*. Each codeword is characterized by its *index* **i**.

Let **n** be the number of bits necessary to code the index, and consider that an uniform assignment is done for the **L** index, so we have

$$L = 2^n \tag{1}$$

In this paper n=10, so the size of the codebook is L=1024.

Taking as a reference the **LBG** clustering algorithm we develop the following algorithm for codebook generation, where the step 3 is introduced to prevent empty cells [5]:

Step 0

Initialization: Choose an initial codebook, $\{y_i, 1 < i < L\}$ with a long training sequence.

Step 1

Classification: Classify the training sequence into the clusters $\mathbf{C_i}$ using a minimum distortion rule. Compute the cluster to which more training vector belong to.

Step 2

Padding: If any cluster is empty, take the most populous cluster and assign his centroid, slightly perturbed by a factor, to the centroid of the empty cluster.

Step 3

Updating: update the centroid of every cluster.

Step 4

Convergence test: compute $(D_{OLD} - D) / D$ and test if it is below a certain threshold. If so stop, otherwise go to step 2.

With this algorithm, different results may be achieved choosing different initial codebooks. We make the initial codebook using a "*splitting*" technique [3].

3. DISTORTION MEASURES

Many Distortion Measures have been proposed in the literature. The problem is which of such distortions achieve the best performance using VQ for speech coding, in the sense of improving the fidelity criterion.

In order to find the best distortion measure, we generate several codebooks using the proposed algorithm over the same training sequence, with the nearest-neighbor search rules listed below:

-Euclidean Distance MSE: This is a conceptually and computationally very simple algebraic distance

$$d_{MSE}(x,\bar{x}) = \left[\sum_{i=1}^{N} (x_i - \bar{x}_i)^2 \right]^{1/2} \qquad (2)$$

where x, $\bar{x}$ denote the input frame and the codeword respectively.

-Weighted Euclidean Distance MSE: In order to get that all the vector dimensions x_i are equally weighted (that is, had unit variance) we use:

$$d_{MSE-W}(x,\bar{x}) = \left[\sum_{i=1}^{N} \omega(i)\,(x_i - \bar{x}_i)^2 \right]^{1/2} \qquad (3)$$

with $\omega(i) = \sigma_x^2(i)$, $i = 0 \cdots n$, where σ is the variance of x_i.

-Likelihood Ratio: Let X(z) be the z-transform of an input frame of speech and $\sqrt{E_p} / A_P(z)$ be the Pth order LPC model of X(z). Let 1/A(z) be any Pth order all-pole filter and E the residual energy given from the inverse filtering of X(z) with A(z). d_{LR} is a non-Euclidean distance. Nevertheless it has a spectral interpretation that makes it adequate for our proposal. Minimizing d_{LR} is equivalent to minimizing the residual energy E since the optimal E_P depends only on the input. A rigorous study of the statistical properties of LPC distance measures can be found in [6] [7]. The d_{LR} for two gain-normalized inverse filters is given by:

$$d_{LR}\left(\frac{1}{A_P},\frac{1}{A}\right) = \int_{-\pi}^{\pi} \frac{|A(e^{j\theta})|^2}{|A_P(e^{j\theta})|^2} \frac{d\theta}{2\pi} = \frac{E}{E_P} - 1 \qquad (4)$$

Another computational expression for d_{LR} can be obtained using [8]

$$d_{LR}\left(\frac{1}{A_P},\frac{1}{A}\right) + 1 = \left\{ r_a(0)\,\frac{r_x(0)}{E_P} + 2 \sum_{n=1}^{P} r_a(n)\,\frac{r_x(n)}{E_P} \right\} \qquad (5)$$

where $r_x(i)$ and $r_a(i)$ denote the short-term autocorrelation sequences of the input speech data and the codeword's polynomial coefficients A(z) respectively.

So, the next five codebooks have been designed:

R.- using d_{MSE}, expression (2), over the short-term autocorrelation normalized coefficients of the input frame and the codeword respectively.

A.- using d_{MSE}, expression (2), over the polynomial coefficients LPC of A(z).

AW.- same codebook **A**, but using d_{MSE-W}, expression (3).

CW.- using d_{MSE-W}, over the LPC Cepstral coefficients **C(i)**, computed recursively from the LPC coefficients **a(i)** by [9]:

$$c(1) = -a(1) \qquad (6)$$

$$c(i) = -a(i) - \sum_{k=1}^{i-1} \left(1 - \frac{k}{i}\right) a(k)\, c(i-k)$$

$$1 < i \le p$$

LR.- using the Likelihood Ratio Spectral Distortion Measure, expression (5).

4. EXPERIMENTAL RESULT

4.1. Analysis Conditions

To generate the codebook, the training sequence was first established. It consists of speech by nine speakers, three female and six male ones, approximately eight million of speech samples, corresponding to more than 15 min of speech.

A sampling frequency of 8 Khz was used to generate 8-bit PCM speech. 15 ms was chosen for the size of the analysis window ,that is, 120 samples per frame. With this analysis conditions a bit rate of 1.2 Kbps is obtained, or equivalently 18 bits/frame. 10 of such bits were established for coding the codeword. The remainder 8 bits/frame was used to code the pitch and the gain.

4.2. Codebook Design

To show the behavior of the proposed nearest-neighbor search rules, in fig. 1 are depicted the distribution of the number of vectors belonging to the same cluster in the training sequence.

As it can be seen, A and AW codebooks have the worst behavior. On the other hand, LR and CW have a fairer distribution and almost have not empty clusters.

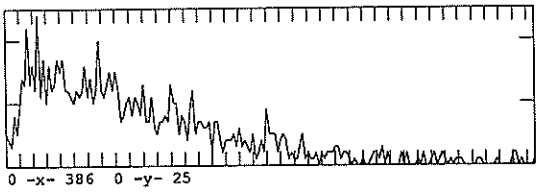


Figure 1.a: Codebook **R**

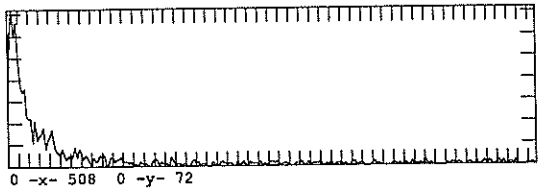


Figure 1.b: Codebook **A**

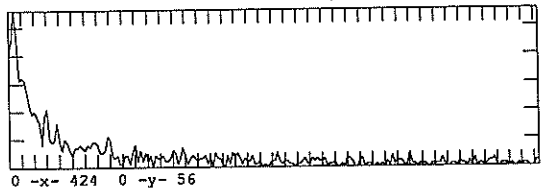


Figure 1.c: Codebook **AW**

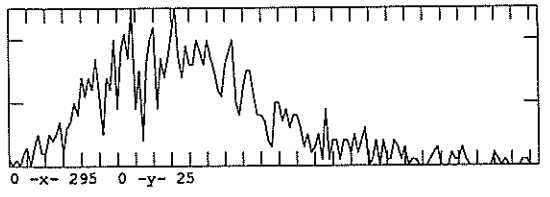


Figure 1.d: Codebook **CW**

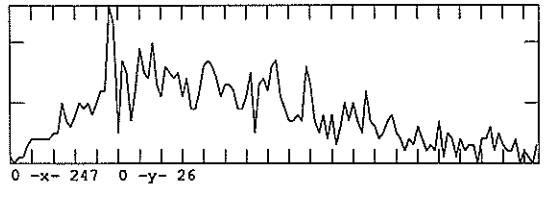


Figure 1.e: Codebook **LR**

4.3. Objective Test

We can quantify the performance of a system by an average distortion **D**, defined as [3]:

$$D_i = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} d_i(x,y) \qquad (7)$$

where **x** denotes the input frame and **y** is the corresponding synthetic one, after the coding process. $\mathbf{d_i}$ is evaluated over the five distortion measures above defined.

Two test sequences of speech were established to evaluate the objective performance of the five codebooks. The first test sequence corresponds to 5 min of speech from speakers included in the training sequence. It is called the **"IN test"**, table I. Unlike the **"OUT test"**, table II, where the speech sequence was evaluated for speakers that do not belong to the training sequence.

IN	Nearest Neighbourg Rule				
Dist.	R	A	AW	CW	LR
R	0.49	0.85	0.86	0.67	0.53
A	1.45	0.77	0.92	1.23	0.90
AW	13.05	7.07	5.65	9.58	8.01
CW	17.34	15.51	15.42	9.06	13.50
LR	2.02	1.57	1.88	1.84	1.32

Table I

OUT	Nearest Neighbourg Rule				
Dist.	R	A	AW	CW	LR
R	0.71	1.00	1.00	0.84	0.52
A	1.78	0.87	1.04	1.62	1.19
AW	10.99	6.99	5.47	9.35	8.67
CW	18.17	16.10	16.12	8.86	13.17
LR	2.19	1.70	1.96	1.93	1.36

Table II

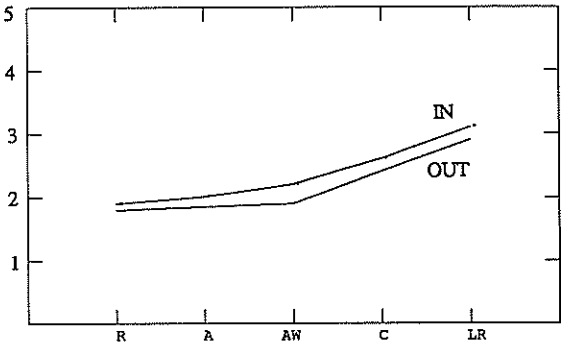


Figure 2: **MOS**

4.4. Subjective Test

While objective distortion measures are necessary and useful tools to design speech coding systems, subjective quality testing is indispensable to make an informed decision on directions for improving system performance [2]. A **MOS** *(Mean Opinion Square)* [10] was performed over the two test: IN and OUT. The result are shown in fig. 2.

5. CONCLUSIONS

- As it can be seen in fig. 2, the best subjective quality in the synthetic speech is attained with the **LR** codebook.

- Objective tests show that each codebook **i** minimize his own average distortion $\mathbf{D_i}$. Furthermore, the **LR** codebook is the second applicant almost any case.

- Therefore spectral distortions, like is the *Likelihood Ratio Distortion Measure*, seem to be more adequates than distances with algebraic and eucludiean sense, like MSE or MSE-W.

REFERENCES

[1] A.J. Viterbi, J.K. Omura: "Principles of Digital Communication and Coding". *McGraw-Hill, 1979.*

[2] J. Makhoul, S. Roucos, H. Gish: "Vector Quantization in Speech Coding". *Proc. IEEE, vol. 73, n 11, pp. 1551-1588. November 1985.*

[3] R. M. Gray: "Vector Quantization". *EEE ASSP Magazine, pp. 4-29. April 1984.*

[4] Y. Linde, A. Buzo, R.M. Gray: "An Algorithm for Vector Quantization Design". *IEEE Trans. on Comm. vol. 28 n.1, pp 84-95. January 1980.*

[5] J.M. Lopez-Soler, A. Rubio-Ayuso, et al: "A Tree Codebook Design Algorithm for Digital Speech Storage Using Vector Quantization". *IV International Symposium on Knowledge Engineering. Barcelona, Spain. May 1990.*

[6] J.M. Tribolet, L.R. Rabiner, M.M. Sondhi: "Statistical Properties of an LPC Distance Measure". *IEEE Trans. on ASSP, vol. 27, n. 5, pp. 550-559. October 1979.*

[7] R.M. Gray, A. Buzo, A.H. Gray, Y. Matsuyama: "Distortion Measures for Speech Processing". *IEEE Trans. on ASSP, vol. 28, n. 4, pp 367-376. August 1980.*

[8] B.H. Juang, D.Y. Wong, A.H. Gray: "Distortion Performance of Vector Quantization for LPC Voice Coding". *IEEE Trans. on ASSP, vol. 30 n. 2 pp 294-303. April 1982.*

[9] Y. Tohkura: "A Weighted Cepstral Distance Measure for Speech Recognition". *IEEE Trans. on ASSP, vol. 35, n. 10, pp. 1414-1422. October 1987.*

[10] N. Kitawaki, M. Honda, K. Itoh: "Speech-Quality Assessment Methods for Speech-Coding Systems". *IEEE Comm. Magazine. vol. 22, n. 10, pp. 26-33. October 1984.*

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

THE SPLIT LEVINSON ALGORITHM FOR EXTRACTING THE LINE SPECTRUM PAIRS

S. SAOUDI*, J. M. BOUCHER* and A. LE GUYADER**

*Département Mathématiques et Systèmes de Communications
ENST-Br. BP 832. 29285 Brest-FRANCE.
**Département Codage et Modèles de Communications
CNET. Route de Trégastel, BP 40. 22301 Lannion-FRANCE.

Abstract : The Line Spectrum Pairs (L.S.P.) provide an efficient representation of the synthesis filter used in Linear Predictive Coding (LPC) of speech. In this paper, different algorithms are derived and compared for the computation of the LSP parameters. We show how we can just use the symmetric (or the antisymmetric) form of the split levinson algorithm to compute new α_k and α_k^* coefficients. Then the LSP parameters are obtained by finding the eigenvalues of tridiagonal matrices derived from the α_k and α_k^*.

1. Introduction

Various speech Linear Predictive Coding methods have been studied for speech transmission at low bit rate. The LSP speech analysis method is known as one of the most powerful LPC technique. The LSP parameters are completely equivalent, in a mathematical sense, to other Linear Predictive Coding coefficients, such as the LPC or PARCOR coefficients. However, the LSP parameters have some additional interesting properties which make them more attractive than the LPC or PARCOR coefficients.
The paper is organized as follows : Section 2 includes a brief review of the LSP speech analysis method. In section 3, we recall the split levinson algorithm and the three term recurrence relation for the singular predictor polynomials for the same type. In section 4, we show how to use the split levinson algorithm to derive the LSP parameters. Another new algorithm to compute the LSP is proposed in section 5. In section 6, we show how to calculate α_k and α_k^* from only one version of the split levinson algorithm. A new algorithm for computing the symmetric and the antisymmetric polynomials from one version of the split levinson algorithm is developed in section 7. Section 8 contains conclusions.

2. LSP parameters

For a given order p, the linear predictive coding analysis of speech results in an all-pole filter H(z), described by :

$$H(z)=\frac{1}{A_p(z)}=\frac{1}{1+\sum_{i=1}^{p} a_i z^{-i}} \qquad (1)$$

The parameters $\{a_i\}_{i=1,2,...,p}$, are commonly referred to as the LPC coefficients. They are computed from the following Toeplitz set of equation[3] :

$$\sum_{i=1}^{p} a_i c_{j-i} = -c_j \quad \text{for } j = 1,..., p.$$

where $(c_i)_{i=0,1,...,p}$ are the autocorrelation coefficients of the speech signal.
The polynomial $A_j(z)$, associated with a j^{th} order LPC analysis, satisfies the following recurrence relationship :

$$A_j(z) = A_{j-1}(z) + k_j z^{-j} A_{j-1}(z^{-1}),\ j = 1,..., p. \qquad (2)$$

where $A_0(z) = 1$.
In (2), the parameters $\{k_j\}_{j=1,2,...,p}$, are called the PARCOR coefficients, they are also interpreted as the reflection coefficients at the boundaries of the acoustic tube model for the vocal tract. For j=p+1 in (2), we have :

$$A_{p+1}(z) = A_p(z) + k_{p+1} z^{-(p+1)} A_p(z^{-1}) \qquad (3)$$

In (3), consider two extrem artificial boundary conditions : $k_{p+1}=1$ and $k_{p+1}=-1$. These conditions correspond to a complete closure and a complete opening at the glottis in the acoustic tube model. Then, the polynomial $A_{p+1}(z)$ can be expressed as : (for $k_{p+1}=+1$ and $k_{p+1}=-1$)

$$P_{p+1}(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) \qquad (4)$$

$$P^*_{p+1}(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) \qquad (5)$$

$P_{p+1}(z)$ is symmetric and $P^*_{p+1}(z)$ is antisymmetric in the sense that they satisfiy :

$$P_{p+1}(z) = z^{-(p+1)} P_{p+1}(z^{-1})$$

$$P^*_{p+1}(z) = -z^{-(p+1)} P^*_{p+1}(z^{-1})$$

The symmetric and antisymmetric predictor polynomials possess some very interesting and important properties summarized in the following [7],[6],[5] :
- for a stable $A_p(z)$, all roots of $P_{p+1}(z)$ and $P^*_{p+1}(z)$ lie on the unit circle (singular polynomials).
- the roots of the two symmetric and antisymmetric

polynomials alternate each other on the unit circle.
It is easily shown that the polynomials $P_{p+1}(z)$ and $P^*_{p+1}(z)$ can be expressed ,for p even, as :

$$P_{p+1}(z) = (1+z^{-1}) \prod_{i\ odd} (1 - 2\cos\omega_i z^{-1} + z^{-2})$$
$$P^*_{p+1}(z) = (1-z^{-1}) \prod_{i\ even} (1 - 2\cos\omega_i z^{-1} + z^{-2}) \quad (6)$$

where :

$$0 < \omega_1 < \omega_2 < \ldots < \omega_p < \pi. \quad (7)$$

Throughout this paper, we will confine our attention to even values of p. The parameters $\{\omega_i\}i=1,2,...,p$ are defined as the Line Spectrum Pair (LSP) parameters.
From eq. 6, it is clear that the LSP parameters are given by the roots of the two polynomials $P^*_{p+1}(z)$ and $P_{p+1}(z)$.Thus they can be obtained by one of the following methods : the direct solving method (if p is less or equal to 8, it is well known that any polynomial of order 4 or less can be solved through its radicals), Newton Raphson, the direct Fast Fourier transform or the inverse cosine transformation.
More recently, DELSARTE and GENIN [1], suggested the split levinson algorithm for reducing the complexity of levinson algorithm. In this paper, we use an analogue formulation for the computation of the LSP frequencies $(\omega_i)_{i=1,...,p}$. First, we obtain intermediate coefficients α_k and α^*_k from the split levinson algorithm and then the LSP are obtained from the eigenvalues of tridiagonal matrices derived from the α_k and α^*_k.

3. The split levinson algorithm

The levinson algorithm has been shown to be redundant in complexity [1]. It can be broken down into a symmetric and antisymmetric form, either of which needs only to be processed to compute the predictor polynomial $A_p(z)$.
The symmetric form of the split levinson algorithm computes recursively the symmetric singular predictor polynomials : (k=1,2,...,p+1)

$$P_k(z) = A_{k-1}(z) + z^{-k} A_{k-1}(z^{-1}) = \sum_{i=0}^{k} p_{k,i} z^{-i} \quad (8)$$

The antisymmetric form of the split levinson algorithm computes recursively the antisymmetric singular predictor polynomials : (k=1,2,....,p+1)

$$P^*_k(z) = A_{k-1}(z) - z^{-k} A_{k-1}(z^{-1}) = \sum_{i=0}^{k} p^*_{k,i} z^{-i} \quad (9)$$

For k=p+1, we obtain the two symmetric and antisymmetric singular predictor polynomials (of eq.6) which roots give the Line Spectrum Pairs.
The symmetric form version of the split levinson algorithm consists of the following operations [1] :
For k=1,2,...,p, compute :

$$\tau_k = \sum_{i=0}^{t-1} (c_i + c_{k-i}) p_{k,i} \qquad \text{for } k = 2t-1,$$
$$= \sum_{i=0}^{t-1} (c_i + c_{k-i}) p_{k,i} + c_t p_{k,t} \quad \text{for } k = 2t$$
$$\alpha_k = \frac{\tau_k}{\tau_{k-1}}$$
$$p_{k+1,i} = p_{k,i} + p_{k,i-1} - \alpha_k p_{k-1,i-1};\ 1 \le i \le t$$
$$\lambda_k = 2 - \frac{\alpha_k}{\lambda_{k-1}}. \quad (10)$$

The initial conditions are :

$$p_{k,0} = 1 \quad \text{for } k \ge 1,$$
$$p_{0,0} = 2,\ p_{1,1} = 1,$$
$$\tau_0 = c_0,\ \lambda_0 = 1. \quad (11)$$

In the case the reflection coefficients are required. They are simply given by :

$$\rho_k = \lambda_k - 1, \quad (12)$$

The predictor polynomial is then determined by :

$$(1-z^{-1}) A_p(z) = P_{p+1}(z) - \lambda_p z^{-1} P_p(z) \quad (13)$$

The antisymmetric version of the split levinson algorithm is obtained along the same lines as the version above [1], with the initial conditions :

$$p^*_{k,0} = 1 \text{ for } k \ge 1,$$
$$p^*_{0,0} = 0,\ p^*_{1,1} = -1,$$
$$\tau^*_0 = c_0,\ \lambda^*_0 = 1. \quad (14)$$

The reflection coefficients are simply given by :

$$\rho_k = 1 - \lambda^*_k \quad (15)$$

The three-term recurrence relations for the singular predictor polynomials for the same type are derived from eq. 10 :

$$P_{k+1}(z) - (1+z^{-1}) P_k(z) + \alpha_k z^{-1} P_{k-1}(z) = 0$$
$$\text{with } \alpha_k = \frac{\tau_k}{\tau_{k-1}}. \quad (16)$$

$$P^*_{k+1}(z) - (1+z^{-1}) P^*_k(z) + \alpha^*_k z^{-1} P^*_{k-1}(z) = 0$$
$$\text{with } \alpha^*_k = \frac{\tau^*_k}{\tau^*_{k-1}}. \quad (17)$$

4. Computing the LSP from the α_k and α^*_k coefficients

DELSARTE and GENIN extend the method to the computation of the Pisarenko frequencies. Here, we use an analogue formulation for the computation of the LSP frequencies. Let us define the two following real functions from the symmetric and antisymmetric singular predictor polynomials using the change of variable

$$z = e^{j\omega} \Rightarrow x = \frac{z^{1/2} + z^{-1/2}}{2} = \cos\left(\frac{\omega}{2}\right) \quad (18)$$

$$F_k(z) = \tau_k^{-1/2} z^{k/2} P_k(z)$$
$$F^*_k(z) = \tau^{*-1/2}_{k+1} \frac{z^{k/2}}{1-z^{-1}} P^*_{k+1}(z) \quad (19)$$

and let us define :

$$\beta_k = \frac{\alpha_k^{1/2}}{2},\ \beta^*_k = \frac{\alpha^{*1/2}_k}{2} \quad (20)$$

The formulas (16) and (17) take the following form using (18) and (19) : (k≥1)

$$\beta_{k+1}F_{k+1}(x) - xF_k(x) + \beta_k F_{k-1}(x) = 0 \quad (21)$$

$$\beta^*_{k+1}F^*_k(x) - xF^*_{k-1}(x) + \beta^*_k F^*_{k-2}(x) = 0 \quad (22)$$

The initial conditions are determined by (11) and (14) :

$$2\beta_1 F_1(x) - xF_0(x) = 0$$
$$\beta^*_2 F^*_1(x) - xF^*_0(x) = 0 \quad (23)$$

From (21) and (22), we can deduce the three-term recurrence for real polynomials of the same parity : (k ≥ 2)

$$\beta_{k+1}\beta_{k+2}F_{k+2}(x) - (x^2 - \beta_k^2 - \beta_{k+1}^2)F_k(x) + \beta_{k-1}\beta_k F_{k-2}(x) = 0 \quad (24)$$

$$\beta^*_{k+1}\beta^*_{k+2}F^*_{k+1}(x) - (x^2 - \beta_k^{*2} - \beta_{k+1}^{*2})F^*_{k-1}(x) + \beta^*_{k-1}\beta^*_k F^*_{k-3}(x) = 0 \quad (25)$$

The initial conditions are (using (23),(11) and (14)) :

$$\beta_2\beta_3 F_3(x) - (x^2 - 2\beta_1^2 - \beta_2^2)F_1(x) = 0$$
$$\beta^*_2\beta^*_3 F^*_2(x) - (x^2 - \beta_2^{*2})F^*_0(x) = 0 \quad (26)$$

The complete set of identities (24),(25) using the initial conditions (26) can be expressed in the classical matrix form as :

$$(J_{p/2} - x^2 I_{p/2})\left[F_1(x), F_3(x), \ldots, F_{p-1}(x)\right]^T = \left[0, \ldots, -\beta_p\beta_{p+1}F_{p+1}(x)\right]^T \quad (27)$$

$$(J^*_{p/2} - x^2 I_{p/2})\left[F^*_0(x), F^*_2(x), \ldots, F^*_{p-2}(x)\right]^T = \left[0, \ldots, -\beta^*_p\beta^*_{p+1}F^*_p(x)\right]^T \quad (28)$$

($I_{p/2}$ is the unite square matrix of order p/2), with :

$$J_{p/2} = \begin{bmatrix} 2\beta_1^2+\beta_2^2 & \beta_2\beta_3 & 0 & & 0 \\ \beta_2\beta_3 & \beta_3^2+\beta_4^2 & \beta_4\beta_5 & 0 & 0 \\ 0 & \beta_4\beta_5 & & & 0 \\ & & & & \\ 0 & & \beta_{p-2}\beta_{p-1} & \beta_{p-1}^2+\beta_p^2 & \end{bmatrix} \quad (29)$$

and

$$J^*_{p/2} = \begin{bmatrix} \beta_2^{*2} & \beta^*_2\beta^*_3 & 0 & & 0 \\ \beta^*_2\beta^*_3 & \beta_3^{*2}+\beta_4^{*2} & \beta^*_4\beta^*_5 & 0 & 0 \\ 0 & \beta^*_4\beta^*_5 & & & 0 \\ & & & & \\ 0 & & \beta^*_{p-2}\beta^*_{p-1} & \beta_{p-1}^{*2}+\beta_p^{*2} & \end{bmatrix} \quad (30)$$

Therefore, the formulas (29) and (30) show that the Line Spectrum Pair parameters (solution of $F_{p+1}(x)=0$ and $F^*_p(x)=0$) are directly available from the eigenvalues of the matrix $J_{p/2}$ and $J^*_{p/2}$. From (18) the eigenvalues are : $x_i^2 = \cos^2(\omega_i/2)$, where $(\omega i)_{i=1,\ldots,p}$ are the LSP parameters.
Then the eigenvalues can be found by using the fast algorithm existing for symmetric tridiagonal matrices [4].

A disadvantage of this method is to obtain $\cos^2(\omega_i/2)$ instead of $\cos(\omega_i)$. In the following section, we will present a new change of variable which leads to direct computation of $\cos(\omega_i)$.

5. Another new algorithm to compute the LSP parameters

Let us define the two following real functions from the symmetric and antisymmetric singular predictor polynomials using the change of variable :

$$z = e^{j\omega} \Rightarrow x = (z + z^{-1}) = 2\cos\omega \quad (31)$$

$$H_k(x) = z^{k/2}P_k(z)$$
$$H^*_k(x) = \frac{z^{k/2}}{1 - z^{-1}}P^*_{k+1}(z) \quad (32)$$

From the formulas (16) and (17), we can deduce the three-term recurrence for the real polynimials $H_k(x)$ and $H^*_k(x)$ of the same parity : (k ≥ 2)

$$H_{k+2}(x) - (x - \alpha_k - \alpha_{k+1} + 2)H_k(x) + \alpha_{k-1}\alpha_k H_{k-2}(x) = 0 \quad (33)$$

$$H^*_{k+1}(x) - (x - \alpha^*_k - \alpha^*_{k+1} + 2)H^*_{k-1}(x) + \alpha^*_{k-1}\alpha^*_k H^*_{k-3}(x) = 0 \quad (34)$$

The initial conditions are : (using (11) and (14))

$$H_3(x) - (x - 2\alpha_1 - \alpha_2 + 2)H_1(x) = 0$$
$$H^*_2(x) - (x - \alpha^*_2 + 2)H^*_0(x) = 0 \quad (35)$$

The complete set of identities (33),(34) using the initial conditions (35) can be expressed in the classical matrix form as :

$$(A_{p/2} - x\,I_{p/2})\left[H_1(x), H_3(x), \ldots, H_{p-1}(x)\right]^T = \left[0, \ldots, H_{p+1}(x)\right]^T \quad (36)$$

$$(A^*_{p/2} - x\,I_{p/2})\left[H^*_0(x), H^*_2(x), \ldots, H^*_{p-2}(x)\right]^T = \left[0, \ldots, H^*_p(x)\right]^T \quad (37)$$

with,

$$A_{p/2} = \begin{bmatrix} 2\alpha_1+\alpha_2-2 & 1 & 0 & & 0 \\ \alpha_2\alpha_3 & \alpha_3+\alpha_4-2 & 1 & 0 & 0 \\ & & & & \\ 0 & & 0\ \alpha_{p-2}\alpha_{p-1} & \alpha_{p-1}+\alpha_p-2 & \end{bmatrix} \quad (38)$$

and

$$A^*_{p/2} = \begin{bmatrix} \alpha^*_2 - 2 & 1 & 0 & & & 0 \\ \alpha^*_2\alpha^*_3 & \alpha^*_3 + \alpha^*_4 - 2 & 1 & 0 & & 0 \\ & & & & & \\ 0 & & & 0\ \alpha^*_{p-2} & \alpha^*_{p-1} & \alpha^*_{p-1} + \alpha^*_p - 2 \end{bmatrix} \quad (39)$$

The LSP parameters (solution of $H_{p+1}(x)=0$ and $H^*_p(x)=0$) are determined from the eigenvalues $x_i=2\cos\omega_i$ of the tridiagonal matrices $A_{p/2}$ and $A^*_{p/2}$.

6. A new algorithm for computing α_k and $\overset{*}{\alpha}_k$ from only one version of the split levinson algorithm

We will show now that the parameters $(\beta_i)_{i=1,...,p}$ and $(\beta^*_i)_{i=1,...,p}$ or $(\alpha_i)_{i=1,...,p}$ and $(\alpha^*_i)_{i=1,...,p}$ could be given by the same version of the split levinson algorithm. From the symmetric version, we have (using (12) and (15)) :

$$\lambda_k + \lambda^*_k = 2 \quad \text{and} \quad \lambda_k = 2 - \alpha_k / \lambda_{k-1} \quad (40)$$

It follows, from these two equations, that :

$$\alpha_k = \lambda^*_k(2 - \lambda^*_{k-1}) \quad (41)$$

and from the antisymmetric version, we have :

$$\lambda^*_k = 2 - \alpha^*_k / \lambda^*_{k-1} \quad (42)$$

which leads to : (using (40))

$$\alpha^*_k = \lambda_k(2 - \lambda_{k-1}) \quad (43)$$

If we use the symmetric form of the split levinson algorithm, the set of (α^*_k) is simply derived by (43), and if we use the antisymmetric version form, the set of (α_k) is simply derived by (41) in the main loop of the algorithm. Then one form of the split levinson algorithm is needed to compute the Line Spectrum Pair parameters.

7. A new algorithm for computing the symmetric and the antisymmetric polynomials from only one version of the split levinson algorithm

The singular predictor polynomial antisymmetric version can also be derived from the symmetric version of the split levinson algorithm. The predictor polynomial can be written as :

$$A_p(z) = \frac{P_{p+1}(z) + P^*_{p+1}(z)}{2}$$

from (13) we deduce :

$$(1 - z^{-1})P^*_{p+1}(z) = 2P_{p+1}(z) - 2\lambda_p z^{-1}P_p(z) - (1 - z^{-1})P_{p+1}(z)$$

or,

$$(1 - z^{-1})P^*_{p+1}(z) = (1 + z^{-1})P_{p+1}(z) - 2\lambda_p z^{-1}P_p(z)$$

then

$$P^*_{p+1,i} = P^*_{p+1,i-1} + P_{p+1,i} + P_{p+1,i-1} - 2\lambda_p P_{p,i-1}$$

with $p^*_{p+1,0} = 1$.

Only half of the coefficients $p^*_{p+1,i}$ is needed to be computed, the second half is deduced by antisymmetric.

The symmetric predictor polynomial is obtained in a similar way from the two last antisymmetric predictor polynomials by :

$$(1 - z^{-1})P_{p+1}(z) = (1 + z^{-1})P^*_{p+1}(z) - 2\lambda^*_p z^{-1}P^*_p(z)$$

The symmetric and antisymmetric predictor polynomials could be computed by the same version of the split levinson algorithm. The roots of the two polynomials can be computed by the methods described in section 2. Complexity reduction is obtained by using the split levinson algorithm [1].

8. CONCLUSIONS

Instead of using the classical levinson algorithm, we have shown how to use only the symmetric or the antisymmetric form of the split levinson algorithm to compute the LSP parameters. We have proved that the LSP parameters can be given by the eigenvalues of some tridiagonal matrices with its elements $(\beta_i)_{i=1,...,p}$ and $(\beta^*_i)_{i=1,...,p}$ or $((\alpha_i)_{i=1,...,p}$ and $(\alpha^*_i)_{i=1,...,p})$ derived from the symmetric (or just the antisymmetric) form of the split levinson algorithm. Note that we can use in a similar way only one version of the split lattice algorithm form ([2]) to extract the LSP parameters with methods described above.

For speech coding applications, we have derived efficient schemes for quantizing the LSP parameters and now we currently use these parameters for CELP coding at bit rate below 8Kbits/s.

REFERENCES

[1] P. DELSARTE and Y. GENIN,"The split levinson algorithm", IEEE Trans. Acoust., speech, and signal processing, Vol. ASSP-34, N°.3, pp. 470-478, June 1986.

[2] P. DELSARTE and Y. GENIN,"On the splitting of classical algorithms in linear prediction theory", IEEE Trans. Acoust., speech, and signal processing, Vol. ASSP-35, N°.5, pp. 645-653, May 1987.

[3] J.D. MARKEL and A.H. GRAY,"Linear prediction of speech", Springer Verlag, 1976.

[4] N. PARLETT,"The symmetric eigenvalue problem", Prentice-Hall, Series in computational mathematics, 1980.

[5] S. SAITO and K. NAKATA,"Fundamentals of speech signal processing", Academic press, 1985.

[6] F. K. SOONG and B.H. JUANG,"Line spectrum pair (LSP) and speech data compression", in Proc. 1984 IEEE, Intern. Conf. Acoust., speech, and signal processing, San Diego, CA, pp.1.10.1-1.10.4, 1984.

[7] N. SUGAMURA and F. ITAKURA,"Speech analysis and synthesis methods developed at ECL in NTT - form LPC to LSP-", Speech communication, North-Holland, Vol.5,N°2, pp.199-215, June 1986.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

SINGLE DSP HIGH QUALITY SPEECH CELP AT 8.0 TO 4.8 KBITS/SEC

D. K. Baghbadrani and Professor C. Xydeas,
Speech and Image Processing Group, Department of Electrical Engineering,
University of Manchester, Dover Street, Manchester M13 9PL, U.K.

S. Morley,
Dowty Advanced Development Centre, Mayze House,
Westmead, Wiltshire, U.K.

This paper describes the theory and implementation of a high speech quality CELP codec operating at 8.0 to 4.8 kbits/sec. A new class of efficient structured codebooks with minimal storage and improved codebook search characteristics is also presented.

I Introduction

Analysis by Synthesis LPC based algorithms (AbS-LPC) currently lead the development of low bit rate speech coding systems. These systems operate on successive short-term speech frames and calculate, at the transmitter, a set of quantised parameters which best reconstruct the input speech signal subject to an error criterion. There are three main categories of such LPC systems namely: Multipulse (MPE) [1], Backward Excitation Recovery (BER) [2] and Codebook Excited (CELP) [3], which differ principally in the way that the excitation signal, used to drive the vocal track filter, is derived. AbS-LPC systems are capable of producing high quality speech at bit rates in the range of 8 to 4.8 kbits/sec, particularly CELP which has been recently confirmed as the 4.8 kbits/sec USA-DOD standard [4] and the 8kbits/sec North America standard [5].

CELP systems exhibit however relatively poor performance at encoding high frequency speech components. This leads to a smooth but "flat" recovered speech quality, particularly for female speech where the "sharpness" of the speech signal is compromised. Currently intense research effort is being directed into "lifting" CELP's quality, closer to toll quality speech, as well as reducing the complexity of the algorithm without deteriorating its high quality speech characteristics.

Systems offering improved CELP speech quality have been already proposed where modifications are suggested to the modelling of the excitation source [4] and where some form of post-filtering is recommended [4,5]. In particular, the HLTP approach, for estimating the parameters of the LTP part of the CELP synthesis filter, is mentioned in [6] as a useful way of improving output speech quality. Undoubtedly other CELP-quality improving techniques will be proposed in the near future, particularly towards the end of the GSM "half rate channel" speech coder standardisation exercise which seeks to deliver a 8.0 to 6.0 kbits/sec coder with better output speech quality than that of the "full rate" RPE-LPC 13 kbits/sec GSM coder.

Effective complexity reduction methods have been also proposed which i) exploit the properties of structured codebooks [5,7,8], and/or ii) approximate the error measure used in the codebook search process [9,10].

This paper describes the theory and implementation of a high quality speech, 8.0 to 4.8 kbits/sec CELP codec which requires only a single DSP. The system is an extension of previous work reported in [9,10] and employs a ternary codebook and a new simplified error measure. This combination results in an efficient two-stage codebook search strategy and allows the implementation of the proposed codec on a single AT&T PSP32C device, operating in full duplex mode at 8.0, 6.0 and 4.8 kbits/sec. Furthermore, the adopted ternary codebook and simplified error measure are used to define a new class of efficient structured codebooks with minimal storage and improved codebook search characteristics.

II System Description

CELP coders consist of an LPC based vocal tract model, a codebook based excitation model and an error criterion which serves to select an appropriate excitation sequence in an AbS optimisation process. The vocal tract model utilises both a short term filter (STF), which models the spectral envelope of speech, and a long term filter (LTF), which accounts for pitch periodicity in voiced speech.

The LTF can be defined "outside" or "inside" the AbS process by minimising i) the energy of the second residual or ii) the energy of the actual (weighted) error between the input and decoded speech, respectively. In the latter case the function of the LTP can be viewed as equivalent to an adaptive excitation codebook [4].

The main excitation codebook was originally designed as a collection of random "Gaussian" vectors [3]. In order to reduce codebook storage requirements and the potentially huge complexity of the codebook search process, efficient "random" and "structured" codebooks have been proposed. These, combined with simplified error measures, resulted in implementable CELP codecs using current DSP technology.

The theory and implementations of such a 6/4.8kbits/sec CELP codec was presented in [9,10]. The codebook used is ternary, with excitation vectors derived from Gaussian sparse excitation vectors by considering only the signs of the np non-zero pulses. Also, the system employed an approximation of the actual error energy criterion.

The codec proposed in this paper retains the ternary codebook and is based on a new and improved excitation codebook search strategy. Consider the codebook search process of a single gain CELP. The process locates the codebook entry which minimises:

$$E^k = \|\gamma^k H C^k - X\|^2 \qquad (1)$$

where $k \in \{1,\ldots L\}$ is the excitation codebook entry, H is the convolutional matrix of the combined LTF and perceptually weighted STF, X is the perceptually weighted input speech vector after the synthesis filter memory has been removed, γ^k is the excitation gain and L is the codebook size. Minimising (1) with respect to γ^k gives:

$$E^k = \|X\|^2 - \frac{(X^T H C^k)^2}{\|H C^k\|^2} \qquad (2) \quad , \quad \gamma^k = \frac{X^T H C^k}{\|H C^k\|^2}$$

Minimisation of (2) is equivalent to maximisation of:

$$\theta^k = \frac{(X^T H C^K)^2}{\|H C^k\|^2} \qquad (3)$$

which is a complex expression to form L times. The problem has been relatively eased with the approximation [9,10]:

$$\|H C^k\|^2 = \mu_o \nu_o + 2 \sum_{i=1}^{N} \mu_i \nu_i^k \qquad (4)$$

N is the length of the excitation vector, μ_i is the autocorrelation of the combined filter impulse response and ν_i^k is the autocorrelation of the C^k codeword.

An inspection of (4) reveals however that $\mu_o \nu_o$ is constant (for unity variance codes) and has usually a greater magnitude than the following summation. This suggests that the influence, in the codebook search process, of the denominator in θ^k is minimal and can be ignored.

The usefulness of using $(X^T H C^k)^2 = \theta^k$ as an error measure for the codebook search has been established by observing that the optimum codeword (which is defined by maximising θ^k in equation 3) is most likely to be within the first 1-2% of the codebook entries with the largest $(X^T H C^k)^2$ contribution. In fact it was found experimentally that such a small subcodebook (holding 1-2% of the codebook vectors) contains for 95% of the synthesis frames the optimum codeword.

This very simple error criterion has been used in the proposed system to "scan quickly" the ternary codebook and thus define a small subset of the codebook entries. This is then searched using the actual error function of equation (3).

This "combination" of the simple and actual error measures gives rise to a CELP which is implementable on a single DSP and whose performance is superior to that of other simplified CELP algorithms [9]. Table 1 shows SEGSNR values obtained from the proposed codec, for different subcodebook sizes ns. It is clear that SEGSNR approaches rapidly the value obtained from a complex CELP which employs equation (3) for all the vectors in the codebook.

Table 2 provides the bit allocation strategy used to produce 8.0, 6.0 and 4.8 kbits/sec codecs respectively. In these systems LPC analysis is performed using the Burg algorithm whereas the three tap LTF is modelled from the short term prediction error.

The codec has been implemented on a single AT&T DSP32C device at the above three bit rates, and operates in a full duplex mode. Furthermore, the complexity of the algorithm is sufficiently small to allow the implementation of an echo cancellation algorithm within the same device.

III Towards a new Class of Structured Codebooks

$\hat{\theta}^k = (X^T HC^k)^2$, used to define the subcodebook in the previous system, can be expressed as

$$\left[\sum_{i=1}^{np} \lambda_{pi} C^k_{pi} \right]^2 \qquad (5)$$

where $\lambda_i = \sum_{n=i}^{N-1} h_{n-i} X_n \qquad i=0,1,\ldots N-1$

and np is equal to the number of non-zero unit magnitude pulses in the excitation vector, P_i indicates the position of the ith (i.e., C_{pi}) pulse. $\{h_n\}$ and $\{X_n\}$ are the impulse response of the synthesis filter and the perceptually weighted input signal (after the memory from previous frames has been removed) respectively. N is the excitation frame length.

According to equation (5) $\hat{\theta}^k$ is maximised when

$$\sum_{i=1}^{np} |\lambda_{pi}| \text{ (6) is maximum}$$

and sign (C^k_{pi}) = sign (λ_{pi}). (7)

$\{\lambda_i\}$, i=0,1,...N-1 can be used therefore as the key element in an excitation search strategy where the positions of the ±1 excitation pulses and their signs are determined from equations (6) and (7). Thus, in theory, equations (6) and (7) define an "optimum" excitation vector which is one of $^{N}C_{np}$ possible vectors. This approach however leads to performance and complexity limitations, due to the simplicity of $\hat{\theta}^k$ and the usually very large number of $^{N}C_{np}$ position patterns, respectively. A better strategy is to employ equations (6) and (7) to formulate a subset, from all possible vectors, which is then searched using equation (3) or equations (3) and (4).

Two schemes following this strategy have been considered. In scheme I $\{\lambda_i\}$ is formed for i=0,1,...,N-1 and the largest npt $|\lambda_{pi}|$ values, i=1,...,npt, are determined where npt > np. A subcodebook of excitation vectors is then defined which consists of all $^{npt}C_{np}$ possible position patterns and the corresponding signs from $\{\lambda_i\}$. The subcodebook is formed "on line" and its size depends on npt and np. Table 3 provides the SEGSNR values of a CELP for different subcodebook sizes ns. The codec employs equation (3) in the subcodebook search procedure and operates at 8.333 kbits/sec with a bit allocation given in table 4. Notice that as N=30 and np=3 the effective size of the excitation codebook is 12 bits for the position pattern plus three bits for the signs of the excitation pulses. As shown however in table 3, satisfactory SEGSNR values are obtained with relatively large subcodebook sizes and this implies an unacceptable level of complexity.

Scheme II overcomes this problem by subsampling $\{\lambda_i\}$ prior to applying the excitation search process of Scheme I. Table 3 shows SEGSNR values obtained for different npt and ns values. $\{\lambda_i\}$ has been subsampled by a factor of two. This reduces the effective size of the codebook to (9+3) bits and the overall bit rate to 7.533 kbits/sec. Furthermore the system produces high SEGSNR values for relatively small subcodebook sizes. Table 3 also shows SEGSNR values when the system employs equations (3) and (4) in the subcodebook search procedure (shown as Scheme III).

Conclusions

A single DSP high quality speech codec operating at 8.0, 6.0 and 4.8 kbits/sec is presented in this paper. The system employs a ternary codebook and an efficient two stage search strategy. Furthermore, a new class of structured codebooks is defined which when coupled with a subcodebook search strategy produce low complexity high speech quality CELP codecs.

References

1. Atal B.S., and Remde, J.R. "A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates", IEEE Proc. ICASSP 1982, pp 614-617.

2. Gouvianakis N., and Xydeas C.S. "Advances in Analysis - By Synthesis LPC Speech Coders", Special Issue of IERE on Mobile Radio, Nov/Dec 1987, pp S272-S286.

3. Atal B.S., and Schroeder M.R. "Stochastic Coding of Speech Signals at Very Low Bit Rates", Conference Record IEEE Int. Conf. on Communications,pp 1610-1613, Amsterdam, May 1984.

4. Campbell, J.P., et al. "An Expandable Error-Projected 4800 bPS CELP Coder (U.S. Federal Standard 4800 bPS Voice Coder)", IEEE Proc. ICASSP 1989, pp 735-737.

5. Gerson I., and Jasiuk M.A., "Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8KbPS", IEEE ICASSP 1990, pp 461-464.

6. Kroon P, and Atal B.S., "Pitch Predictors with High Temporal Resolution", IEEE ICASSP 1990, pp 661-664.

7. Lin D., "Speech Coding Using Efficient Pseudo-Stochastic Block Codes", IEEE Proc. ICASSP 1987, pp 1354-1357.

8. Ireton M.A., and Xydeas C.S, "On Improving Vector Excitation Coders Through the Use of Spherical Lattice Codebooks (SLC's)", IEEE ICASPP 1989, pp 57-60.

9. Xydeas C., Ireton M.A., and Baghbadrani D.K., "Theory and Real-Time Implementation of a CELP Coder at 4.8 and 6.0 kbps using Ternary Code Excitations", Proc. IERE 5th Int. Conf. on Digital Processing of Signals in Comms.", University of Loughborough, September 1988, pp 167-174.

10. Baghbadrani D.K., Xydeas C.S., and Morley S., "A Single DSP High Speech Quality CELP Codec", IEE Colloquium on "Speech Coding" organised by Professional Group E5 (Signal Processing), 9 October 1989, Digest No. 1989/112, pp 611/615.

Codebook Size=512, np=4, Bit Rate=6.0 kbits/sec

ns :	1	3	5	7	9
SNR	9.2	9.94	10.27	10.54	10.75

SNR for complex CELP: 10.9

Table 1

npt	ns	Scheme I (SNR)	Scheme II (SNR)	SCHEME III (SNR)
6	20	11.7	12.63	12.45
7	35	12.3	13.0	12.86
8	56	12.6	13.36	13.2
9	84	13.01	13.4	13.28
10	120	13.2	13.65	13.4
11	165	13.37	13.67	13.46
12	220	13.6	13.73	13.52
13	286	13.79	13.78	13.58
14	364	13.88	13.79	13.64
15	455	14.0	13.8	13.68

Table 3

4.8 kbits/sec CODEC.ns=8,np=6

Parameter	Framelength	bits/30msec
12 STF-COEFFS	240	51
3 LTF-COEFFS	120	2*9
1 LTF-DELAY	120	2*7
1 EXL-INDEX	60	4*9
1 EXC-GAIN	60	4*6
1 SYNC	240	1
Total :		144

6.0 kbits/sec CODEC ns=8, np=4

Parameter	Frame Length	bits/30msec
12 STF-COEFFS	240	51
3 LTF-COEFFS	120	2*12
1 LTF-DELAY	120	2*7
1 EXL-INDEX	40	6*9
1 EXC-GAIN	40	6*6
1 SYNC	240	1
Total :		180

8.0 kbits/sec CODEC ns=8, np=3

Parameter	Frame Length	bits/30msec
12 STF-COEFFS	240	51
3 LTF-COEFFS	120	2*12
1 LTF-DELAY	120	2*7
1 EXL-INDEX	24	10*9
1 EXC-GAIN	24	10*6
1 SYNC	240	1
Total :		240

Table 2

Parameter	Frame Length	bits/30msec
12 STF-COEFFS	240	51
3 LTF-COEFFS	120	2*12
1 LTF-DELAY	120	2*7
1 EXL-INDEX	30	8*12
1 EXC-GAIN	30	8*5
3 SIGNS	30	8*3
1 SYNC	240	1
Total :		250

Table 4

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ROBUST LPC VECTOR QUANTIZATION BASED ON KOHONEN'S DESIGN ALGORITHM

José A. Rodríguez-Fonollosa, Enrique Masgrau, A.Moreno

Departament de Teoria del Senyal i Comunicacions
Universitat Politècnica de Catalunya
Apdo. 30.002, 08080 Barcelona, Spain.

This paper describes a Multistage Vector Quantization scheme in which the codewords are adapted to follow the input statistics. This adaptation is computationally very simple and requires no additional bit transmission. The adaptation algorithm is shown to be closely related with the vector quantizer design techniques known as LBG and Kohonen's. We have studied the application of the developed scheme to quantize the LPC parameters and some results are presented in which the resulting adaptive structure is shown to outperform not only the non-adaptive multistage vector quantizer but also the conventional full-search vector quantizer.

1. INTRODUCTION

Vector quantization (VQ) is a simultaneous quantization of a sequence of samples or vector. This process allows to make effective use of the interrelations among the different vector components and performance arbitrarily close to the ultimate rate-distortion can be achieved by VQ if the vector dimension is high enough [1].

Nevertheless, the exponential growth in complexity forces the use of low-dimensionality VQ in practical systems. In this case some kind of adaptation is necessary to obtain adequate performance, specially when we deal with non-stationary inputs as in speech coding.

Several kinds of adaptive VQ-based coding schemes can be found in the literature. For example, the popular CELP [2] can be seen as an adaptive vector quantizer whose codewords are frame to frame adapted to the input statistics (autocorrelation) by linear filtering.

In the work we describe here we have followed a different approach in which the adaptation algorithm is directly derived from the VQ design techniques. The resulting scheme will be useful when some local stationarities are expected in the input vector statistics. In this communication we explore its application to quantize the LPC parameters, Video coding is another area in which this adaptive quantizer has been shown to be useful [10].

It is well known that, for a given bit rate, a speaker dependent codebook, i.e. designed for the specific speaker whose speech is being coded, would work better than a speaker-independent codebook. Therefore, and adaptive quantizer is also expected to outperform a non-adaptive one when it is used to quantize the LPC parameters. This scheme would also be able to adapt to other variations in the speech spectrum as those due to the acoustic environment of the speaker and to the recording conditions, and the result will be an increase both in performance and in robustness.

The adaptive vector quantization of the LPC parameters have been also studied by other authors. In [3] a system that changed the codewords in time is described. Nevertheless this change creates the necessity of transmitting the new vectors to the receiver with a significant increase in bit-rate. In the scheme that we present the multistage VQ structure [4] is shown to allow the redesign of the quantizer of one stage using the information given by the rest of the stages, therefore, no additional bit transmission is needed.

In [5] vector linear prediction is used to make effective use of the considerable redundancy between different speech frames within one phoneme. Our approach can also be viewed as a vector predictor of reduced complexity. Furthermore the quantizer is expected to follow not only the local stationarities but also the long term (speaker) statistics of the input vectors.

In the following section of the paper we will first present the adaptive multistage vector quantization (AMSVQ) scheme. This scheme was previously reported in [6] where the adaptation algorithm was derived as an LMS type minimization. In this paper, we extend our preliminary report and show the close relation between the proposed algorithm and the VQ design techniques known as LBG [7] and Kohonen's [8].

In the last part of the paper the developed scheme is applied to the quantize the LPC parameters. We will first compare the performance of the adaptive with the non-adaptive multistage structure and the full-search VQ. Then this quantized LPC parameters are used in the CELP and the Multipulse coder and both the SNR results and the subjective quality is discussed.

2. DESCRIPTION

Multistage vector quantization (MSVQ) has always been seen as a suboptimal VQ scheme with reduced complexity and storage. It consists of successively approximating the input vector in several cascaded VQ stages, where the input vector from each stage is the quantization error from the preceding stage. In [4] the MSVQ is applied to the quantization of the LPC parameters, and it is shown that, in the case of Euclidean distance measures such as the log-area ratio, that quantizer is very close to a theoretically predicted asymptotically optimal rate distortion relationship.

In the scheme we present the multistage structure has been used to develop a continuously adaptive VQ. The objective of the adaptive algorithm is to update the last used codevector $\mathbf{c}_i$ in order to minimize the exponentially weighted mse E_i defined as

$$E = \sum_{j=0}^{\infty} \beta^j \, ||\mathbf{x}_i(n-j)-\mathbf{c}_i(n+1)||^2 \qquad (1)$$

where $\mathbf{x}_i(n-j)$ are the input vectors previously quantized by $\mathbf{c}_i$. The minimization of (1) gives

$$\mathbf{c}_i(n+1) = \frac{\sum_{j=0}^{\infty} \beta^j \, \mathbf{x}_i(n-j)}{\sum_{j=0}^{\infty} \beta^j}$$

$$\mathbf{c}_i(n+1) = (1-\beta) \left[\sum_{j=1}^{\infty} \beta^j \, \mathbf{x}_i(n-j) + \mathbf{x}_i(n) \right] \qquad (2)$$

Note that $\mathbf{x}_i(n)$ is the last input vector $\mathbf{x}(n)$ and the infinite sum can be identified as the previous $\mathbf{c}_i$, so we can express the above equation as

$$\mathbf{c}_i(n+1) = \beta \, \mathbf{c}_i(n) + (1-\beta) \, \mathbf{x}(n)$$

$$\mathbf{c}_i(n+1) = \mathbf{c}_i(n) + (1-\beta) \, (\mathbf{x}(n) - \mathbf{c}_i(n)) \qquad (3)$$

or

$$\mathbf{c}_i(n+1) = \mathbf{c}_i(n) + (1-\beta) \, \mathbf{e}(n) \qquad (4)$$

where $\mathbf{e}(n)$ is the quantization error.

This update equation is the well known Kohonen's algorithm, used to design what he calls self-organizing feature maps [8], i. e., vector quantizers. This algorithm, as it appears in (4), is not useful for adapting the quantizer as it is used because the receiver does not known the value of the quantization error. Nevertheless in a MSVQ the error of one stage is quantized by the following stages, therefore, this quantized error is available at the decoder and is used instead of the actual error. The resulting updating equation is then

$$\mathbf{c}_i(n+1) = \mathbf{c}_i(n) + \mu \mathbf{e}_q(n) \qquad (5)$$

where $\mu=(1-\beta)$ is the step size and $\mathbf{e}_q(n)$ is, in general, the sum of the outputs of the following stages. In the most simple case, (2 stages), the VQ equations are

$$\mathbf{c}_i = q_1(\mathbf{x}) \qquad (6)$$

$$\mathbf{e}_q = q_2(\mathbf{x} - \mathbf{c}_i) \qquad (7)$$

$$\mathbf{z} = \mathbf{c}_i + \mathbf{e}_q \qquad (8)$$

where $\mathbf{x}$ is the input vector, $\mathbf{z}$ the quantized vector, $\mathbf{c}_i$ the output of the first quantizer and $\mathbf{e}_q$ the contribution of the second codebook (Fig. 1).

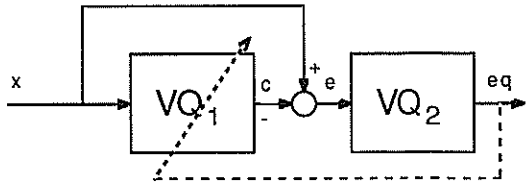


Figure 1. AMSVQ system with 2 stages

Then $\mathbf{e}_q$, that is an estimation of the quantization error of the first codebook, is used to adapt the first quantizer as equation (5) indicates.

As it is shown in [6] this algorithm can also be obtained applying a LMS-type minimization algorithm to the error at the output of the first codebook. In our simulations we only adapted the first stage that was a full-search codebook. Nevertheless, a similar algorithm can be derived for gain-shape vector quantizers and other kind of structures as the tree-searched vector quantizer [9].

Although robust to variations in the signal statistics the above adaptation make the quantizer more sensitive to channel errors that the conventional multistage structure. This problem can be reduced using a leakage factor similar to that used in the

LMS-type algorithms of practical ADPCM schemes. Then (5) is modified to give:

$$\mathbf{c}_i(n+1) = (1-\gamma)\,\mathbf{c}_i(n) + \mu \mathbf{e}_q(n) + \gamma\,\mathbf{c}_{i0}(n) \qquad (10)$$

where γ is a small value that control the memory of the system and $\mathbf{c}_{i0}$ is the initial (design) value of the codevector $\mathbf{c}_i$.

3. CODEBOOK DESIGN

The codebook design approach is similar to that used in conventional multistage VQ [1]. We need a representative speech training set from different speakers, but thanks to the adaptive nature of the AMSVQ system the results are not expected to be very conditioned by the election of the training sequence.

Starting with the first codebook and using the LBG clustering algorithm [7], the codebooks are constructed in succession. The training sequence for the second codebook is obtained as the first codebook quantization error. The problem is that , to obtain this quantization error, we need the output of the second codebook to apply the adaptation algorithm (5) to the first codebook. As $\mathbf{e}_q$ is not available to obtain this training sequence, the actual quantization error $\mathbf{e}$ is used to adapt the first codebook (equation (4)). Further details can be found in [6].

4. RESULTS

The AMSVQ system has been applied to the quantization of LPC parameters with important improvements respect to the conventional multistage structure. Quantization of the LPC coefficients has been extensively studied. Generally, the inverse sines of the reflection coefficients or the log-area ratio values are quantized. We chose the log-area ratio values with the mse distortion for our experiments. Therefore, we will define the LAR-SNR as

$$\text{LAR-SNR} = \frac{\sum_{m=1}^{M}\sum_{i=1}^{N} v_i^2(m)}{\sum_{m=1}^{M}\sum_{i=1}^{N} (v_i(m)-v_{qi}(m))^2}$$

where v_i is the i-th log-area ratio.

Vector quantizers have proven to be very efficient in encoding the predictor parameters. Nevertheless, for high quality coding, 24-30 bits are required and a full search codebook become impractical. Structured codebooks, as the multistage, must be used to reduce the complexity. The developed adaptation algorithm makes the suboptimal multistage structure efficient and robust against different speakers, languages and environments.

To design the quantizers a training speech sequence formed by 80 sentences of 7 female and 7 male speakers was used. And to test the quantizers we chose 24 sentences of a different male speaker (not included in the training sequence). The silent segments (background noise) were not processed.

The speech signal was not preemphasized and the correlation method with a Hamming window of 200 samples (25 ms) was used to obtain ten log area parameters every 160 samples (20 ms).

4.1. High-rate quantization

The first results we present have been obtained using a codebook of 8 codewords (3 bits/frame) in the first stage and 3 codebooks of 256 codewords (8 bits/frame) in the second, third and fourth stage. The number of bits for each frame is thus 27 (3+8+8+8) and at 50 frames/s, the bit rate equals 1350 bps. To obtain a fast adaptation and robustness against channel errors, only the 8 codewords of the first quantizer were adapted.

First of all, we studied the performance of a single codebook of 8 codewords (3 bits) when it was adapted using the actual error $\mathbf{e}$ (forward adaptation). The results showed an increase in signal to noise ratio (LAR-SNR) of more of 2 dB for values of the step size between 0.1 and 1, and local maximum were observed near μ=0.1 and μ=0.6, so these values and μ=0 were used to design three different sets of codebooks.

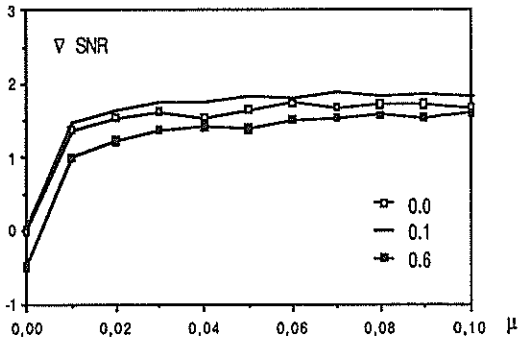


Figure 2. Increase in SNR for the three designed AMSVQ

Then, the designed codebooks have been used to quantize the test sequence with different values of the step size μ. Fig. 2 illustrates the increment in LAR-SNR respect to the 19 dB of the conventional (μ=0), multistage vector quantizer. The best results

are now obtained with μ=0.4 and the AMSVQ system designed for μ=0.1. Nevertheless smaller values of the step size as μ=0.1 or μ=0.01 give also a significant increase of the performance and can be a good choice for providing robustness against channel errors.

4.2. Low-rate quantization

The aim of this experiment was to compare the full-search VQ with the AMSVQ. Due to the complexity of the full-search scheme only 10 bits per frame were used. We tried different distributions of bits between the first and the second stage. Fig. 3 compares the LAR-SNR obtained with each scheme and in table I the complexity of the full-search and the best AMSVQ schemes is also shown .

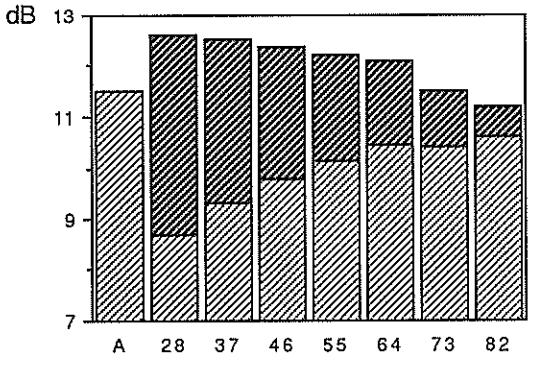


Figure 3. LAR-SNR for some AMSVQ schemes with different bit distributions and the Full-search VQ (A).

Scheme	bits/frame	Comp. Cost	LAR-SNR
VQ	10	1024	11,52
AMSVQ	3+7	136	12,53
AMSVQ	2+8	260	12,62

Table I. Computational cost and LAR-SNR for some 10 bits VQ schemes.

4.3. Speech coding

The developed AMSVQ of the LPC parameters has been integrated in a CELP and a MultiPulse coder. As, it was expected, the results show that an increase in LAR-SNR gives also an increase in the global performance of the coder and both speech quality and SNR is improved (Table II).

bits/frame	CELP	Multipulse
oo	10,57	14,50
AMSVQ-27	10,00	13,30
MSVQ-27	9,65	12,90
AMSVQ-19	9,38	12,80
MSVQ-19	9,15	12,20

Table II. SNR for different coders and some LPC VQ schemes

5. CONCLUSIONS

The application of the developed adaptation algorithm to the multiple stage vector quantizer allows to increase the performance and reduce the complexity of previous VQ-based schemes. This algorithm is very simple and requires no additional bits. The quantizer is continuously redesigned and it is less sensible to the chosen training sequence. Nevertheless, one of the most important results is the increase in robustness across different speakers, languages and environments that the adaptation algorithm provides.

The results show that the AMSVQ increases the performance of conventional multistage quantizers and can be a good choice for coding the LPC coefficients with high quality.

REFERENCES

[1] J. Makhoul, S. Roucos, and H. Gish. "Vector Quantization in Speech Coding". Proc IEEE, vol. 73, November 1985.

[2] M.R. Schroeder, and B.S. Atal. "Code-excited linear prediction (CELP): High-quality Speech at very low bit rates". Proc. ICASSP, Vol. 3, pp 941-944, March 1985.

[3] D.B. Paul. "An 800 bps adaptive vector quantization vocoder using a perceptual distance measure". IEEE Proc. ICASSP, Boston, April 1983.

[4] B.H. Juang, A.H. Gray Jr. "Multiple stage Vector Quantization for Speech Coding". IEEE Proc. ICASSP, pp 597-600, November 1982.

[5] M. Young, G. Davidson, and A. Gersho. "Encoding of LPC Spectral Parameters using Switched-Adaptive Interframe Vector Predict.". IEEE Proc. ICASSP, pp 402-405, April 1988.

[6] J.A. Rodríguez-Fonollosa. "Adaptive Multistage Vector Quantization". Proc. MELECON, pp 225-228, Lisboa, April 1989.

[7] Y. Linde, A. Buzo, and R.M. Gray. "An Algorithm for Vector Quantizer Design". IEEE Trans on ASSP, vol. 34, no. 4, August 1986.

[8] T. Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, Berlin 1988.

[9] J.A. Rodríguez-Fonollosa. "Cuantificación Vectorial Adaptativa aplicada a la codificación de voz". PhD dissertation, DTSC-UPC, Barcelona, Spain, July 1989.

[10] J.R. Fonollosa, and J.A. Rodríguez-Fonollosa. "Parallel Adaptive Multistage Vector Quantization for digital Video compression". Proc. EUSIPCO'90.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

6.5 kbps Self-Excited/Code-Excited Linear Prediction Speech Coder

H. B. Hansen, H. Nielsen
Telecommunications Research Laboratory
DK-2970 Hørsholm, Denmark.

Y. Wu, J. Aa. Sørensen
Electronics Institute, Technical University of Denmark
DK-2800 Lyngby, Denmark.

Abstract: *This paper describes a combined self-excited/code-excited linear prediction speech coder with time-varying bit allocation based on non-fatal voiced/unvoiced classification of the speech frames. The coder configuration and the bit allocations are chosen so as to comply with the dynamic properties of speech. The performance of the classification unit is demonstrated, and a new codebook generation method is presented.*

1 Introduction

Several speech coding principles have been developed for medium-to-low bit rates, but for most of them speech quality degrades rapidly below 8 kbps. At low bit rates around 4 kbps the conventional fixed bit allocation schemes are unable of ensuring that all parts of the speech information are always represented with sufficient accuracy to produce high-quality synthetic speech. To retain high quality, when the bit rate is lowered, it becomes more important that bit allocation reflects the dynamic properties of speech.

The CELP principle [1] is quite successful in providing high quality at low bit rates yet at the expense of a heavy computational load and slightly degraded quality of vowel sounds. On the other hand self-excitation (SE) [2] has shown to be effective in voiced segments, where recent excitations are likely to reappear.

In order to alleviate the problems with the CELP and to exploit speech dynamics we propose a combined approach incorporating time-varying bit allocation based on voiced/unvoiced decisions. The computational complexity is reduced by designing specialized codebooks and by using recursive codebook search procedures [3].

2 Coder Description

The main principle of the speech coder is that of a CELP with an SE long-term prediction section implemented by means of an adaptive codebook. The speech signal is segmented into frames of 20 ms length, and each frame is classified as being either voiced or unvoiced. The function of the SE-section is controlled by the voicing classification, and the voicing status also forms the basis for changing the bit allocation. The coding strategies are outlined in Figure 1. Clearly, it is assumed that voiced speech is highly predictable and stationary and that for unvoiced speech these properties are less pronounced.

	voiced	unvoiced
LPC order	10	6
self-excitation	yes	no
subframe length	5 ms	2.5 ms

Figure 1: *Functional differences between coding strategies.*

Figure 2 shows the functional block diagram (LPC analysis, subsegmentation and quantization blocks are omitted). During voiced speech the SE-section is used to remove periodicity, and the residual signal is vector quantized by the CELP with a small fixed codebook, which is optimized to handle voiced residuals only. In unvoiced segments the CELP does the job by itself, and as the prediction gain of unvoiced segments is very poor the model order is reduced and most of the bits are allocated to the excitation. A separate codebook containing unvoiced excitation sequences is now employed and the subframe length is reduced, because some important features of unvoiced speech are known to be of very

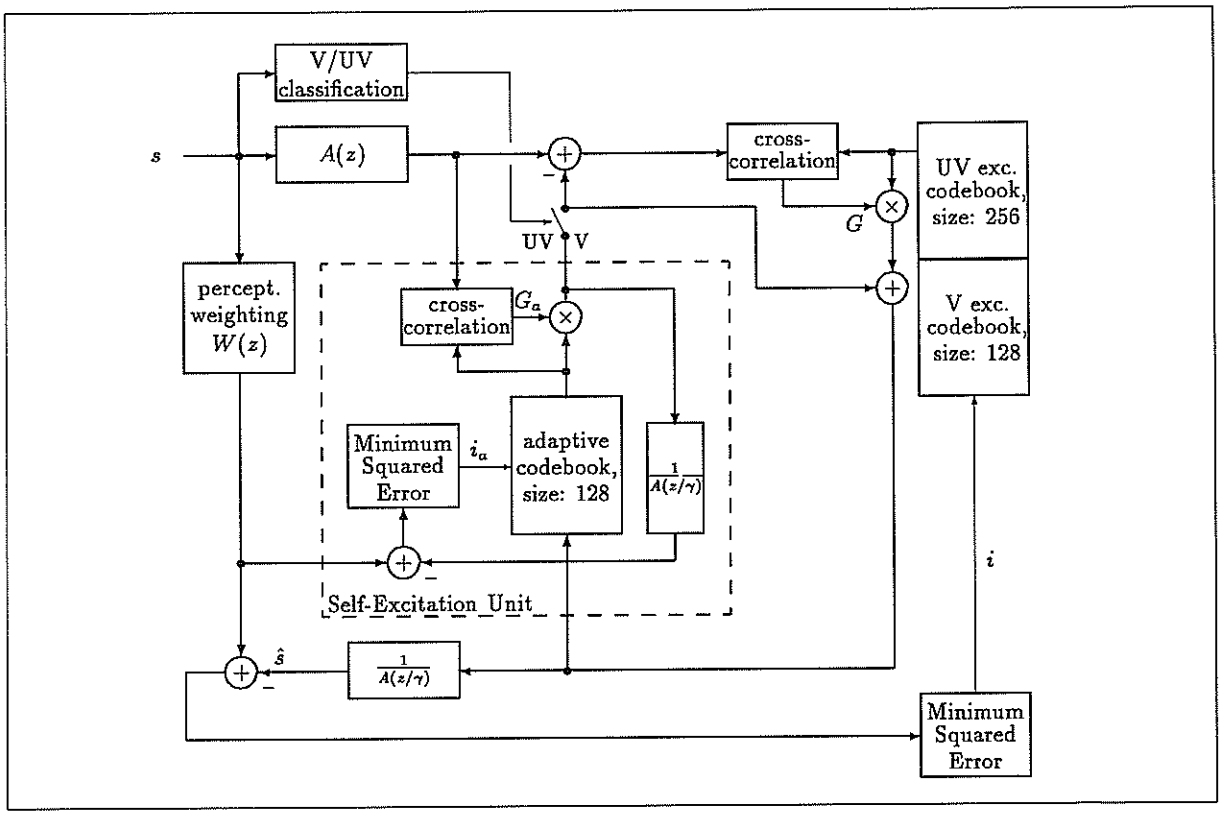


Figure 2: *Self-Excited/Code-Excited Linear Prediction speech coder.*

short duration. The use of small specialized codebooks enhances speech quality as well as computational efficiency compared to the original 1024-CELP with standard long-term prediction. The LPC parameters are obtained using the autocorrelation method [4], and the speech segment is selected by a Hamming window of 25 ms length, thus overlapping the last 5 ms of the previous frame. The LPC coefficients are transformed into Log-Area-Ratios (LARs), which are attractive when the system order is not fixed. Besides the LARs the voicing status bit and the index and gain parameters of the codebooks are transmitted.

It goes without saying that certain precautions need be taken, when hard decisions like voiced/unvoiced are introduced. Furthermore, the positioning accuracy of the voicing transitions is limited because of the fixed framelength. The remedy for this deficiency is trimming the classification unit by adjusting its thresholds, retaining the unvoiced state until fairly strong indications for voicing are recognized, and finally, always keeping the adaptive codebook updated, even in unvoiced segments where the SE-section is not used. In the start-up phase where the adaptive codebook is empty, status is forced to unvoiced allowing the codebook to be updated prior to usage. In addition to this, the LPC parameters are carefully interpolated such as to smooth away voicing transitions as well as ordinary frame transitions. The unvoiced excitation codebook may consist of random vectors, but as it will have to deal with some voiced onset residuals optimization might be necessary. For the voiced excitation codebook, which is used for voiced residuals exclusively, optimization is mandatory.

3 Time-varying Bit Allocation

Throughout the literature it is often pointed out, that time-varying bit allocation could be used to enhance speech quality [5, 6]. The major problem, however, still seems to be classification, because it is difficult to discriminate more than a few classes reliably when no restrictions can be imposed on the speech material. We therefore chose only two classes: voiced (V) and un-

voiced (UV) speech, and as voicing is a very prominent feature it forms a good basis for employing time-varying bit allocation.

3.1 V/UV Classification

The V/UV classification is performed by investigating the autocorrelation function, which is also calculated for the LPC analysis. As shown on Figure 3, the autocorrelation function is normalized with respect to frame

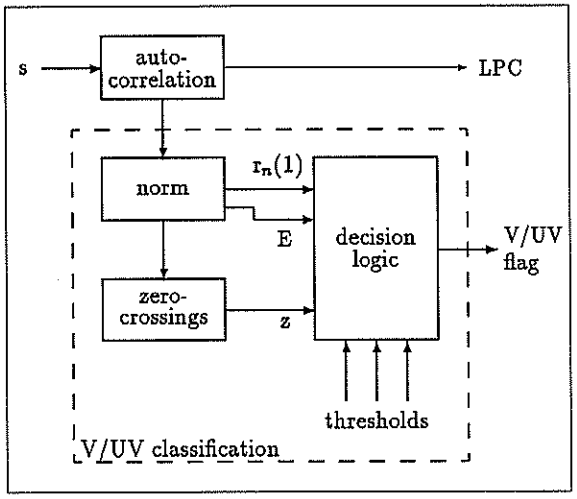


Figure 3: *V/UV classification unit.*

energy, and the number of zero-crossings is determined. Then the frame energy E, the unit delay correlation coefficient $r_n(1)$ and the zero-crossing count z are passed to a very simple decision algorithm, which compares the values to the thresholds and assembles the results for the final decision. Despite its simplicity the classification unit performs well, and some lack of precision is not disastrous when the codebooks are optimized properly.

3.2 Dynamic Bit Allocation

Quantization tables for all parameters and gain factors were obtained by statistical analysis carried out separately in the voiced and unvoiced cases. Examples of results typical of LAR coefficients 1–6 are given in Figure 4. In general, the density peaks are found at different values and the voiced coefficients show lower variance, while for the higher-order coefficients only small differences remain. Utilizing these statistics non-uniform quantizers were designed for the LAR coefficients and the adaptive codebook gain. For the fixed codebook gain, which has a nice symmetrical and single-peaked density, a logarithmic design was chosen.

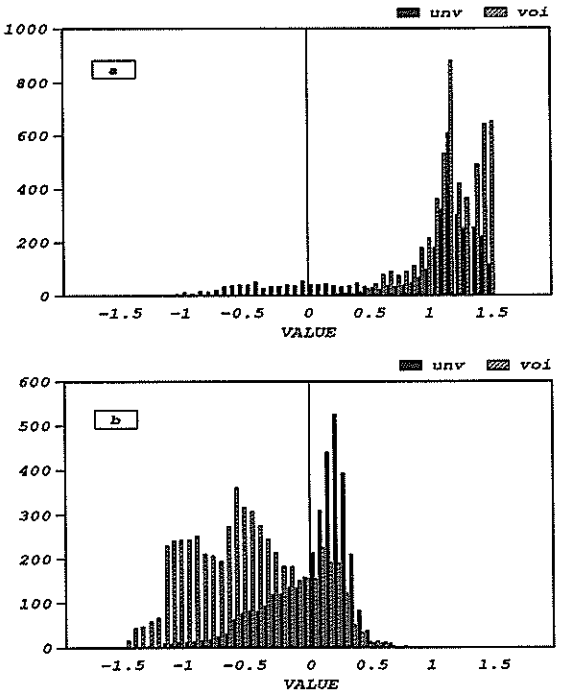


Figure 4: *Histograms comparing voiced and unvoiced states of a) LAR1 and b) LAR2.*

The parameters to be transmitted are encoded in accordance with voicing status as shown in Figure 5. Keeping the GSM half-rate application in mind, this bit allocation leaves 4.9 kbps for the error correction system, in relative terms almost the same amount as is allotted in the full-rate standard.

As an experiment, the LPC parameters were coded using the line-spectrum-pair representation [7], and it was found that speech quality was unaffected by these changes.

4 Codebook Generation

The adaptive codebook contains overlapping code vectors and the virtual search procedure [3] is used. This codebook structure is applicable to any codebook which is not subject to training, and it enables the use of recursive search procedures. As both fixed codebooks in this configuration are trained, the code vectors in these codebooks are stored separately.

After having tried some conventional codebook training algorithms like LBG [8] with rather disappointing

	voiced	unvoiced
LPC	41	25
SEV	$4 \cdot (7+4) = 44$	0
CELP	$4 \cdot (7+4) = 44$	$8 \cdot (8+5) = 104$
V/UV	1	1
bits/frame	130	130
bit rate (20ms frame): 6500 bps		

Figure 5: *6.5 kbps bit allocation. Parentheses contain vector quantizer bits per subframe for index and gain, respectively.*

results, we decided to develop a new method based on statistical selection. Referring to Figure 2 the method can be described in four steps:

1. Initially, the codebooks are filled with gaussian vectors.

2. Two large pools of residual vectors are generated by simulating the analysis part of the coder, extracting the inverse filter output signal for the unvoiced residual pool and the difference between this signal and the self-excitation for the voiced residual pool.

3. The coder is rerun with the pools substituted for the codebooks in order to test the vectors on different speech inputs. During the test indices of the optimal code vectors are written to file.

4. For each pool the index file is analysed statistically, the resulting histogram is sorted and a number of the most frequent vectors corresponding to the desired codebook size are copied from the pool to constitute the new codebook.

Usually, for practical reasons the size of the pools must be limited. This is done by applying the mean-squared vector distortion measure in the extraction phase such that a new vector is added to the pool only if the distance to its nearest neighbour in the codebook exceeds a threshold; otherwise the vector is discarded. A fixed threshold can be used if the vector is normalized prior to each comparison; and before a vector is placed in the pool it is scaled in order to obtain constant power level of the resulting codebook in compliance with the gain factor quantization in the codec.

5 Conclusion

In this paper a speech coder configuration combining two well-known principles, code-excited and self-excited linear prediction, is presented. This combination is suited for employment of time-varying bit allocation based on voiced/unvoiced classifications.

It is demonstrated that even a very simple classification unit is capable of providing useful information for the benefit of quantization efficiency. Informal listening tests have confirmed that hard voiced/unvoiced decisions, which make abrupt changes in bit allocation, cause no audible impairment of the reconstructed speech when the basic requirements mentioned in Section 2 are met.

The codebook generation method differs from LBG in that no averaging is performed. Like all other codebook training procedures this one requires a very long training sequence, which makes especially the vector test phase very time consuming, but with these codebooks speech quality is improved as compared to what is obtainable with LBG codebooks. Note that if the second stage is left out, the procedure could be used to optimize stochastic codebooks as well.

References

[1] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *ICASSP'85*, pp. 937–940, IEEE, 1985.

[2] R. C. Rose and T. P. Barnwell III, "The self excited vocoder - an alternate approach to toll quality at 4800 bps," in *ICASSP'86*, pp. 453–456, IEEE, 1986.

[3] W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, "An efficient stochastically excited linear predictive coding algorithm for high quality low bit rate transmission of speech," *Speech Communication*, vol. 7, pp. 305–316, 1988.

[4] L. R. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[5] S. Wang and A. Gersho, "Phonetically-based vector excitation coding of speech at 3.6 kbps," in *ICASSP'89*, pp. 49–52, IEEE, 1989.

[6] N. S. Jayant and J.-H. Chen, "Speech coding with time-varying bit allocations to excitation and LPC parameters," in *ICASSP'89*, pp. 65–68, IEEE, 1989.

[7] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *ICASSP'84*, IEEE, 1984.

[8] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. COM-28, pp. 84–95, Jan. 1980.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A full hand-free radiotelephone with vocal dialing*

C. Baillargeat, J. Boudy, I. Lecomte, L. Lelievre, A. Baron,
C.Parment, P. Lockwood, A. Gilloire*

MATRA-Communication
Speech Processing Group
Rue J.P. Timbaud, B.P. 26
78 392 Bois d'Arcy FRANCE

* CNET-LAA/TSS/CMC
route de Trégastel, B.P. 40
22301 Lannion FRANCE

With a full hand-free radiotelephone with vocal dialing in a car, the speaker can concentrate completely and safely on his driving task while composing and talking with a desired suscriber. Nevertheless some drawbacks arise: first, acoustical echoes and ambiante car noise are perceived by the far-end suscriber, and on the other side, for the car driver, dialing by speech recognition becomes less efficient and reliable in a noisy environment. This paper describes some methods for reducing acoustical echoes and ambiante car noise in transmission and vocal dialing schemes. Then, a hardware implementation of the full hand-free and vocal dialing functions has been performed in a prototype connected with a usual radiotelephone.

1. Introduction

Achieving a good quality of speech transmission and reliable performance speech recognition in the car is an important challenge specially in the context of mobile telephony applications where the user can access the telephone functions by voice. The break through such technology is appealing, since the driver can concentrate completely and safely on his task while composing and conversing in a full handfree mode.

For full hand-free radiocommunications, when the speaker is phoning while driving, the far-end listener is disturbed by the ambiante car noise and by the echoes of its own speech. Various noise signal processing and LMS-echo cancellation techniques have been proposed during the last ten years to improve the speech quality transmission in a car [1]. Noise reduction techniques are generally based on mono-sensor frequential approaches like Wiener filtering and derivatives techniques [2][3] or temporal approaches like Sambur's method [4]. Recently Kim and Un [5] proposed a mono-sensor speech enhancement technic for colored wideband noise contexts: their approach, based on linear prediction filtering, affords a weak complexity and it can be easily implemented on DSP hardware.

Voice access to an agenda by uttering a single word (typically the name of the desired suscriber) avoids a time consuming and cumbersome dialing phase. Dialing necessitates reliable performance at the string level which implies a very high recognition score at the utterance level. This might be somewhat contradictory with the expected use of the system in general conditions where the signal to noise ratio (SNR) is close to 0 db. A 0db SNR was measured at 130 km/h in a relatively high range car, using a hand free microphone.. This type of condition will unfortunately become mostly common in the near future, especially if the telephone is used in medium to low range cars.

In an agenda-driven access mode, a speaker dependent approach must be used. In this case the system has to deal with the mismatch condition problem, in the sense that training is made in a quiet and unstressful environment while the recognition is made in completely different conditions. The mismatch is not only due to noise that will corrupt the test signal, but also to the increased variability of pronounciation caused by various factors such as stress, tireness or simply the noise level (Lombard effect). In our system recognition phase is performed by the DTW algorithm: some robust-noise weighting [7] of euclidian or projection distances were trained and retained for the hardware implementation.

This paper adresses on one side the approaches used for echo cancellation and noise reduction, and on an another side the problem of speaker-dependent discrete utterance recognition. Section 2 gives a description of the recording conditions as well as the database used for the experiments. Section 3 describes the LMS-based echo cancellation technic and the Kim's noise reduction approach [5] based on temporal linear filtering.The speech enhancers described hereafter are based on the use of a single hand-free microphone.

Section 4 presents the speaker-dependent recognition stage based on the dynamic time warping technique (DTW) with robust weighting schemes. Finally, Section 5 is devoted to the description of the hardware hand-free and vocal dialing functions implementation.

2. Context

The experiments have been carried out in one high-range vehicle. The acoustical characteristics have been analysed in various conditions: several speeds (O,70,90,110,130 km/h), several types of road (urban areas, highways, open country roads). The recordings have been made using a SONY PCM system. The signal was then downsampled to 8 kHz. An unidirectional microphone was used and positioned on the left of the windscreen. Preliminary experiments showed that this position appeared to have the best SNR ratios. Four speakers, two males and two females have been used for the tests. The vocabulary is an agenda of thirty french family names words .

* This work was supported by CNET-FRANCE TELECOM (contract n°87.35.111)

3. Hand-free radiotelephone functions for transmisssion

The hand-free radiotelephone is constituted of a microphone and a loud-speaker installed in the car cockpit instead of a classical hand-set. Therefore, during a radiocommunication, the far-end talker's voice is reflected inside the car and is picked up by the microphone. In that case, for the far-end talker, an echo is heard and the conversation is disturbed. Moreover, Larsen effects may occur between the loud-speaker and the microphone inside the car.

In order to study this problem, some recordings have been made in car with an unidirectional microphone and two speakers. For several microphone and loud-speaker positions, microphone SNR and acoustical coupling evaluations were computed. The best places, retained after measurements, were on the left of the windscreen for the microphone and under the dashboard for the loud-speaker.

In order to avoid Larsen effects, to suppress the echo, and to achieve the noise reduction problem, we propose two techniques: first the LMS-echo cancelling technique, then a speech enhancement technique introduced by Kim and Un [5].

3.1 Echo cancellation algorithm

In the hand-free telephone systems, acoustic echo is a very annoying problem . The best way to handle it appears to be echo cancellation. This technique is developped here and implemented with a digital adaptive filtering. The acoustic echo canceller first gives an adaptive estimation of the echo path impulse response using the far-end talker speech reflected inside the car and then subtracts the echo component from the signal picked up by the microphone.

The block diagram of an echo cancellation system is shown in figure 1.

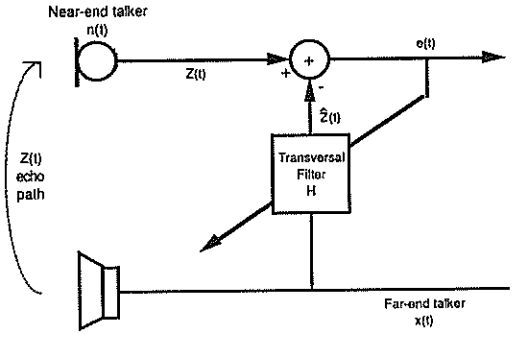


Figure 1 : Adaptive Acoustic Echo Cancelling Block Diagram

Let be $H=\{h_0,h_1,.....,h_{N-1}\}$ the vector of the N coefficients $h_i(n)$ of the filter and $X=\{x(n),x(n-1),....,x(n-N+1)\}$the vector of the N most recent input signal samples. These coefficients $h_i(n)$ are adjusted to minimize the mean square error MSE defined as,

$$(1) \qquad MSE = E((\, z(n) - \hat{z}(n)\,)^2)$$

with the following equations,

$$(2) \qquad \hat{z}(n) = \sum_{i=1}^{N} h_i(n)\; x(n-i) = H^T . X$$

T denotes the transpose of the vector.

and the filter H is updated according to the LMS algorithm of Widrow [9] with a normalization by the short-time input signal energy.

This method can eliminate the echo without cutting the voice paths and makes comfortable conversation instead of the half-duplex technique. Nevertheless, we encountered several difficulties which will be considered below.

First, the real-time implementation of the echo path is difficult because of the filter lengh which mainly determines the computational complexity of the algorithm. In a car application, it requires several hundred taps to achieve a significant reduction of the echo signals level, since the impulse response of the acoustic echo path is several tens of milliseconds long.

On an another side, the acoustic echo path is a nonstationary channel because of the possible driver's motions: in that case a step-size optimization of the LMS algorithm is necessary. Moreover, the LMS algoritm is sensitive to the noise which will reduce the performances of the echo canceller.

An other problem arises if the driver talks simultaneously with the far-end talker (double talk: DT). In general cases, for adaptive echo cancellers, occurence of DT necessitates an accurate endpoint detection and freezing of the filter coefficients. Specifically, in the simultaneous talk between the near-end and far-end speakers, the echo canceller has to freeze adaptive update.

Nevertheless, the two first problems are less critical in a car than in a teleconferencing room [10] since dimensions, people motions and wall absorption are different. On the other hand, it is very difficult to detect DT. Speech detection, in itself, is a difficult problem in speech recognition or noise enhancement. Some standard techniques of speech activity detection in echo cancellation applications exist in the bibliography as proposed by [10].

Simulation results

A simulation was done on a SUN-4 workstation which modeled the system shown in figure 1. The amount of echo cancellation is measured by the Echo Return Loss Enhancement (ERLE) which is defined by a ratio of the average power of pure echo zc(n) to the average power of residual error ec(n) as follows:

$$(3)\ ERLE(dB) = 10\ \log_{10}\ \{E\,[(zc(n)^2]\ /\ E\,[ec(n)^2]\}$$

The simulation results are illustrated in the following two figures 2 a-b. When the car engine is off (fig 2a), an ERLE of 23 dB is obtained, and when a speed of 90 Km/h is attained on a road with medium roughness (fig 2b), an ERLE of 10 dB is obtained.With this technique, we can have quite good results for small and medium speeds but when the car speed reaches 130 Km/h, the ambiante car noise is so high that the echo signals are masked and efficient echo cancellation is no more necessary.

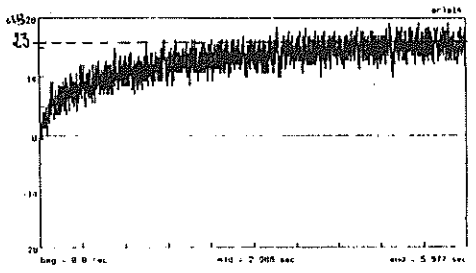


fig 2 (a) : ERLE without car noise (car engine off)

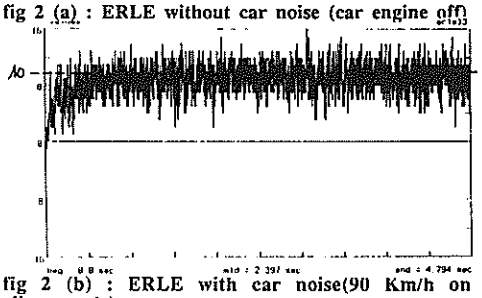


fig 2 (b) : ERLE with car noise(90 Km/h on medium road)

3.2 Speech enhancement : Kim's noise approach

From a signal processing point of view, the sources of noise can be classified in two groups: low frequency narrowband noises, essentially due to mechanical sources (engines, tyres...), and flat spectrum noises produced by aerodynamic disturbances.

The global structure of the system, used here for enhancing speech corrupted by car noise, is characterized by a filtering block (Forward/Backward adaptive digital filtering F/B ADF) driven by a speech detector stage shown in figure 3: noisy speech signals are processed by a single sensor and the difficulty is to find a temporal filtering technique which can separate noise contributions from speech signals. In many cases of mono-sensor noise suppression approaches, the discrimation between silence and speech is required. In our case, if the speech detector detects speech, the on/off switch activates the filtering block to adjust its parameters. The adjustment is stopped during speech pause. The enhanced speech is the filter output signal.This method requires an accurate endpoint detection, which may be a difficult task in noisy environments.

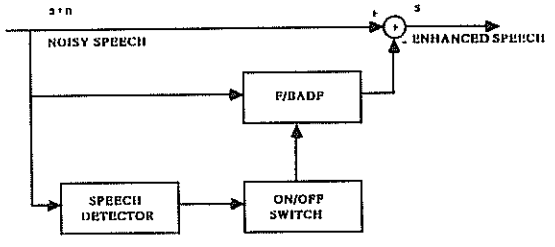


fig 3 : Enhancement of noisy speech with F/B ADF

The forward/backward adaptive digital filtering method (F/B ADF) is studied for the noisy speech enhancement. The enhancement algorithm uses both past and future samples to estimate the current sample. A block diagram and a description of this technique is presented in [5].

Simulation results and discussion

The simulations, done on a SUN_4 workstation, allowed us to compare different results obtained with various order values (N=4, 8, 16) for the four speakers of the database. Segmental filter output SNR have not been improved with regard to input noisy signals SNR. Only subjective improvements have been perceived by listening tests. Better subjective results were obtained for an order value of 8 and for female speakers: in fact, the F/BADF filter has been observed to have some low-frequencies cutting effects.

4. Speaker-dependent recognition

Our recognition system is based on a speaker-dependent approach performed with the classical DTW algorithm. It is well known that speaker-dependent isolated word recognition rates fall when the noise -i.e. car speed- increases: from our simulations recognition scores drastically decrease -from 20% to 40%- when the speed becomes higher than 90 km/h (56 mph).

In the DTW recognition context, the "usual" Euclidian distance is commonly used to compare short time speech cepstral representations. It has been shown that this distance is not optimal, and that weightings like the bandpass lifter or exponential lifters improve significantly the performance in clean conditions [7]: from simulations performed in real noisy conditions of the database, we obtained recognition improvements by 13% to 20% for car speeds higher than 90 km/h.

The cepstral projection technique had given encouraging results in distances values comparisons [6]. On the other side, the global recognition scores (averaged from all the speeds conditions) with the weighted projection have not been improved compared to weighted Euclidian distances, but it was interesting to note that weighting the projection distances can increase the recognition scores by 13%. The increase is only of 8% for the weighting of the Euclidian distances.

Short-time modified coherence representations [8] with euclidian distances have not given better results.

5. Hardware implementation

5.1 A full hand-free radio-telephone system

5.1.1 Presentation

The prototype developped (called "RTML à commande vocales") supports the implementation of the algoritmics described above and operates in real time. It is directly connectable to the RADIOCOM 2000 Radio-Telephone ("alpha-numérique" model).

This product allows hand-free communication and dialing by speech recognition. Digital data exchanges between radio-telephone and prototype require an I2C bus. Two other connections are used for the analog data. This prototype must work in adverse environment with electrical, mechanical, climatic and electromagnetic constraints. The board is installed in a metallic box, the volume size is 30cm *20cm * 7cm.

5.1.2 Prototype architecture

The prototype is built around a DSP32C (a 32 bit floating point digital signal processor of ATT, which perfoms 12.5

million instructions per second). The memory map is made of 32k words of RAM (35ns access time) and 64k words of EPROM (access time 150ns) where are stored all the programs which will be transfered to the RAM for their executions. The user's vocal prints are saved in 32k words of FLASH EPROM (electrically erasable memories, 150ns acces time).
The DSP 56200 (echo cancelling chip with adaptive filter) is used as a peripheral of the DSP32C. The parallel data bus 0 to 7 is connected to a micro-controller (PCB 80C652) data bus which serves as an interface with an I2C bus and a terminal (PC portable...). This controller addresses 64k octet of EPROM and 32k octet of RAM, it drives LEDS and switches which program all the mode of operation desired. Data can be directly loaded from the serial port of the controller into the DSP's RAM in DMA mode.
The analog part is divided in two sub-parts: on one side, the speech acquisition and restitution stage, and on an another side the interface between the analog input and output parts of the radio-telephone and the DSP32C. A microphone input is used to pick up speech sounds. The analog signals are band-pass filtered before the linear-14 bit A/D conversion. The sampling frequency is 8 kHz and the data are sent to the serial input of the DSP32C. Data from the serial output port are converted in analog waveform and, then, they are sent to the audio-amplifier of the radio-telephone. The other transfers between DSP32C and the radio-telephone use PCM A-law 8 bit converters.

5.2 Noise reduction and echo cancelling implementation

The Forward-Backward adaptive digital filtering and speech detection have been implemented on the ATT DSP32C. The echo cancellation algorithm has been implemented in the DSP 56200 chip.
The results of real tests in the car can be summarized as follows:
-for the implemented echo cancellation function: some measures of echo reduction have been made in the car for various input levels of the microphone and we could observe that theoretical values have been approximately reached. Subjective tests performed in the context of real radio-communications confirmed this observation.
-concerning the noise reduction, the F/BADF filter does not correctly cancel the disturbing noise on speech signals, but speech is not altered.

5.3 DTW algorithm hardware implementation

The recognition system, based on a DTW algorithm, runs on the floating-point DSP32C which is driven by an AT-Personal Computer in the prototype. The experimenter controls the recognition process through this PC-AT. The system is speaker-dependent, so the user must perform a learning phase before beginning the recognition phase.

Description of the learning phase:
The user is seated in the parked car with the engine off. Then he has to utter twice all the words (about 50 words) belonging to its personal vocabulary, family names and some instructions; the second utterances are made in reverse order. Vocabulary words are stored as sets of vectors containing the cepstral coefficients. These coefficients are computed in real time during the acquisition process and they are stored in ROM. When acquisition is terminated, the user can enter the recognition mode. During the recognition process, the pronounced word is compared to all the words of the dictionnary. Then decision rules (the k-nearest neighbour) are applied with a rejection threshold.

The recognition stage, implemented on the DSP32C, is presently under real tests in the car and will be able to be confronted with results obtained during the simulations phase (see part 4).

6. Conclusion

This paper has given a description of a full hand-free radiotelephone with a vocal dialing function. First, the problem of the echo cancellation and noise reduction for transmission has been presented. Some techniques based on linear filtering have been proposed and tested in simulations. Then the speech recognition in ambiante car noise was addressed and a speaker-dependent approach based on the classical DTW was simulated on a database performed in real conditions. Finally, different algorithmical techniques were retained after the simulations phase, and were implemented on a hardware prototype. Real tests in a car has confirmed that the echo cancellation function was well efficient in medium noisy conditions. Nevertheless, the implemented noise reduction function has not given improvements, which was expected from the simulations phase. The vocal dialing function has been also implemented in the prototype and is presently under real tests in the car.

References

[1] A. Gilloire, J.F. Zurcher:"Achieving the control of the acoustic echo in audio terminals", EUSIPCO-88, Grenoble, pp. 491-494, September 1988.
[2] J.S. Lim, A.V. Oppenheim:"Enhancement and bandwidth compression of noisy speech", Proc. of the IEEE, vol. 67, n°12, pp.1586-1604, December 1979.
[3] P. Vary:"Noise suppression by spectral magnitude estimation-mechanism and theoretical limits-", Signal Processing 8, pp. 387-400, 1985.
[4] M.R. Sambur:"Adaptive noise cancelling for speech signals", IEEE ASSP-26, pp. 419-423, October 1978.
[5] J.W. Kim, C.K. Un:"Enhancement of noisy speech by forward/backward adaptive digital filtering", ICASSP'86, Tokyo, pp.89-92.
[6] I. Lecomte et al.:"Car noise processing for speech input", ICASSP'89, Glasgow, pp. 512-515, 1989.
[7] Y. Tohkura:"A weigthted cepstral distance measure for speech recognition", IEEE ASSP-35, n°10, pp.1414-1422, October 1987.
[8] D. Mansour, B.H. Juang:"The short-time modified coherence representation and its application for noisy speech recognition", IEEE ICASSP'88, 1988.
[9] B. Widrow et al.:"Stationary and nonstationary learning characteristics of the LMS adaptive filter", Proc. of the IEEE, vol. 64, n°8, August 1976.
[10] J.L. Botto:"Etude des algorithmes transversaux rapides: application à l'annulation d'écho acoustique pour l'audioconférence", Thèse de docteur-ingénieur, Univ. de Rennes, 1986.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A NOISE REDUCTION FOR SPEECH RECOGNITION SYSTEMS

SHOGO NAKAMURA, SYUUJI KUROKAWA, YOSHIHIKO HORIO, MAKOTO KOTANI

DEPARTMENT OF ELECTRONIC ENGINEERING, TOKYO DENKI UNIVERSITY
2-2, KANDA, NISHIKI-CHO, CHIYODA-KU, TOKYO 101, JAPAN

Algorithm for speech recognition systems which uses template matching on time-spectrum patterns are only effective in restricted environments because the recognition rate rapidly decrease when there is noise from the surroundings. In order to prevent this we have to eliminate the noise. So we have directed our research toward developing a simple noise reduction technique.

1. INTRODUCTION

In recent years, much research has been done on various speech recognition systems which use template matching on time-spectrum patterns and which function very effectively as human-machine interfaces. However, due to the algorithm used in these systems, outside restrictedenvironments their recognition rate rapidly decreases; this is caused by the interference of noise from the surroundings. In order to prevent the detetiotation of the recognition, we have to eliminate or grately reduce the effect of the noise. this is where we have directed our research. This paper describes a simple noise reduction technique we have developed for use with the time spectrumpatterns which are used in many speech recognition systems.

2. OUTLINE OF PROPOSED SYSTEM

The block diagram of the proposed system shown in Fig.1 consists of a preprocessor with two inputs, a noise reduction sectiorand a speech recognition section. Each of them is explained below in detail.

2.1 Pre-processing section

Primary input Ip(t) placed in front of a speech signal S(t) corrupted by environmental noise N(t). Reference input Ir(t) placed at the back of the speaker receives noise k(t)N(t+td). In comparison with noise, speech signals compose a small parcentage of Ir(t); this is because of signals in Ir(t) were not considered. k(t)N(t+td) is a correlated version of the noise received by Ip(t), where k(t) represents a relative level ration of reference noise Ir(t) to environmental noise N(t) in Ip(t) and td is the time difference between Ip(t) and Ir(t). The two input signals enter an 8ch digital Band-Pass Filter Bank (BPFB), and 8ch envelopes are obtained as the output of a low-pass filter bank. The envelopes are sampled and held at 100 Hz. The resultant time spectrum signals (Ip(nT)) of Ip(t) is represented as

$$Ipi(nT)=Si(nT)+Ni(nT) \qquad (1)$$

The Iri(nT) of Ir(t) is

$$Iri(nT)=k(nT)Ni(nT+td) \qquad (2)$$

where i refers to BPFB channel number. Noise reduction is performed for every frequency channel using these time spectrum data.

2.2 Noise reduction algorithm

The block diagram of the proposed noise reduction algorithm shown in Fig.2 consists of three sections. The first determines and renews the ratio k(nT) between Ni(nT) and Iri(nT), and calculates mean value (km) using this data. The second section detects speech, and the third smoothes cancelling output using a Cascaded Integrator Comb (CIC) low-pass filter.

We do not compensate for td because the distance between the two microphones in the system is at most 60 cm, and td is not significant for speech recognition using a time spectrum pattern sampled every 10 msec. Also td detection usually requires a great deal of computation time, e.g.for calculation of the frequently used correlation function.

Next, we explain the proposed noise reduction algorithm. First, level ratio k(nT) is computed using Ipi(nT) and Iri(nT). At the same time, the mean value (km) of k(nT) at certain successive points is computed. While Ipi(nT) includes speech signals, k(nT) is not computed and fixed on km. Therefore, we need to distinguish speech signals from noise. This has been accomplished in the following manner.

Spectrum differences Dpi and Dri in Fig.3 are computed by Eq.(3) and Eq.(4) respectively:

$$Dpi=[Ipi(nT)-Ipi((n-1)T)] \qquad (3)$$

and

$$Dri=[Iri(nT)-Iri((n-1)T)] \quad (4)$$

From these, the difference (Ddi) between Dpi and Dri is obtained. The beginning point of a speech signal can be obtained using Ddi. Thus, we can consider that speech has occured when Ddi exeeds the preset threshold level THi.
The procedure for computing and renewing the level ratios k(nT) is:

(1) For non-speech section, the level ratio k(nT) is computed by the Ipi(nT) and the Iri(nT) at every sampling.
(2) For mean value km, by use of the preceeding data.
(3) For speech, mean value km is used for noise reduction.

The procedure for detecting speech is:

(1) Ddi is computed by the spectrum differences of Ipi(nT) and Iri(nT).
(2) It is judged that speech has started when Ddi exeeds preset threshold level THi.
(3) Speech is considered finished when power level of the noise-reduction output nearly equals the level of the beginning of speech.

Thus, we can obtain the noise reduced speech time spectrum by compensating Ipi(nT) with Iri(nT). And finally, The noise-reduced time spectrum pattern is smoothed by the CIC low-pass filter as shown in Fig.4. The transfer function H(z) of this filter is represented as follows:

$$H(Z)=[(1-Z^{-4})/(1-Z^{-1})]^2/16. \quad (5)$$

This has a low-pass property to smooth the noise reduced signals.

3. SIMULATION

The proposed noise reduction algorithm has been aplied to a word recognition system using a Binary Time Spectrum Pattern (BTSP) which can recognize several hundreds words with a speaker-dependent recognition rate of over 98% when there is no noise from the surroundings. A block diagram of the recognition system is shown in Fig.5.
Table 1 shows the location of microphones and noise source. Figure 6 represents an example of time spectrum pattern of (a) speech signal only, (b) speech signal corrupted by noise, and (c) result of proposed reduction method. The simulation results which have been applied to the speech recognition system are shown in Table 2. Noise reduction was carried out for an input signal corrupted by noise and then speech recognition was tested for 30 different words. The results would be satisfactory for some applications.

4. conclusion

We have described a simple noise reducer for a speech recognition system using BTSP. The proposed method is based on cancellation in the frequency domain whose band is divided into several subbands. The proposed algorithm is very simple and makes it easy to implement a noise reduction system. Also, based on computer simulations it is found that this system could have practical applications. our next research goal is to enhance the noise reduction rate and to integrate this algorithm into a practical systems.

Acknowledgment

We would like to thank all collegues for many heipful discussions. Also, Mr.Naruse has made valuable contributions to the computer simulations of noise reduction.

Reference

1. B.Widrow, et.al " ADAPTIVE SIGNAL PROCESSING": Prentice-Hall Inc.
2. T.Muroi, et.al "Speaker-independent word recognition using partial expantion and weighted avereage templates": The Trans. of IECE vol.J69-A, No.1, P150
3. S.Kurokawa et.al "An approach of noise canceller on frequency domain using band pass filters": National Conference Record 1987 Information and System, IEICE155
4. S.Kurokawa et.al "An approach to noise cancellation for speech recognition systems": 1988 Autumn National Convention Record, IEICE-10
5. S.Nakamura et.al "A speech recognition by the method of tracking local peak powers": 1987 National Conference Record,
6. H.Takase et.al "A pre-processor for speech recognition Part 2": 1989 National Conference Record,IEICE
7. A.V.Oppenheim "Introduction to discrete time signal processing": Tutorial Lecture at the Joint Meeting of the ASA and ASJ Nov. 14, 1988
8. M.J.Shensa "Non-wiener solutions of the adaptive noise canceller with a noise reference": IEEE Trans. on Acoustics, Speech, and Signal Processing, vol.ASSP-28, no.4, august, 1980
9. B.Widrow et.al "Adaptive signal processing": Prentice-Hall, Dec. 1975

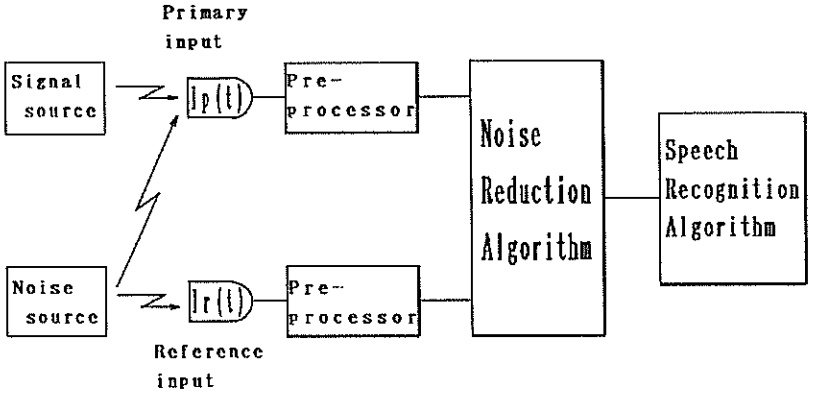
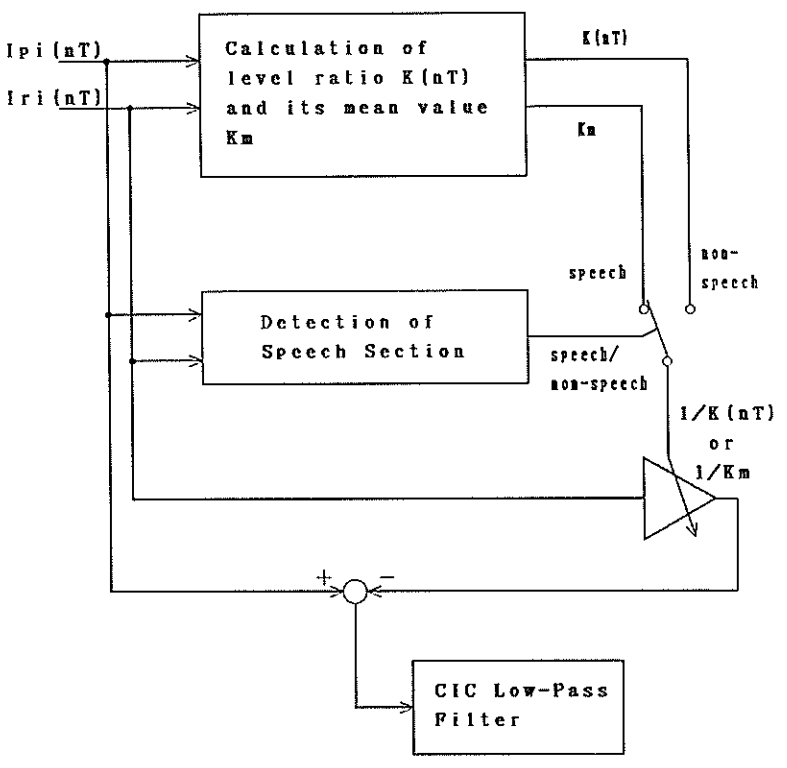


Fig.1 The Proposed Noise Reduction System



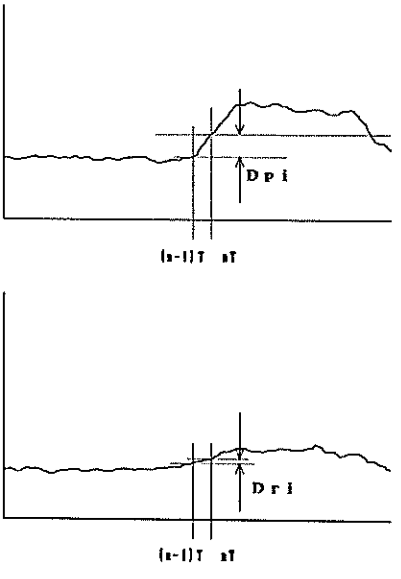
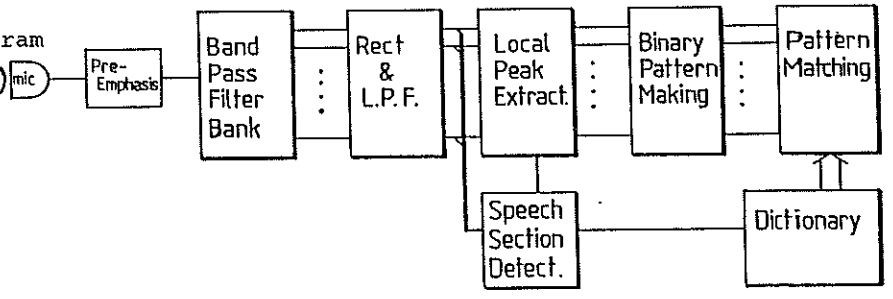


Fig.3 Detection of the Speech Section using time-spectrum difference

Fig.2 The block diagram of noise reduction algorithm

Fig.5 A block diagram of speech recognition system using BTSP



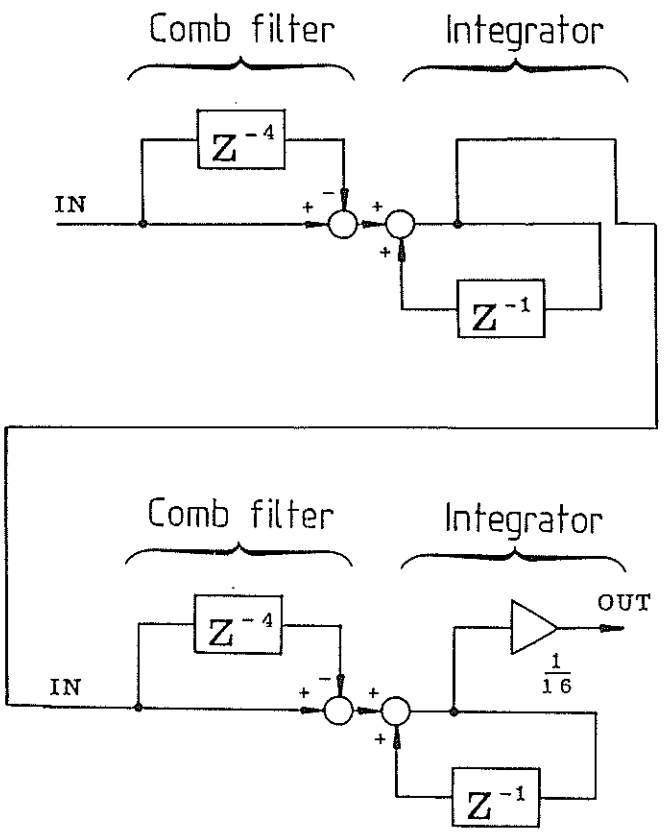


Fig. 4 A CIC low-pass filter

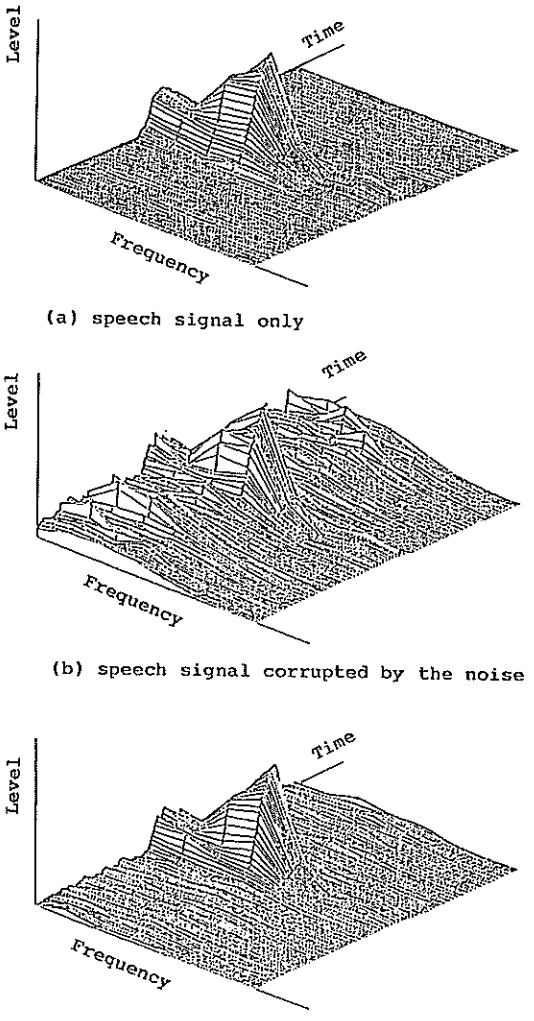


Fig.6 An example of the time-spectrum pattern

Table 2 Results of simulated recognition test which used the names of 30 cities in Japan.

	RECOGNITION RATE	
NOISE LEVEL	0dB	70dB
speaker #1	96.6%	72.4%
speaker #2	100.0%	83.3%
speaker #3	96.7%	90.0%
speaker #4	100.0%	86.7%
mean	98.3%	83.2%

Table 1. The conditions of simurations

Location of microphones	
Primary	Front of the speaker, about 30(cm)
Reference	Rear of the speaker, about 30(cm)
Noise sources	
In the automobile, In the subway, etc..	

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

MULTI-BAND ADAPTIVE CODEBOOKS FOR VXC

C. García-Mateo*, L.A. Hernández-Gómez**, A. Pena-Gimenez** and F.J. Casajús-Quirós**

* ETSI Telecomunicación-USC, Aptdo. 62, 36280 Vigo (Pontevedra) SPAIN
** ETSI Telecomunicación-UPM, Cdad. Universitaria, 28040 Madrid, SPAIN

[1] *In this contribuition we analyse some possibilities for Vector Excitation Coding of the speech signal using frequency domain representations. We have considered some possibilities based both on different characteristics of the frame of speech to be represented and on different representations depending on the frequency band. Our final interest is twofold; first, to include some of the results of this work into real-time time-domain CELP coders; and second, to propose more efficient VXC schemes for communication-quality speech at rates below 4.8 Kbps.*

1. INTRODUCTION

Stochastic Coding or Code-Excited Linear Prediction Coding (CELP) has been recently proposed as the DoD 4.8 Kbps Federal Standard [1] after a complete evaluation of 4.8 Kbps voice coders [2]. The final quality for this scheme and its robustness in acoustic noise, channel errors and tandem coding conditions lead us to a speech coder outperforming all the US Government standards at rates below 16 Kbps and comparable to 32 Kbps continuosly variable slope delta modulation (CVSD).

CELP coders can be considered as a two-stage Vector Quantizer (VQ) of the speech signal, based on analysis-by-synthesis search procedures using filtering operations with a synthesis filter obtained by means of linear prediction. The two-step VQ for the speech signal is obtained from:

1) An adaptive codebook, constructed from past samples of the synthetic speech, that models the long-term signal periodicity.
2) A fixed stochastic codebook, used to represent the remaining error after step 1.

Both stages are based on a search for the optimum excitation to an LPC synthesis filter.

One of the major drawbacks for this coding scheme has been a high computational cost concentrated on the search procedures. And one of the proposed alternatives for reducing computational load was the use of frequency domain representations [3]. Starting from this alternative the algorithm we refer to as Vector Adaptive Transform Coder (VATC) presented in [4] was a preliminary experience looking for an improvement in performance of conventional ATC by using VXC in the frequency domain. VATC was also derived as a two-stages Vector Quantizer but now for the Short-Time Fourier Transform of the speech signal. This has faced us with some differences if we compare the VATC scheme with the conventional CELP scheme. The main point to be exploited by using frequency domain representations in Vector Excitation Coding (VXC) is the possibility of different strategies for representing different frequency bands. And in the paper this is presented as a result of some experiments related to the use of frequency representations in time-domain VXC. In particular we will concentrate our attention on using different frequency bands for representing the past information of the speech: i.e. in the frequency-domain VQ stage where an adaptive codebook is used.

The use of different coding strategies for frames of speech with phonetically distintive characteristics is another improvement for conventional VXC (as CELP coders) we have considered. As stated in [5] for rates below 4.8 Kbps the scarcity of bits makes a finer control of bit-allocation achievable with a phonetically-based frame segmentation essential for preserving natural speech quality. In our scheme we include a segmentation strategy close to the proposed in [5] but with only 4 phonetically distinctive segment categories, so that different coding strategies are used for different categories. Those analysis frames corresponding to voiced sounds are more critical in our representation, and we have concentrated most of our discussion on the representation of this kind of segments. On the other hand we avoid the use of long-term prediction in representing unvoiced segments of speech.

The remainder sections of the paper are organised as follows. We first present (section 2) a description of some of our simulation results using frequency-domain representations in time-domain CELP schemes. Section 3 describes the possibilities for a basic frequency-domain VXC scheme. Different coding strategies for different segment categories are proposed in Section 4, focussing our attention on the representation of voiced sounds and on multi-band procedures for representing the information from an adaptive codebook. Some concluding remarks are given at the end of the paper.

2- FREQUENCY-DOMAIN SEARCH

In this section we discuss some simulation results for time-domain VXC using frequency-domain representations for the search procedures. The scheme we have considered, for research purposes only, divides the whole spectrum into three frequency bands; including different adaptive codebooks for different frequency bands. The search procedure is thus moved to the frequency-domain (carrying out 80 DFT points for a frame size of 40 samples) (see a similar scheme for the stochastic codebook in [3]) while choosing the adaptive codevectors by minimizing a squared error between spectral lines. Once the corresponding band-limited frequency-domain adaptive codevectors are selected, the adaptive contribution is obtained by taking the inverse DFT of the multi-band spectral representation (using the results for each sub-band). The search procedure for the stochastic codebook is

[1] This work has been supported by the PLANICYT under the Project "Tratamiento Avanzado de la Información".

made in a similar way after combining the Fourier transform of the sucessive stochastic codevectors with the Fourier transform of the impulse response of the synthesis filter.

Initial attempts with this scheme were focussed on determining how many spectral lines were necessary in order to obtain high-quality synthetic speech. We started reducing the number of spectral lines involved in the search. And we realize that a search based on a small number of lines sometimes improves the segmental SNR but loosing some accuracy in the spectrogram plot. For instance, by using only 7 spectral lines (over 41) in the adaptive codebook search, and 20 in the stochastic search we obtained high-quality speech with good spectrographics representations. At this point one question arises: what must we do with the remaining spectral lines? We have try two different procedures. First, we filled up those spectral lines with zeros both for the adaptive and the stochastic codebooks. Second, we use non-zero values for the adaptive codebook and zero values for the stochastic codebook. The result was a better spectral resolution at higher frequencies for the second alternative, and a low-pass filtering effect perceptually noticeable in the synthetic speech for the first one.

As summary, we can say that frequencies above 2500 Hz have a small contribution during the adaptive codebook search, and good results can be obtained by searching only for a reduced number of spectral lines. Thus the use of selective frequency-domain search procedures can be, as we will see in the next sections, an efficient way to improve the basic VXC structure in terms of both quality and computational cost.

3. POSSIBILITIES FOR A FREQUENCY-DOMAIN VXC

If we move now to frequency representations, what we obtain is a similar two-stage VQ but now for the Fourier transform of a frame of speech. And if we look at what we usually do in CELP coders we can notice how generally the Fourier transform to be represented in step 2 has some different characteristics if compared to the Fourier transform to be represented in step 1. In particular it is important to note the difference between the spectral envelope for the original speech and the spectral envelope necessary for the error to be represented in step 2. An the important point is to realise than in spite of this difference time-domain CELP coders use the same LPC envelope in both stages.

Figure 1 includes four graphics corresponding to: 1.a the Fourier transform of the original speech; 1.b the Fourier transform of the long-term contribution ; 1c the Fourier transform of the error to be represented in step 2 : difference between the original Fourier transform of the speech and the Fourier transform obtained in step1 ;and 1.d the Fourier transform of the contribution of step 2. The LPC spectral envelope of the original speech is included in the four plots.

Based on these results and in order to consider different possibilities from a frequency-domain representation, we have used a combination of conventional ATC techniques with VXC procedures. The aim of the proposed scheme is to represent the short-time Fourier transform (STFT) of the speech signal from two components [4] :

1) A Fourier transform obtained from past information of the synthetic speech. We call this component "log-term contribution".

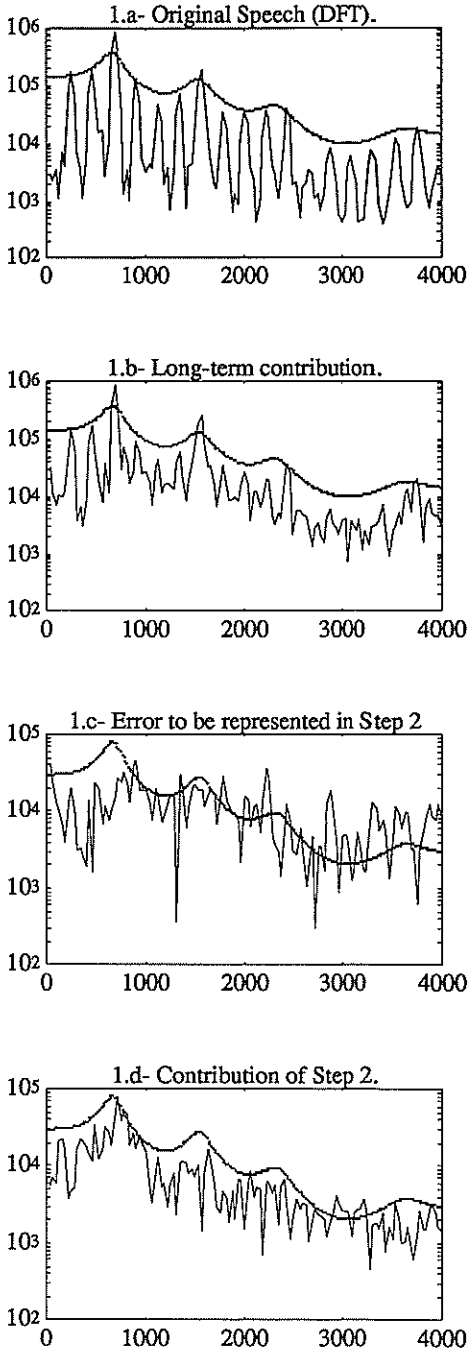


Figure 1. Different Fourier Transforms corresponding to different signals involved in time-domain VXC (see text) -the sampling rate is 8000 Hz.

2) A Fourier transform obtained using the optimum innovation sequence from a codebook of Fourier transformed excitation sequences. We will refer to this second component as "excitation codebook contribution".

This two components must be able to represent all the information included into the STFT of the original speech: the spectral envelope, the fine structure (voiced sounds) and the remaining details of the original Fourier transform.

Although this scheme seems similar to the corresponding time-domain VXC scheme we can now discover some important differences between them. Time-domain VXC algorithms include linear prediction to decorrelate the speech signal, and VQ to represent an excitation signal with a nearly flat spectrum. In addition to the use of linear prediction for representing a spectrum envelope, a frequency-domain VXC algorithm can combine its full-band division into multiple-bands for a more efficient representation of the speech signal. Two different combination strategies are possible: A) the use of a multi-band division of the full-band for the long-term contribution, trying to represent the variation of the spectral envelope from frame to frame, and an LPC envelope for the excitation codebook contribution. Or B) the use of an LPC envelope for the long-term contribution (and thus including an excitation sequence -its STFT- similar to the one used in time-domain VXC algorithms) and a multi-band division for the excitation codebook contribution. This two strategies can overcome the use of the same spectral envelope for both stages of the VXC coding algorithm (as in Figure 1) and thus it can provide a more efficient scheme.

The scheme we have used in this work corresponds to possibility A; i.e. division of the full-band for the long-term contribution and LPC envelope for the excitation codebook contribution. Furthermore, for those frames with pitch periods less than the analysis frame length a fine structure is included so as to avoid the addition of random noise between sucessive harmonics of the fundamental frequency.

Another important difference between time-domain VXC and the frequency-domain scheme we have used is the way in which the continuity between sucessive frames of synthetic speech is guaranteed. In time-domain VXC the synthesis filter memory carried over from the previous frames is enough to guarantee the continuity in the reconstructed signal. On the other hand, in VATC the "unfiltered" long-term contribution is a potential source of discontinuities in the synthetic speech. So in order to reduce block edge effects a proper overlapped structure in frequency-domain VXC must be included.

4. CODING STRATEGIES

One of the most important constraints in VXC algorithms is its rigid representation for successive frames of speech. In VXC techniques speech is generally divided into frames with a fixed frame size, and every analysis frame is represent under the same model and the same bit-allocation regardless of the particular phonetic information of the frame. Therefore the coder strategy must be general enough to be able to represented all the different phonetic categories in the speech signal resulting in a very inefficient representation.

The coding algorithm we propose includes a phonetic segmentation directed to the use of different representation strategies for frames with different phonetic features.

4.1 Segmentation.

In order to distinguised between different phonetic categories a segmentation algorithm has been included in our scheme. We have used four signal measurements to obtain the phonetic category of each frame. These measurements are: zero-crossing rate, low-band speech energy, sum of the sample magnitudes over the speech frame, and first reflection coefficient. The actual segmentation process can be considered as a simplified version of the proposed in [5], using the linear discrimination classifier reported in [6]. The segmentation algorithm resuls in the following speech categories:

- **Unvoiced** (including silence): noise like waveforms.
- **Onset**: transitions between voiced and unvoiced.
- **Voiced**: which are further subclassified into two categories: low-pass voiced; voiced sounds with a high resonance at low frequency; and "generalised" voiced, which corresponds to the remaining segments of voiced speech.

This segmentation process requires a maximum of 2 bits per frame. We are now considering the possibility of including a similar segmentation algorith but only requiring 1 bit per frame in time-domain CELP coders; and thus exploiting the possibilities offered by the "future expansion bit" suggested by B.S. Atal in the DoD 4.8 Kbps Federal Standard [1].

Once this phonetic classification is done, different coding strategies can be applied for representing different speech categories. The different coding strategies we have used are described in the following paragraphs.

4.2 Coding of Unvoiced Segments.

Unvoiced segments of speech have a noise-like waveform without significant correlation between samples. This "noisy" behaviour of unvoiced segments introduces two main considerations in our scheme. First, long term correlation does not show significant values, so it is wasteful to look for past samples of the synthetic speech in order to represent the actually speech frame. And we have omit the use of the first step in VATC (long-term contribution) for those segments declared unvoiced. During unvoiced sounds we found more useful to reduce the frame length and to concentrate our attention on the excitation codebook contribution.

And second, the spectral envelope of this "noise-like" signal has most of the spectral energy concentrated in the high frequency band with a more simple structure than the corresponding to voiced sounds. Therefore a 6th-order LPC model is enough for the proper representation of the spectral envelope in unvoiced frames. This spectral representation is based on a specific quantization procedure based on LSP (Line Spectrum Pairs).

4.3 Coding of Onset Segments.

The coding strategy we have used in representing those frames which corresponds to onset segments is similar to the one used in unvoiced frames. It is necessary to update the LPC parameters at a higher rate than in steady-state situations and the spectral envelope for this kind of segments has been chosen to be a 10th-order LPC model trying to have an accurate representation of these fast transitions. The spectral representation has again been made by using a specific quantization procedure based on LSP.

4.4 Coding of Voiced Segments.

For voiced segments both long-term contribution and excitation codebook contribution are needed to represent the speech signal.

The search in the adaptive codebook is performed by considering the full-frequency band divided into several

frequency bands. For the excitation codebook contribution an LPC spectral envelope for the error between the original speech and the long-term contribution is first obtained. And then the optimum excitation sequence is selected by exhaustive search through an excitation codebook. This procedure tries to be an alternative for providing an spectral envelope modification for step 1 (adaptive contribution) different from the one used in step 2 (stochastic contribution). And thus avoiding the use of the same spectral envelope for both steps with the annoying results showed in Figure 1 for time-domain VXC.

Taking advantage of our multi-band scheme we have used different frequency bands for representing different segments of voiced speech according to a phonetic criterion. We have sub-divided the class of voiced sounds in two sub-classes: one corresponds to voiced sounds with high spectral resonances at low frequencies, that we have called *"low-pass voiced segments"* ; and the other class corresponds to the remaining voiced sounds we have called *"generalised voiced sounds"* . Just because voiced sounds have most of their energy concentrated at the low frequency band, we have design our coder strategy for voiced frames based on long-term prediction for selected low frequency bands. Thus frequency-selective long-term prediction produces a long-term prediction error with a flatter spectrum then the original signal spectrum, and thus suitable for a characterisation based on a low-order LPC envelope.

Finally it is important to point out that a noticeable reduction in quantization noise during voiced speech can be achieved by including a fine structure, together with the LPC envelope and the selected innovation, in the excitation codebook contribution. This fine structure was defined in our multiband VXC from the delay obtained in the evaluation of the long-term contribution.

A more detailed description of the procedures and particular parameters together with the bit-allocation for every category can be found in [7].

5. CONCLUDING REMARKS

Preliminary simulation results show almost transparent synthetic speech at 4.8 Kbps and the potential of producing telephonic-quality speech at rates lower than 4.8 Kbps. The possibilities for maintaining the speech quality at lower bit-rates is now being considered by means of two major alternatives concerning with the coding strategy of voiced sounds. One alternative is to reduce the number of frequency bands used in the evaluation of the long-term contribution. This reduction could be based on a band-selective long-term prediction using only high-energy frequency-bands from previous frames. And the second alternative we are considering is based on more elaborated prediction strategies and the reduction of the number of bits assigned to the excitation codebook contribution.

Finally we are working towards the extension of some of the results of this research over time-domain VXC trying to analyse different possibilities to reduce the bit rate of the 4.8 Kbps US Government Standard while keeping its present performance. In particular we are dealing with some alternatives to improve real-time implementations of time-domain VXC and the use of the "future expansion bit" suggested by B.S. Atal [1].

6. REFERENCES

[1] J.P. Campbell, T.E. Tremain. and Jr; V.C. Welch "The DoD 4.8 Kbps Standard (Proposed Federal Standard 1016)". Advances in Speech Coding. Kluwer Academic Publishers 1990.

[2] D. Kemp, R. Sueda and. T.E. Tremain."An Evaluation of 4800 bps Voice Coders,". Proceedings of the IEEE Intl. Conf. ASSP 1989, p. 200-203.

[3] I.M. Trancoso and B.S. Atal "Efficient Procedures for Finding the Optimum Innovation in Stochastic Coders," in Proc. ICASSP, April 1986.

[4]L.A. Hernández Gómez, F.J. Casajús Quirós and R. García Gómez. "High-Quality Vector Adaptive Transform Coding at 4.8 Kbps," in proc. ICASSP, paper S4.7, New York, April 1988.

[5] S. Wang and Allen Gersho. "Phonetically-Based Vector Excitation Coding of Speech at 3.6 Kbps," in Proc. ICASSP, paper S2.3, Glasgow, May 1989.

[6] J.P. Campbell, and T.E. Tremain. "Voiced/Unvoiced Classification of Speech with Applications to the US Government LPC-10E Algorithm," in Proc. ICASSP, vol1, pp 437-476, Tokyo, April 1986.

[7]C. García-Mateo, F.J. Casajús Quirós and L.A. Hernández Gómez. "Multi-Band Vector Excitation Coding of Speech at 4.8 Kbps". in proc. ICASSP, paper 5S1.4, April 1990.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

An Efficient Approximation-Elimination Algorithm for Fast Nearest-Neighbour Search Based on a Spherical Distance Coordinate Formulation

V. Ramasubramanian and K. K. Paliwal

Computer Systems and Communications Group
Tata Institute of Fundamental Research
Homi Bhabha Road, Bombay – 400 005, INDIA

Abstract An efficient approximation-elimination search algorithm for fast nearest-neighbour search is proposed based on a spherical distance coordinate formulation, where a vector in K-dimensional space is represented uniquely by its distances from $K+1$ fixed points. The proposed algorithm uses triangle inequality based elimination rules which is applicable for search using metric distance measures. It is a more efficient fixed point equivalent of the Approximation Elimination Search Algorithm (AESA) proposed earlier by Vidal [2]. In comparison to AESA which has a very high $O(N^2)$ storage complexity, the proposed algorithm uses only $O(N)$ storage with very low approximation-elimination computational overheads while achieving complexity reductions closely comparable to AESA. The algorithm is used for fast vector quantization of speech waveform and is observed to have $O(K+1)$ average complexity.

Introduction

Nearest neighbour search consists in finding the closest point to a query point among N points in K-dimensional space. This search is widely used in several areas such as pattern classification, nonparamteric density estimation and data compression using vector quantization. Reducing the complexity of nearest-neighbour search is of considerable interest in these areas, and particularly in vector quantization encoding where the complexity increases exponentially with dimension K [1]. In this paper, we discuss fast nearest-neighbour search in the context of vector quantization. Vector quantization encoding is the minimum-distortion quantization of a vector $\mathbf{x} = (x_1, \ldots, x_K)$, (referred to as the *test vector*) using a given set of N K-dimensional *codevectors* called the *codebook* $\mathbf{C} = \{\mathbf{c}_j\}_{j=1,\ldots,N}$, of size N, under some distance measure $d(\mathbf{x}, \mathbf{y})$. This involves finding the nearest-neighbour of $\mathbf{x}$ in $\mathbf{C}$, given by $q(\mathbf{x}) = \mathbf{c}_k : d(\mathbf{x}, \mathbf{c}_k) \leq d(\mathbf{x}, \mathbf{c}_j)$, $j = 1, \ldots, N$, which requires N vector distance computations $d(\mathbf{x}, \mathbf{c}_j)$ using the exhaustive full-search.

One of the common approaches to fast nearest-neighbour search is the use of inexpensive elimination rules to eliminate codevectors which cannot be nearer to the testvector $\mathbf{x}$ than the current nearest-neighbour $\mathbf{c}_n$. Among these, the triangle inequality based elimination rule, which is applicable when the distance measure is a metric is a popular one [2]. If $d_n = d(\mathbf{x}, \mathbf{c}_n)$ is the current-nearest neighbour distance, then the triangle inequality based elimination rule rejects $\mathbf{c}_j$ if $d(\mathbf{c}_j, \mathbf{a}) > d(\mathbf{x}, \mathbf{a}) + d_n = r''$ or $d(\mathbf{c}_j, \mathbf{a}) < d(\mathbf{x}, \mathbf{a}) - d_n = r'$, where $\mathbf{a}$ is some point in $\mathcal{R}^K$. This elimination corresponds to a hyperannulus constraint of the search space, where codevectors lying outside the hyperannulus region formed between the two concentric hyperspheres of radius r' and r'', centered at $\mathbf{a}$ (henceforth referred to as 'anchor point') are eliminated. The hyperannulus has a width $r'' - r' = 2d_n$, and encloses the current nearest-neighbour ball of radius d_n. This elimination requires the N codevector - anchor point distances $\{d(\mathbf{c}_j, \mathbf{a})\}_{j=1,\ldots,N}$, and the test vector - anchor point distance $d(\mathbf{x}, \mathbf{a})$. The efficiency of this triangle-inequality (or hyperannulus) elimination increases with the use of more number of distinct anchor points, $\{\mathbf{a}_m\}_{m=1,\ldots,M}$, the codevectors retained after elimination being those present inside the intersection volume of the multiple hyperannulus corresponding to $\{\mathbf{a}_m\}_{m=1,\ldots,M}$. However, use of M anchor points requires the precomputation and storage of the MN codevector - anchor point distances $\{\{d(\mathbf{c}_j, \mathbf{a}_m)\}_{m=1,\ldots,M}\}_{j=1,\ldots,N}$ and the computation of the M test vector-anchor point distances $\{d(\mathbf{x}, \mathbf{a}_m)\}_{m=1,\ldots,M}$.

The AESA (Approximation Elimination Search Algorithm) proposed earlier by Vidal [2] employs this multiple anchor point hyperannulus elimination by using the codevectors themselves as the anchor points. Since the test vector-anchor point distances are then the test vector to codevector distances, only those codevectors whose distance to the test vector is known can be used as an anchor point. The codevectors are introduced as anchor points in an approximation-elimination framework, where the search consists of successive approximation to the actual nearest neighbour using repeated application of two main steps: i) Approximation: Selecting a codevector using an approximation criterion and ii) Elimination: Eliminating codevectors using the triangle inequality rule. The role of the approximation step is to select codevectors as close to the test vector as possible using some computationally inexpensive approximation criterion and serves as an efficient alternative to choosing codevectors in random or in a prespecified sequence. In this scheme, since the codevectors forming the anchor point set and the number of such codevectors are dependent on the test vector, any subset of the full codebook can play the role of the anchor point set. Therefore all the codevector-codevector distances $\{d(\mathbf{c}_i, \mathbf{c}_j)\}, i, j = 1, \ldots, N$ have to be precomputed and stored in a triangular array of size $N(N-1)/2$. This exhorbitant $O(N^2)$ storage complexity of AESA severely limits its practical use for large codebook sizes [2] and in DTW-based fast isolated word recognition for vocabulary sizes upto about only 400 [3].

In this paper, we propose an efficienct equivalent algorithm of AESA which uses only $O(N)$ storage with $O(K+1)$ average search complexity and worst case complexity closely comparable to AESA. The proposed algorithm uses only $K+1$ fixed anchor points and is based on the representation of a vector in K-dimensional space by its $K+1$ distances to some $K+1$ fixed points. This representation and the

approximation-elimination procedure based on this is given in the following sections.

Spherical Distance Coordinate Representation

Let $\{\mathbf{a}_m\}_{m=0,\ldots,M-1}$ be M points in $\mathcal{R}^K$. For any $\mathbf{x}$ in $\mathcal{R}^K$, let $d_m^x = d(\mathbf{x}, \mathbf{a}_m)$, where d_m^x is the Euclidean distance between $\mathbf{x}$ and $\mathbf{a}_m$, given by $(d_m^x)^2 = \sum_{j=1}^K (x_j - a_{mj})^2$. Denoting d_m^x as simply d_m, the M distance specification corresponds to M equations $e_m : \sum_{j=1}^K (x_j - a_{mj})^2 = (d_m)^2 \Rightarrow e_m : ||\mathbf{x}||^2 + ||\mathbf{a}_m||^2 - 2\sum_{j=1}^K x_j a_{mj} = (d_m)^2, m = 0, \ldots, M-1$. Subtracting expression e_m from e_0 gives,

$$\sum_{j=1}^K x_j a'_{mj} = \left[(d_0^2 - d_m^2) - (||\mathbf{a}_0||^2 - ||\mathbf{a}_m||^2)\right]/2 = D_m^x \quad (1)$$

where $a'_{mj} = a_{mj} - a_{0j}$. Given only $\{\mathbf{a}_m\}_{m=0,\ldots,M-1}$ and $\{d_m\}_{m=0,\ldots,M-1}$, this can be viewed as a set of $M-1$ linear equations in the K unknowns $(x_j)_{j=1,\ldots,K}$ and can be expressed as $\mathbf{Ax} = \mathbf{D}^x$, where $\mathbf{A}$ is a $(M-1) \times K$ matrix with the 'difference anchor vector' $(\mathbf{a}_m - \mathbf{a}_0)^T$ as row m, for $m = 1, \ldots, M-1$ and $\mathbf{D}^x = (D_1^x, \ldots, D_{M-1}^x)^T$. This set of $M-1$ equations in K unknowns will have a unique solution for $(x_j)_{j=1,\ldots,K}$, (i.e., $\mathbf{x}$) only for $M-1 = K$ (i.e., $M = K+1$) and the rank of $\mathbf{A}$ is K — which requires linear independence of its row vectors $\{(\mathbf{a}_m - \mathbf{a}_0)^T\}_{m=1,\ldots,M-1}$. Thus, a vector $\mathbf{x}$ in $\mathcal{R}^K$ can be uniquely represented by its $K+1$ distances $\mathbf{d}^x = \{d_m^x = d(\mathbf{x}, \mathbf{a}_m)\}_{m=0,\ldots,K}$, (henceforth referred to as the *(spherical) distance coordinates* of $\mathbf{x}$) from $K+1$ fixed points $\{\mathbf{a}_m\}_{m=0,\ldots,K}$ (called the anchor points, henceforth) which are such that the the K difference vectors $\{(\mathbf{a}_m - \mathbf{a}_0)^T\}_{m=1,\ldots,K}$ are linearly independent. When either $M < K+1$, or $M = K+1$ but the anchor points do not satisfy the linear independent condition, the solution will not be unique and the M distances of $\mathbf{x}$ from the M anchor points will not locate $\mathbf{x}$ uniquely. We refer to these anchor points as 'sub-optimal' (with respect to unique point representation). The locus solution of $\mathbf{x}$ for sub-optimal anchor points then corresponds to (1) being essentially underdetermined.

Based on this, the N codevectors $\{\mathbf{c}_j\}_{j=1,\ldots,N}$ are represented by their $K+1$ distance coordinates $\mathbf{d}^{c_j} = \{d_m^{c_j} = d(\mathbf{c}_j, \mathbf{a}_m)\}_{m=0,\ldots,K}$ using some specific optimal anchor point configuration. Given a test vector $\mathbf{x}$, its $K+1$ spherical distance coordinates $\mathbf{d}^x$ are also computed from the given optimal $K+1$ anchor points. Then, the approximation criterion, as used by Vidal [2] for selecting candidate codevectors $\mathbf{c}_a$ close to the test vector $\mathbf{x}$, given for the general case of M anchor points here will be: $\mathbf{c}_a = \arg\min_{\mathbf{c}_j \in \mathbf{C}} \alpha_M(\mathbf{d}^x, \mathbf{d}^{c_j})$, where, $\alpha_M(\mathbf{d}^x, \mathbf{d}^{c_j}) = \sum_{m=0}^{M-1} |d_m^x - d_m^{c_j}|$ is essentially a M-dimensional L_1 distance between $\mathbf{d}^x$ and $\mathbf{d}^{c_j}$. For optimal anchor point configuration, when $M = K+1$, these coordinates locate $\mathbf{c}_j$ and $\mathbf{x}$ uniquely and $\alpha_{K+1}(\mathbf{d}^x, \mathbf{d}^{c_j})$ will be a good approximation of the actual distance between $\mathbf{x}$ and $\mathbf{c}_j$ with a high degree of correlation. With respect to elimination, when $M = K+1$, the $K+1$ hyperannulus formed for a given current nearest-neighbour ball will have a convex intersection volume locally inscribing the current nearest-neighbour ball. However, when $M < K+1$, the sub-optimal number of distance coordinates of $\mathbf{x}$ and $\mathbf{c}_j$ do not represent them uniquely. Consequently, $\alpha_M(\mathbf{d}^x, \mathbf{d}^{c_j})$ will be a poor approximation of the actual distance and the approximation criterion can select codevectors which are actually very far from $\mathbf{x}$, though having smaller α_M values than the codevectors which are closer to $\mathbf{x}$. With respect to elimination, since $\mathbf{x}$ is not represented uniquely, the current nearest-neighbour ball volume specified in terms of the M perturbations in $\mathbf{d}^x$ and the corresponding M hyperannulus intersection volume will not be localized around $\mathbf{x}$. During elimination, this will result in retention of codevectors which are actually far from $\mathbf{x}$, but contained within the M hyperannulus intersection volume which now encloses the sub-optimal locus solution of $\mathbf{x}$.

Anchor Point Configurations

For the representation of the codevectors and test vector by their distance coordinates in the proposed approximation-elimination algorithms, we use the following two specific anchor point configurations which satisfy the optimal conditions for unique point location:

A1: $\mathbf{a}_0$ at the origin and $(\mathbf{a}_1, \ldots, \mathbf{a}_K)$ along the coordinate axes at equal distances ρ from the origin.

A2: $\mathbf{a}_0$ at the origin and $(\mathbf{a}_1; \ldots, \mathbf{a}_K)$ at equal distances ρ from the origin along the K principal component directions of the test vector data obtained as the directions of the eigen vectors of the covariance matrix of the test vector data. Here, anchor point $\mathbf{a}_m = \rho \mathbf{e}_m$, is located along the direction of the eigen vector $\mathbf{e}_m$ whose corresponding eigen value λ_m is the m^{th} maximum among the K eigen values $\lambda_m, m = 1, \ldots, K$.

The principal component directions represent the orthogonal directions of maximum variance in the data, proportional to the corresponding eigen values. Moreover, among $\{\mathbf{e}_i\}_{i=r,\ldots,K}$, $1 \leq r \leq K$, $\mathbf{e}_r$ is the direction of maximum variance and is also the solution for the best perpendicular least-square error (or eigen vector) fit, with minimum cross-sectional volume orthogonal to this direction. Therefore, an hyperannulus with its center placed well away from the origin along $\mathbf{e}_1$ will yield the minimum intersection (cross-section volume) with the data and the codevector distribution, implying high approximation-elimination efficiency, and this efficiency decreases gradually from $\mathbf{e}_1$ to $\mathbf{e}_K$. Thus, the A2 configuration will be more efficient than A1 in approximation and elimination, particularly in the case of speech waveform vector quantization where the vectors are highly correlated across their components.

Fixed Anchor Point - Approximation Elimination Search Algorithm (FAP-AESA)

The FAPAESA described here uses a constant number of M fixed anchor points all at the same time. This uses the approximation criterion given above for M anchor points to select the codevector at each approximation step.

Given $\mathbf{C} = \{\mathbf{c}_j\}_{j=1,\ldots,N}$, and $A_M = \{\mathbf{a}_0, \ldots, \mathbf{a}_{M-1}\}$, $\{\{\mathbf{d}^{c_j}\}_{m=0,\ldots,M-1}\}_{j=1,\ldots,N}$ are precomputed and stored with NM storage. Given a test vector $\mathbf{x}$, its M distance coordinates $\mathbf{d}^x = \{d_m^x\}_{m=0,\ldots,M-1}$ and the full approximation values $\alpha_M(\mathbf{d}^x, \mathbf{d}^{c_j}) = \sum_{m=0}^{M-1} |d_m^x - d_m^{c_j}|$, for $j = 1, \ldots, N$, are computed. Approximation - elimination search in FAP-AESA then involves repeating steps 1 to 3 until the codevector set $\mathbf{C}$ is empty:

Step 1: Approximation: $\mathbf{c}_a = \arg\min_{\mathbf{c}_j \in \mathbf{C}} \alpha_M(\mathbf{d}^x, \mathbf{d}^{c_j})$;
$\mathbf{C} = \mathbf{C} - \{\mathbf{c}_a\}$

Step 2: Distance computation and nearest-neighbour update
$d_a = d(\mathbf{x}, \mathbf{c}_a)$; *If* $d_a \geq d_n$ *then* continue at step 1
else $\mathbf{c}_n = \mathbf{c}_a$ and $d_n = d_a$

Step 3: Elimination: $\mathbf{C} = \mathbf{C} - \{\mathbf{c}_j : d_m^{c_j} < d_m^x - d_n$ or
$d_m^{c_j} > d_m^x + d_n$ for $m = 0, \ldots, M-1\}$

The main computational overhead in FAP-AESA is the full approximation carried out in the first step for all the N codevectors incurring N M-dimensional L_1 distance computation cost and the full M anchor point elimination step on all the N codevectors. In order to reduce this overhead cost incurred due to the use of all M anchor points, we describe in the following section, an algorithm which uses the anchor points incrementally as done in AESA [2].

Incremental FAP-AESA (IFAP-AESA)

The incremental realization of FAP-AESA, termed IFAP-AESA, is similar to AESA with the main difference that only a prespecified number of anchor points M is used. The anchor points are introduced at each approximation step upto a maximum of M anchor points (or until the codevector is exhausted, whichever happens earlier), after which the number of anchor points used is held constant at M, and the approximation - elimination search proceeds until the codevector set is exhausted.

The IFAP-AESA differs from FAP-AESA mainly in the approximation step with the introduction of an 'increment' control step which increments the anchor point set upto the specified maximum number of anchor points. Given $\mathbf{x}$, the anchor point set and approximation values are intialized as $A_{-1} = \{\}$ and $\alpha_0(\mathbf{d}^x, \mathbf{d}^{c_j}) = 0$ for all $\mathbf{c}_j \in \mathbf{C}$; step count $p = 0$. The modified approximation step with the increment control is:
Increment: $p = p+1$; *if* $p \not> M$, *then* [$q = p-1$; $d_q^x = d(\mathbf{x}, \mathbf{a}_q)$ and $A_q = A_{q-1} \cup \{\mathbf{a}_q\}$; $\alpha_{q+1}(\mathbf{d}^x, \mathbf{d}^{c_j}) = \sum_{\mathbf{a}_m \in A_q} |d_m^x - d_m^{c_j}|$ is computed incrementally as $\alpha_{q+1}(\mathbf{d}^x, \mathbf{d}^{c_j}) = \alpha_q(\mathbf{d}^x, \mathbf{d}^{c_j}) + |d_q^x - d_q^{c_j}|$]; *Approximation*: $\mathbf{c}_a = \arg\min_{\mathbf{c}_j \in \mathbf{C}} \alpha_{q+1}(\mathbf{d}^x, \mathbf{d}^{c_j})$; The distance computation and nearest-neighbour update are as in FAP-AESA, and for every update the elimination is done using the current anchor point set A_q with $q+1$ anchor points. The incremental accumulation of α_q is done upto $q = M-1$, i.e., M anchor points have been introduced. Prior to the incrementing of α_q value, elimination with the new anchor point $\mathbf{a}_q$ is done by rejecting $\mathbf{c}_j$ if $|d_q^x - d_q^{c_j}|$ is greater than d_n. For $p > M$, the approximation value of the remaining codevectors is saturated at α_M – the full M approximation value as α_q is not incremented further and the approximation involves only finding the codevector with the minimum full-approximation value α_M.

Here, the incremental accumulation of the approximation value at each step p requires only one coordinate difference computation for the codevectors in the current codevector set. As this codevector set size decreases progressively at each elimination step and for every addition of a new anchor point, the single component approximation cost decreases rapidly as the search progresses and the effective approximation overhead cost is much less than for FAP-AESA. Moreover, since the anchor point $\mathbf{a}_0$ introduced in the first step is known a-priori, the first approximation candidate $\mathbf{c}_a$ is found by a binary search on the ordered $\mathbf{a}_0$ codevector coordinates after computing $d(\mathbf{x}, \mathbf{a}_0)$. Elimination is also done by a binary search based truncation on this ordered list using the lower and upper hyperannulus bounds corresponding to the current nearest-neighbour radius $d(\mathbf{x}, \mathbf{c}_a)$. This cuts down the $O(N)$ initial approximation-elimination cost to $O(\log(N))$ resulting in significant computational overhead cost savings, while requiring only an additional N scalar storage for the ordered $\mathbf{a}_0$ coordinates of the N codevectors.

Experiments and Results

Here we present results characterizing the performance of FAP-AESA and IFAP-AESA and compare it with AESA [2] in the context of vector quantization of speech waveform data. The algorithms FAP-AESA and IFAP-AESA are studied for number of anchor points $M \leq K+1$ as well as $M > K+1$. For $M > K+1$, the $K+1$ optimal configurations A1 and A2 are extended with the anchor points $\{\mathbf{a}_m\}_{M>K+1}$ located in separate quadrants at equal distances from the origin.

Fig. 1. shows the performance efficiency of FAP-AESA for dimension $K = 10$ and codebook size $N = 1024$ using 5000 vectors. This is shown for M, the number of anchor points used, varying from 1 to 15. The basic elimination efficiency of M anchor points is measured by $\overline{nc}_a$, the average number of codevectors retained after elimination using the *actual* nearest-neighbour ball radius $d(\mathbf{x}, q(\mathbf{x}))$ to define the hyperannulus. This is the average number of codevectors inside the smallest M hyperannulus intersection volume achievable using the given M anchor points for a given test vector. This can be seen to decrease with increase in M indicating better localization of the intersection volume around the current nearest-neighbour ball with addition of anchor points. The elimination efficiency saturates for $M > K+1$, showing that use of optimal $K+1$ anchor points is quite sufficient to produce a saturating localization of the intersecting volume.

The approximation efficiency in using M anchor points is shown using $\overline{nc}_i$ – the average number of codevectors retained after elimination with the current nearest-neighbour as the *initial* codevector $\mathbf{c}_a$ selected by the approximation criterion in the first step. Given that use of the actual nearest-neighbour $q(\mathbf{x})$ achieves the complexity $\overline{nc}_a$, the difference between $\overline{nc}_i$ and $\overline{nc}_a$ is an indirect measure of the approximation efficiency as it indicates how close the codevector $\mathbf{c}_a$ selected by the approximation criterion is to $q(\mathbf{x})$. This is quite high for $M < K+1$ and indicates poor approximation efficiency. However, as M increases towards $K+1$, $\overline{nc}_i$ shows a drastic fall close to $\overline{nc}_a$, indicating the excellent approximation efficiency at $K+1$. The approximation efficiency however saturates for $M > K+1$. It can also be noted that configuration A2 performs significantly better than A1 with higher approximation and elimination efficiency, particulary for $M < K+1$, due to its principal component advantage in having smaller intersections with the speech data.

The actual complexity of number of distance computations carried out in FAP-AESA is given by $\overline{nc}$. This is very close to $\overline{nc}_a$ complexity for all M, indicating the high efficiency of the combined approximation-elimination steps in reducing the first step complexity of $\overline{nc}_i$ very close $\overline{nc}_a$ – the best complexity achievable by FAP-AESA if it starts with the actual nearest-neighbour itself.

In Fig. 2, we show the performance of IFAP-AESA for M (the maximum number of anchor points used) varying from 1 to 15 for $K = 10$ and $N = 1024$. In Fig. 2(a), $\overline{nc}$ gives the average number of distances computed and is the basic complexity of IFAP-AESA. The net complexity is $\overline{nc} + M$ – including M, the average number of test vector - anchor point distances computed. Also shown is $\overline{nc}_q$ – the average number of approximation steps required to locate the actual nearest-neighbour. This characterizes the effective approximation efficiency of the incremental scheme with smaller $\overline{nc}_q$ indicating more effective approximation. $\overline{nc}$ and $\overline{nc}_q$ decrease as M increases towards $K+1$ and saturates for $M > K+1$, indicating the saturation of both the approximation and elimination efficiency. It can also be noted that anchor point configuration A2 performs significantly better than A1 for all M. The average number of anchor points $\overline{M}$ actually used in the search is shown in Fig. 2(b) for various M. $\overline{M}$ is much less than M, indicating that the search terminates usually well before all the M anchor points are used. Configuration A2 uses less number of anchor points on an average than A1 demonstrating its higher approximation-elimination efficiency.

In Table I, we compare the performances of FAP-AESA, IFAP-AESA (for A1 and A2 configurations) with AESA [2] for dimension $K = 10$ and codebook size $N = 1024$ using $M = K+1 = 11$ anchor points in terms of computational and storage complexity and overhead costs. Here, the following can be noted: For FAP-AESA and IFAP-AESA,

A2 performs better than A1 in terms of complexity reduction and overhead costs. For a given configuration, IFAP-AESA performs very closely to FAP-AESA, but with considerbly reduced approximation-elimination overhead costs. FAP-AESA and IFAP-AESA have an additional storage of $N(K+1)$ and $N(K+1)+N$ respectively (i.e., only $(O(N)$ storage complexity) as against the $N(N-1)/2$, $O(N^2)$ storage complexity of AESA. In addition, the overhead costs are much lower for IFAP-AESA than AESA. Both FAP-AESA and IFAP-AESA (with the more efficient A2 configuration) perform very closely to AESA while using only a maximum of 11 anchor points as compared to a much larger (116) number of maximum anchor points required by AESA. The standard *macs* measure also shows FAP-AESA and IFAP-AESA to have less overall complexity than AESA. The significant complexity reduction offered by FAP-AESA and IFAP-AESA over full-search can be noted.

Table II shows the average and worst case complexities of FAP-AESA and IFAP-AESA using anchor point configuration A2, for codebook sizes $N = 256, 512, 1024$ with dimension $K = 10$, and $K = 8, 9, 10$ with $N = 1024$. The constant complexity performance of these algorithms can be observed across the dimensions and codebook sizes, with the additional overhead of $K+1$ distance computations between the test vector and $K+1$ anchor points resulting in their $O(K+1)$ average complexity.

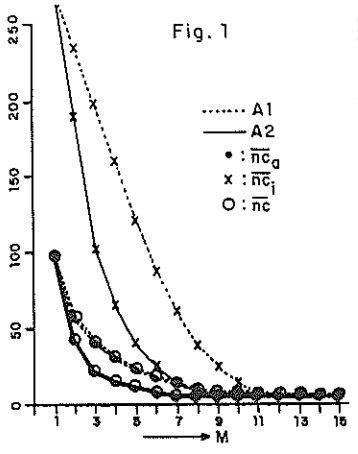
We have shown here that fast nearest-neighbour search in a K-dimensional space can be done efficiently using only $K+1$ optimal anchor points, the distances from which represent any vector uniquely. This scheme is an efficient alternative to AESA as it achieves comparable complexity reductions using significantly less storage. However, the proposed scheme does not extend readily as an alternative to AESA for fast DTW-based IWR search [3], as this involves the use of M optimal anchor points such that every word in the utterance in the given vocabulary is represented uniquely by its distances to these anchor points. This however raises the issue of finding M words from the given vocabulary set of size N (with $M << N$) to serve as the optimal anchor points under unique representation criterion. This also leads to the more interesting possibility of representing words in a regular M-dimensional space with the M distances to the M optimal anchor points as the coordinates of a word. This has important implications in terms of fast search and in applying standard pattern recognition techniques of regular vector spaces to words and word-like units.

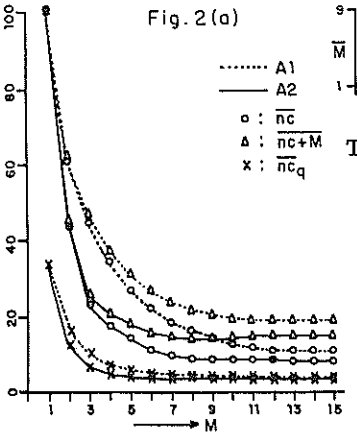
References

[1] J. Makhoul, S. Roucos and H. Gish, "Vector quantization in speech coding", Proc. IEEE, vol. 73, pp. 1555–1588, Nov. 1985.

[2] E. Vidal, "An algorithm for finding nearest neighbours in (approximately) constant average time complexity", Pattern Recognition Letters, No. 4, pp. 145–157, 1986.

[3] E. Vidal, H. M. Rulot, F. Casacuberta, and J. M. Benedi, "On the use of a metric-space search algorithm (AESA) for fast DTW-based recognition of isolated words", IEEE Transactions of Acoustics, Speech and Signal Processing, vol. 36, No. 5, pp. 651–660, May 1988.





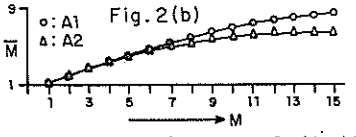


Table II — Performance of FAP-AESA & IFAP-AESA
Anchor point configuration: A2

K = 10					N = 1024				
	FAP-AESA		IFAP-AESA			FAP-AESA		IFAP-AESA	
N	$\overline{nc}$	$\widehat{nc}$	$\overline{nc}$	$\widehat{nc}$	K	$\overline{nc}$	$\widehat{nc}$	$\overline{nc}$	$\widehat{nc}$
256	4.5	76	6.5	79	8	4.0	85	6.5	86
512	5.2	88	7.6	91	9	4.8	126	7.5	129
1024	5.9	117	8.7	120	10	5.9	117	8.7	120

$(\overline{nc}, \widehat{nc})$: (average, maximum) number of codevector distances computed

Table I — Performance of FAP-AESA, IFAP-AESA and AESA [2]; Dimension $K = 10$, Codebook size $N = 1024$

Algorithm	AC	M		nc		nc+M		Additional Storage	Approximation Cost	Elimination Cost	m	a	c
		ave	max	ave	max	ave	max						
FAP-AESA	A1	11	11	7.9	137	18.9	148	11264	1118	1956	18.9	1982	2066
	A2	11	11	5.9	117	16.9	128	11264	1081	1699	16.9	1978	1805
IFAP-AESA	A1	7.4	11	11.8	142	19.2	153	12288	101	339	19.2	239	441
	A2	6.2	11	8.7	120	14.9	131	12288	67	213	14.9	163	281
AESA		11.4	116	11.4	116	11.4	116	523776	172	741	11.4	366	914
FULL-SEARCH		–	–	1024	1024	–	–	0	–	–	1024	1946	1023

AC: anchor point configuration; **M**: number of anchor points used; **nc**: number of codevector distances computed; **(m, a, c)**: average number of (multiplications, additions and comparisons)

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

FAST SOURCE-INDEPENDENT VECTOR QUANTIZERS AND THEIR APPLICATION IN SPEECH PROCESSING

H. Brehm and M. Herbert
Lehrstuhl für Nachrichtentechnik, Universität Erlangen-Nürnberg
D-8520 Erlangen, Cauerstr. 7, Fed. Rep. of Germany

The real time application of vector quantization is limited to low vector dimensions and coding rates due to the implementation complexity. In this paper we present a compander system for source-independent vector quantization of spherically invariant random processes. Therein lattice quantizers are used which have the advantage of fast quantization algorithms and save of codebook memory. The theoretical approach is confirmed by experimental results for synthetical as well as for speech signals. The performance of the compander system combined with an adaptive vector quantization scheme is compared with the rate-distortion function. Furthermore a speech coding system is presented which combines lattice quantizers and conventional vector quantizers. Thereby a better speech quality and a reduction of implementation complexity can be achieved.

1 Introduction

Vector quantization is widely used in speech and image coding. Vector quantizers exploit the linear and non-linear dependencies of subsequent samples. Therefore their performance is better than that of scalar quantizers. In practice vector quantizers are designed by the iterative Linde-Buzo-Gray (*LBG*) algorithm using a training sequence to match the quantizer to the statistical properties of the signal source. An important disadvantage of *LBG* codebooks is the lack of structure which could be used for fast quantization algorithms. The computational expense of the necessary exhaustive search as well as the codebook memory grow exponentially with the vector dimension and the coding rate. Already the application of a modest vector dimension and codebook size is limited.

The alternative to the *LBG* quantizers are lattice quantizers. However, they are only optimal for sources with uniformly distributed signal vectors. The codevectors of the lattices are arranged on a regular grid. They can be described by a linear combination of k basic vectors in $\mathbb{R}^n$, $k \leq n$. Usually k equals n. Using matrix notation the so-called generator matrix $\underline{M}$ is built up by the basic row vectors and the lattice vectors are determined by

$$\underline{c} = \underline{l} \cdot \underline{M}, \tag{1}$$

where $\underline{l}$ is a k-dimensional row vector whose components are integer numbers.

The simplest lattice, the Z-lattice, consists of all points with integer coordinates. Thus, signal vectors are quantized by rounding their components. Another important lattice is the D-lattice. Its codevectors are the subset of the Z-lattice with the constraint that the sum of the vector's components is even. The quantization error of the D-lattice is smaller than that of the Z-lattice and the complexity of the coding algorithm is only a little larger. Many other lattice types exist [1]. Our investigations [2] of various lattices show that the D-lattice is the most favourable. The other lattices have a complexer structure and therefore more complicated quantization algorithms with only slight improvement in coding quality.

Lattice quantizers have two advantages compared to *LBG* quantizers. Firstly, the quantization procedure is performed by the manipulation of the signal vector and is independent of the codebook size and the coding rate. The expense depends on the vector dimension only linearly. Thus the 'per sample' expense is constant. Secondly, neither in the transmitter nor in the receiver a stored codebook is required. Despite of these advantages, lattice quantizers were not used so far, because they are optimal for uniformly distributed signal vectors only.

2 Source-independent vector quantization

Now we consider the class of spherically invariant random processes (SIRPs) which are widely used as signal models in speech and image processing. Important examples are processes with Gaussian, Laplacian or Gamma probability density functions (*pdf s*). Band-limited speech signals can be modelled excellently by SIRPs as shown in [3].

Recently we have derived a transformation for SIRPs which maps the decorrelated source signal to another one with a uniform distribution in a k-dimensional sphere [2]. Thus, we can use the lattices for quanti-

zation and map the quantized signal back by the inverse transformation. The processing scheme can be interpreted as a k-dimensional compander system. The compressor transformation is followed by a uniform vector quantizer and the inverse expander transformation. The system is source-independent, because the transformation is identical for all different SIRPs, and can be applied to speech coding.

In [2] the transformation was derived by theoretical considerations on SIRPs. The result can be summarized as follows:

Given n samples of the decorrelated signal, the vector norm

$$r = \left[\sum_{i=1}^{n} x_i^2\right]^{1/2} \tag{2}$$

is calculated. Then the samples x_i are transformed to y_i by

$$y_i = \frac{x_i}{r} \qquad for \quad i = 1, \ldots, n-2. \tag{3}$$

The new samples $y_i, i = 1, \ldots, m$ with $m = n-2$ form a vector $\underline{y}$ in the m-dimensional unit sphere. If the signal source can be discribed by a SIRP, as assumed, the vectors $\underline{y}$ are uniformly distributed within the sphere.

The *pdf* of the transformed signal was measured for the one- and two-dimensional case to verify the theoretical result. Both for a *SIRP* signal with univariate Gamma desity and for a decorrelated speech signal the expected uniform distribution was achieved very well [2].

The measurement of the higher dimensional *pdf*s is not practical due to the computational expense and a missing representation possibility. Therefore the *pdf* of the vector norm

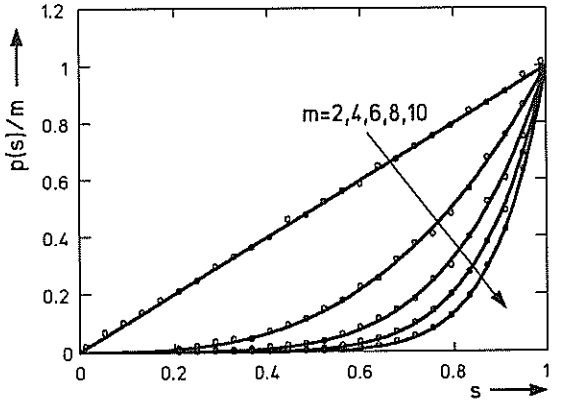


Figure 1: Theoretical and measured *pdf* of the vector norm after transformation for a Gamma SIRP (dimensions m=2,4,6,8,10)

$$s = \left[\sum_{i=1}^{m} y_i^2\right]^{1/2} \tag{4}$$

is considered. It is given by

$$p(s) = m \cdot s^{m-1} \tag{5}$$

for a uniform distribution in the m-dimensional unit sphere.

In figure 1 and 2 the theoretical and measured *pdf*s are depicted for a Gamma SIRP and a decorrelated speech signal. The measured data agree well with the expected curves and confirm that the uniform distribution is achieved. This is true for both the Gamma SIRP and the decorrelated speech signal. Thus the transformation and the vector quantization system is independent of the statistics of the source signal as stated above.

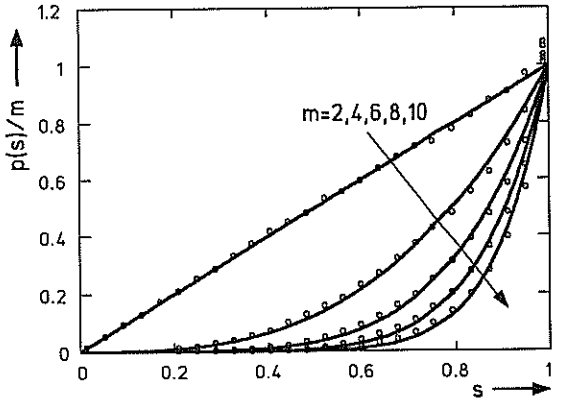


Figure 2: Theoretical and measured *pdf* of the vector norm after transformation for a bandlimited speech signal (dimensions m=2,4,6,8,10)

3 Quantization of SIRPs

In this section we discuss experimental results on spherically invariant speech-model processes. The proposed coding scheme was implemented and compared to LBG vector quantization ($LBG - VQ$) in [2]. The compander system outperforms the $LBG-VQ$ at coding rates above one bit/sample. At rates below this value the lattice suffers from edge effects since their Voronoi regions (the m-dimensional quantization intervals) cannot fill up the signal sphere completely. At higher rates the number of codevectors is larger and the influenence of the edge effects deminishes. Therefore the application seems to be restricted to high rate coding, for example high quality speech coding, where lattice quantizers can improve coding quality and reduce coder complexity.

In the field of low bit rate coding a more sophisticated approach is necessary. The comparison of the results of the compander system with the bounds of the rate

distortion theory shows a gap of about 10 dB at a rate of 1 bit/sample. This gap can be closed by an adaptive vector quantization scheme as proposed in [4]. The fixed rate vector quantizer is replaced by an ensemble of vector quantizers with different rates.

Using the fact, that a SIRP can be represented as a mixture of Gaussian processes of different variances, the signal is divided into frames of length L. Each frame is considered as a realization of a Gaussian process whose variance equals the estimated short time variance v^2. The coding rate of the frame is chosen according to

$$r(v) = log\frac{v}{w} \qquad for \quad v \geq w \tag{6}$$

This rule was derived from the rate-distortion theory of SIRPs [4]. A given overall coding rate can be achieved by suitable choice of the parameter w.

Instead of Gaussian LBG codebooks as in [4] we now use the lattice compander system for quantization. Thereby we benefit from the fast quantization algorithms and the save of memory space. Furthermore lattice quantizers have a better performance at rates above 2 bit/sample. We coded $4 \cdot 10^5$ samples of a Gaussian source by D-lattice and LBG vector quantizers. The results are compared with the rate-distortion function (RDF) of the Gaussian process in figure 3.

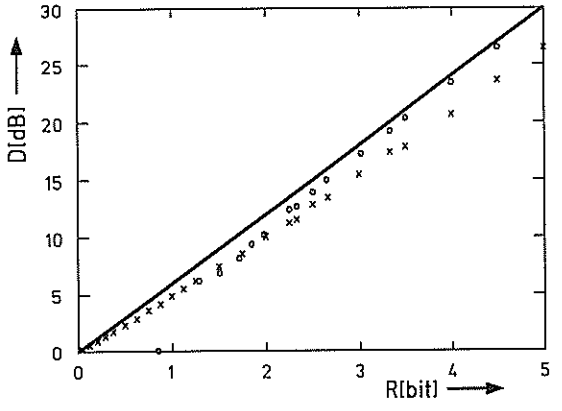


Figure 3: Distortion of Gaussian $LBG - VQ$ (x) and the lattice compander system (o) compared with the RDF of the Gaussian source

At rates above 2 bit/samples the signal to noise ratio (SNR) of $LBG - VQ$ differs from the RDF considerably because only vector dimensions below 4 are realizable. The improvement by the lattice quantizers is between 2 and 3 dB. The rate-distortion function is reached up to 0.5 dB. At rates below 2 bit/samples the performance of lattices decreases due to the mentioned edge effects. Here the Gaussian vector quantizers are better and the rate-distortion function is almost attained.

These results suggest to use a composite ensemble of Gaussian codebooks at low rates and lattices at high rates. Gaussian codebooks are employed in a range where their complexity is still low and their performance is superior to lattices.

The adaptive quantization system with the composite ensemble is utilized to quantize SIRPs with univariate Laplacian, K_0 and Gamma distribution. Additionally a process matched to the distribution of a speech signal is considered. Its pdf is given by the G-function $0.056 \cdot G_{02}^{20}(0.133 \cdot x^2 | -1/3, 0.3)$ [3].

We use the D-lattice of dimension 8 and Gaussian codebooks of vector dimensions between 2 and 24. The ensemble contains codebooks with rates between 0.1 and 5.0 bit/sample. The adaption frame lenght is 48. Table 1 shows the results for an overall coding rate of 1 bit/sample. Below the switching rate $R_{G/L}$ the Gaussian codebooks, above it the lattice quantizers are employed.

SIRP	$R_{G/L}$			RDF
	∞	1.50	2.00	
Laplace	8.01	7.85	8.12	8.19
K_0	9.59	9.46	9.81	10.01
Gamma	10.91	11.08	11.35	11.74
G_{02}^{20}	12.76	13.25	13.43	13.99

Table 1: SNR in dB for different SIRPs and switching rates $R_{G/L}$ compared with the RDF

The improvements for the Laplacian and K_0 SIRPs are only moderate due to the fact that already the Gaussian ensemble ($R_{G/L} = \infty$) almost reaches the RDF. For the other, more "speech-like" distributed processes the improvement by the composite ensemble is greater. The RDF is reached up to a gap of 0.5 dB. The computational expense is reduced by the factor two for the composite ensemble.

So far, the transmission of the side information was not taken into account. For the Gaussian ensemble the short time variance has to be transmitted. For the lattice quantizer the side information of the compander system has to be coded. Note that in the lattice case the variance can be calculated approximately from the compander side information and need not to be transmitted.

We use a 2 bit Max-Lloyd quantizer to code the compander side information and a 6 bit logarithmic quantizer for the variance. Additionally 1 bit/frame is required for signalling whether a Gaussian codebook or a lattice is selected. Using the lattice dimension 8 and the frame length 48 the resulting side information rate is 0.125 bit/samples for the Gaussian and 0.161 bit/sample for the composite ensemble, because only 25 percent of the frames are coded by lattices.

The SNR of the composite ensemble decreases between 0.5 and 1.0 dB due to the quantization of the side information, because the uniform distribution of the transformed signal is disturbed in the lattice case. Thereby the advantage over the Gaussian ensemble is lost. Recent investigations indicate that the decrease could be prevented, if a better coding of the side information is performed. Alternatively a higher vector dimension could be used.

4 Coding of speech signals

In this section we discuss a system for speech coding using adaptive vector quantization with the composite ensemble. The speech signal is decorrelated by a cascade of a 10 tap short term and a 1 tap long term prediction filter. The frame variance v^2 of the residual signal is estimated and the actual rate is calculated by (6). According to the switching rate $R_{G/L}$ either a Gaussian codebook or a lattice quantizer is selected. In the Gaussian case the variance v^2 is quantized by a 6 bit logarithmic quantizer, in the lattice case the side information of the compressor transformation is determined and quantized by a 2 bit Max-Lloyd quantizer. An approximate value of v^2 is computed. The residual is scaled by v and coded by the corresponding vector quantizer. The receiver is built up in the reverse order.

The filters are updated every 144 samples. The short term filter coefficients are coded by a 7 bit vector quantizer, the pitch lag by 6 bit and the pitch filter coefficient by a 2 bit Max-Lloyd quantizer. The rate of the residual vector quantizer is adapted in frames of 48 samples. Using a sampling rate of 8 kbit/s the filter information and the signal bit for the quantizer type (Gaussian/lattice) take 1 kbit/s. The side information for scaling requires between 1 and 2 kbit/s depending on the switching rate $R_{G/L}$. Typically 1.25 kbit/s are required.

A sequence of 180000 speech samples (22.5s) of a male speaker was coded with different rates. The results of the Gaussian and the composite ensemble ($R_{G/L}$=1.5 bit/samples) are compared in table 2.

desired rate	8 kbit/s		12 kbit/s	
VQ type	G	G/L	G	G/L
actual rate	7.95	7.97	11.99	11.94
SNR [dB]	6.87	6.39	9.91	10.15
SNRQ [dB]	10.12	10.34	13.85	14.25
RDF [dB]	12.7		17.2	
complexity	1.00	0.76	1.33	0.77

Table 2: Performance and relative complexity of the Gaussian and the composite ensemble ($R_{G/L}$=1.5 bit/samples)

Therein SNR denotes the signal to noise ratio of the whole coder (transmitter input to receiver output), SNRQ the signal to noise ratio between prediction filter output and synthesis filter input (i.e. the quantization error). Obviously the SNRQ is about 3 dB higher than the SNR but these two values are based on different error criteria due to the filter transfer function. Only the SNRQ can be compared with the RDF. The results show a slight improvement by the composite ensemble. If the compressor side information is not quantized, the values are about 0.5 dB higher because the uniform distribution is better attained by the transformation. A better coding of the side information may improve the results.

The subjective quality with side information quantization is the same for both ensembles, without it the composite ensemble is better. The main advantage of the Gauss-lattice ensemble is the reduction of the complexity.

5 Conclusion

We have presented a source-independent vector quantizer system for $SIRPs$ using fast lattice quantizers which can be used for speech signals as well. It seems well suited to high rate coding because the rate-distortion function can be reached and the computational expense and codebook memory is dramatically reduced. For low rate coding we used an adaptive vector quantization scheme and a composite ensemble of conventional Gaussian codebooks and lattice quantizers. We coded both synthetic and speech signals and compared the composite ensemble with a Gaussian ensemble. The new system results in an improvement of SNR which is however compensated by the quantization of the side information. Thus we attain nearly the same results in both cases but the computational complexity and the codebook memory is reduced. The performance of the new system could be improved, if a better coding of the side information or higher dimensional lattice quantizers are used.

References

[1] Conway, J.H. and Sloane, N.J.A., Sphere Packings, Lattices and Groups (Springer, New York, 1988)
[2] Brehm, H. and Herbert, M., Lattice Quantizers in Speech Coding, Proc. Int. Conf. ASSP, Glasgow 1989, pp. 140-143
[3] Brehm, H. and Stammler, W., Description and Generation of Spherically Invariant Speech-model Signals, Signal Processing, 1987, pp. 119-141
[4] Brehm, H. and Trottler, K., Adaptive Vector Quantization of Band-limited Speech and Speech-like Waveforms, Signal Processing IV, Proc. EUSIPCO, Grenoble, 1988, pp. 887-890

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Information-Theoretic Performance Bounds for Adaptive Speech Coding

Hans Kalveram, Peter Meissner

Institut für Allgemeine Nachrichtentechnik, Universität Hannover,
Appelstr. 9A, D-3000 Hannover 1

Composite source models for speech waveforms provide rate distortion bounds which are especially applicable for adaptive coding schemes, because various adaptation techniques can be reflected by the source model. The computation of these performance bounds leads to practical rules for designing coding schemes with time varying transmission rates.

1 Introduction

Composite source (CS) models for speech signals consist of a finite number of subsources describing specific spectral characteristics and a switch process representing the variation of these characteristics [2].

Rate distortion theory [1] provides performance bounds for any source coding scheme working on a given source model. Modeling speech by a CS allows the computation of rate distortion functions (RDF's) with respect to the mean squared error criterion, which are significantly below those for simple stochastic models. Previously, we examined the relationship between Itakura-Saito distance (ISD) and rate distortion theory for CS models [4, 3]. These results are extended to a gain-adaptive CS, where the switch state estimation is accomplished by the gain-optimized version of the ISD.

2 Gain-adaptive composite source model for the speech waveform

We consider a CS as a family of K vector-valued subsources $X^{(1)},\ldots,X^{(K)}$ and a switch S, which connects one of the subsources to the output Y. Given the switch is in state $S_t = s_t \in \{1,2,\ldots,K\}$ at time t, then the output will be the random vector $Y_t = X_t^{(s_t)}$. The subsources and the switch process are assumed to be independent identically distributed (i.i.d.) and mutually independent. Each $X_t^{(k)}$ being multivariate Gaussian distributed, the output $X_t^{(s_t)}$ obeys a mixture of these distributions.

In modeling a speech waveform, each output vector Y_t represents a single speech frame y_t of M samples. The number of vectors coming from a specific subsource depends on the duration, the switch connects that subsource to the composite output. We consider Gaussian distributed subsources with zero mean and a covariance matrix, which is identical to the $M \times M$ Töplitz covariance matrix of a stationary autoregressive process of order p. This covariance matrix $\Sigma(a,v)$ can be parametrized by the predictor coefficients $a = (a_1,\ldots, a_p)$, $(a_0 = 1)$ and the residual variance v.

The most natural assumption, made in our previous work [3, 4], is to provide a specific predictor $a^{(k)}$ and residual variance $v^{(k)}$ for each subsource $X^{(k)}$, $k = 1, \ldots, K$ representing a specific autoregressive model spectrum. In this way, CS models have the capability to represent spectral variations as well as the non-Gaussian probability density of speech signals.

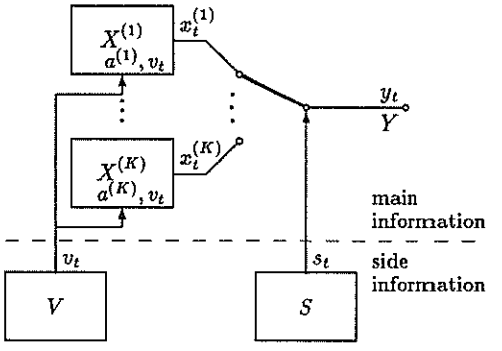


Fig. 1: Gain-adaptive composite source.

In this paper, we assume a CS model (cf. Fig. 1) which is still based on a set of K predictors $a^{(1)},\ldots, a^{(K)}$ selected by s_t, but represents the residual variance of each frame y_t individually by v_t. So the covariance matrix of the vector $X_t^{(k)}$ is given by $\Sigma(a^{(k)}, v_t)$, $a^{(k)} = (a_1^{(k)}, \ldots, a_p^{(k)})$. The output Y_t produces the main information of a waveform coding scheme, which may use additional side information from estimates of S and V, representing the selection of spectral shape and gain separately. In the following, no assumptions about these processes are required, because only conditional probabilities, given the state sequence $\mathbf{s} = (s_1,\ldots, s_T)$ and the sequence of residual variances $\mathbf{v} = (v_1,\ldots, v_T)$, are concerned. Let

$\Theta_a = (a^{(1)}, \ldots, a^{(K)})$ denote the subsource parameters and $\varphi(.;\mu,\Sigma)$ the Gaussian p.d.f., the p.d.f. of speech, represented by a concatenation $\mathbf{y} = (y_1, \ldots, y_T)$ of output vectors y_t is

$$\begin{aligned} p(\mathbf{y}|\mathbf{s},\mathbf{v},\Theta_a) &= \prod_{t=1}^{T} \varphi(y_t; 0, \Sigma(a^{(s_t)}, v_t)) \qquad (1) \\ &\approx \prod_{t=1}^{T} (2\pi v_t)^{-\frac{M}{2}} \exp\Big(-\frac{M}{2v_t} Q(y_t; a^{(s_t)})\Big). \end{aligned}$$

The approximation is valid for $p \ll M$, using

$$Q(y_t; a) = \sum_{j=0}^{p} a_j \sum_{i=0}^{p} a_i c_{i-j}(y_t), \qquad (2)$$

where

$$c_i(y_t) = \frac{1}{M} \sum_{m=1}^{M-|i|} y_{t,m} y_{t,m+|i|}, \qquad (3)$$

denotes the autocovariance function of frame y_t.

In principle, this model may be seen as a special case of the model considered in [3, 4] by representing each combination of $a^{(k)}$ and v_t, $k = 1, \ldots, K$, $t = 1, \ldots, T$ in a specific subsource; however, the parameter estimation must be revised. The advantage of the gain-adaptive CS, especially with regard to coder design and the computation of performance bounds, is its enhanced capability to model the time-varying spectral shape and gain of speech.

3 Maximum-Likelihood estimation and the gain-optimized Itakura-Saito distance

For a fixed number of subsources K, we take the Maximum-Likelihood approach

$$p(\mathbf{y}|\tilde{\mathbf{s}},\tilde{\mathbf{v}},\tilde{\Theta}_a) = \max_{\mathbf{s},\mathbf{v},\Theta_a} p(\mathbf{y}|\mathbf{s},\mathbf{v},\Theta_a) \qquad (4)$$

to estimate the model parameters $\Theta_a = (a^{(1)}, \ldots, a^{(K)})$ of the subsources as well as the sequences of states $\mathbf{s} = (s_1, \ldots, s_T)$ and residual variances $\mathbf{v} = (v_1, \ldots, v_T)$ that correspond to the observed sequence of speech frames. Using the approximation (1), the logarithmic likelihood function becomes

$$\begin{aligned} \log p(\mathbf{y}|\mathbf{s},\mathbf{v},\Theta_a) &= -\frac{MT}{2}\log 2\pi \qquad (5) \\ &\quad -\sum_{t=1}^{T}\Big(\frac{M}{2}\log v_t + \frac{M}{2v_t} Q(y_t; a^{(s_t)})\Big). \end{aligned}$$

For fixed $\mathbf{s}$ and Θ_a, the residual variances are simply estimated by

$$v_t = Q(y_t; a^{(s_t)}), \quad t = 1, \ldots, T, \qquad (6)$$

leading to

$$\begin{aligned} \max_{\mathbf{v}} \log p(\mathbf{y}|\mathbf{s},\mathbf{v},\Theta_a) &= \qquad (7) \\ -\frac{MT}{2}\log 2\pi e - \frac{M}{2}\sum_{t=1}^{T} \log Q(y_t; a^{(s_t)}) \end{aligned}$$

This can be maximized with respect to $\mathbf{s}$ by choosing s_t for each t separately according to

$$Q(y_t; a^{(s_t)}) = \min_k Q(y_t; a^{(k)}). \qquad (8)$$

For fixed $\mathbf{s}$ and $\mathbf{v}$, maximizing (5) with respect to Θ_a, i.e. $a^{(k)}$, $k = 1, \ldots, K$, leads to the normal equations

$$\sum_{j=0}^{p} a_j^{(k)} \sum_{t=1}^{T} \delta_{s_t,k} \frac{c_{i-j}(y_t)}{v_t} = 0, \qquad i = 1, \ldots, p \qquad (9)$$

which are based on the short-time autocovariance $c_i(y_t)$, normalized by v_t, and averaged over all frames y_t having labels $s_t = k$ (indicated by Kronecker-δ).

Using an arbitrary initial parameter set $\tilde{\Theta}_a^{[0]}$, an iterative algorithm solves the ML-approch (4) by the equations

$$p(\mathbf{y}|\tilde{\mathbf{s}}^{[n]}, \tilde{\mathbf{v}}^{[n]}, \tilde{\Theta}_a^{[n-1]}) = \max_{\mathbf{s},\mathbf{v}} p(\mathbf{y}|\mathbf{s},\mathbf{v},\tilde{\Theta}_a^{[n-1]}), \qquad (10)$$

$$p(\mathbf{y}|\tilde{\mathbf{s}}^{[n]}, \tilde{\mathbf{v}}^{[n]}, \tilde{\Theta}_a^{[n]}) = \max_{\Theta_a} p(\mathbf{y}|\tilde{\mathbf{s}}^{[n]}, \tilde{\mathbf{v}}^{[n]}, \Theta_a). \qquad (11)$$

They are computed by first optimizing jointly with respect to s_t (8) and v_t (6) for each $t = 1, \ldots, T$ and then updating the $a^{(k)}$ by solving the system of equations (9), which is linear for fixed $\mathbf{s}$ and $\mathbf{v}$.

As already shown in [4], there is a strong relationship between ML estimation for CS models, cluster analysis, and vector quantization using the Itakura-Saito distance (ISD)

$$d_{\mathrm{IS}}(y_t; a, v_t) = \frac{Q(y_t; a)}{v_t} + \log v_t - \log v(y_t) - 1. \qquad (12)$$

Here, $v(y_t) = \min_a Q(y_t; a)$ is the residual variance of the optimal predictor for frame y_t. The gain-adaptive CS model leads directly to the gain-optimized ISD

$$d_{\mathrm{GO}}(y_t; a) = \min_{v_t} d_{\mathrm{IS}}(y_t; a, v_t) = \log Q(y_t; a) - \log v(y_t). \qquad (13)$$

The ML-approach (4) is equivalent to the minimization of the average gain-optimized Itakura-Saito distance

$$\begin{aligned} \overline{d_{\mathrm{GO}}} &= \frac{1}{T}\sum_{t=1}^{T} d_{\mathrm{GO}}(y_t; a^{(s_t)}) \qquad (14) \\ &= \frac{1}{T}\sum_{t=1}^{T} \log Q(y_t; a^{(s_t)}) - \frac{1}{T}\sum_{t=1}^{T} \log v(y_t), \end{aligned}$$

because the right-hand sides of (14) and (7) differ only by a factor and some constants, that neither depend on $\mathbf{s}$ nor on Θ_a. In the usual VQ-design for the gain-optimized ISD, i.e. minimizing (14) alternately with respect to $\mathbf{s}$ and Θ_a, the computation of centroids requires a nonlinear system of equations (cf. (9) after substituting v_t from (6)). The usual way of circumventing this problem [7] is to approximate the gain-optimized ISD by the gain-normalized ISD

$$d_{\mathrm{GN}}(y_t; a) = d_{\mathrm{IS}}(y_t; a, v(y_t)) = \frac{Q(y_t; a)}{v(y_t)} - 1. \qquad (15)$$

The algorithm given by (10) and (11), which is equivalent to VQ-design by the 'individually optimized algorithm' [6], avoids this approximation without substantial increase in computational cost.

4 Resulting performance bounds for adaptive speech coding

In our previous paper [4], the average Itakura-Saito distance could be interpreted in such a way that it describes the asymptotic behaviour of the rate distortion function with regard to the mean squared error distortion measure. This fact is illustrated in Fig. 2 by an additional axis, which is scaled in units of the ISD. The reference point of this scale is the asymptote of the conditional RDF

$$R_{Y|s}^{K=T}(D) \geq \frac{1}{2}\frac{1}{T}\sum_{t=1}^{T}\log v(y_t) - \frac{1}{2}\log D. \qquad (16)$$

of a somewhat extreme CS model, which consists of as many subsources as there are frames. For a given number of subsources, the minimization of the Itakura-Saito distance measure leads to a composite source model which has the lowest possible conditional RDF in the range of small values of distortion. The resulting performance bound for 8 subsources is given in Fig. 2 corresponding to $\overline{d_{IS}} = 0.873$ [4]. The model is based on about 60 sec. of speech (male speaker), sampled at a rate of 16,000 samples per second ($M = 160$) and normalized to a variance of 1.

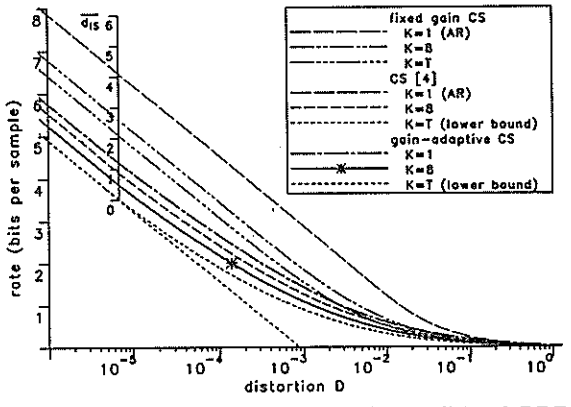


Fig. 2: Quantitative description of the conditional RDF asyptote by the (gain-optimized) Itakura-Saito distance.

Since the gain-adaptive CS model is a special case of the usual CS model, the same method as in [4] may be used in computing its RDF. Averaging the RDF for the Gaussian models of all frames at points of equal slope, we get the conditional RDF given s and v, which is, in the asymptotic range, very tight to the bound

$$R_{Y|s,v}(D) \geq \frac{1}{2}\frac{1}{T}\sum_{t=1}^{T}\log Q(y_t; a^{(s_t)}) - \frac{1}{2}\log D. \qquad (17)$$

Using (14) or (7), the right-hand side of (17) can be expressed in terms of $\max_v \log p(y|s,v,\Theta_a)$ and $\overline{d_{GO}}$, respectively. From this it becomes obvious that the solution of the combined estimation problem (4) gives the lowest possible conditional RDF in the asymptotic range for all state sequences s. The average gain-optimized ISD establishes the position of the asymptotic range of the conditional RDF, given s and v. In Fig. 2, $R_{Y|s,v}(D)$ for a gain-adaptive CS with $K = 8$ is plotted. It corresponds to $\overline{d_{GO}} = 0.484$, which can be measured by the ISD-axis, because the reference point is still given by (16). It must be noted, that in this case the variance of the model obtained by ML estimation is not equal to the average sum of squares of the signal, so that the curve reaches the distortion axis beyond $D = 1$.

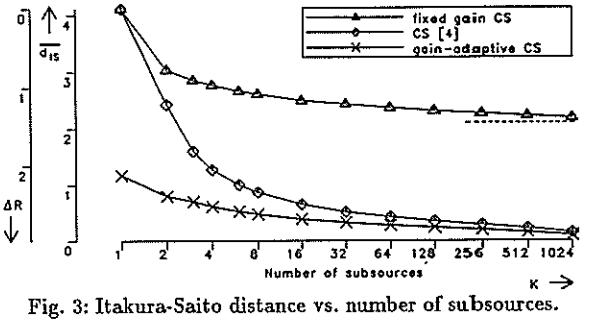


Fig. 3: Itakura-Saito distance vs. number of subsources.

Fig. 3 depicts the minimum average (gain-optimized) ISD attainable for a given speech signal by a (gain-adaptive) CS having K subsources. For fixed K, $\overline{d_{GO}}$ is clearly lower than $\overline{d_{IS}}$ due to the fact, that there is the additional freedom of adapting the residual variance frame-by-frame. For $K \to \infty$, both of them approach the limiting value 0. It is worth noticing, that even for $K = 1$, i.e. a fixed predictor, the gain-optimized ISD is significantly lower than the ISD. This fact clearly stresses the importance of gain-adaption.

The structure of the CS model can also be used in a switched coding scheme, i.e., a blockwise forward adaptive waveform coder with finitely many adaptation states [5]. The important design parameters of the coder, eg. size and order of a filter codebook, can be mapped into corresponding modelling assumptions, here K and p, resulting in a variety of models having different complexity. The comparison of their RDFs gives hints about suitable choices for these parameters. Analogously, the optimality of state estimation is tranferred to a switched coding scheme, if it selects a filter and a gain factor by using the ISD and applies a time-varying transmission rate depending on the state [4].

These implications are extended here for gain-adaptive CS models. The performance bound (17) is valid for the main information in adaptive coding schemes with K filters and a transmission of both the state s_t and the residual variance v_t as side information.

So the gain-optimized ISD is suitable for forward adaptive coding schemes, which are blockwise gain-adaptive. For a speech frame y_t, the predictor $a^{(s_t)}$ is determined according to (8) and the residual variance v_t may be computed by (6). In the rate distortion computation, different subsource rates are implicitely assumed for each frame, depending on v_t. For a gain-adaptive CS (K=8), those subsource rates which result in a composite rate of 2 bits per sample (marked point in Fig. 2) are plotted in Fig. 4, forming one line for each subsource. An adaptive

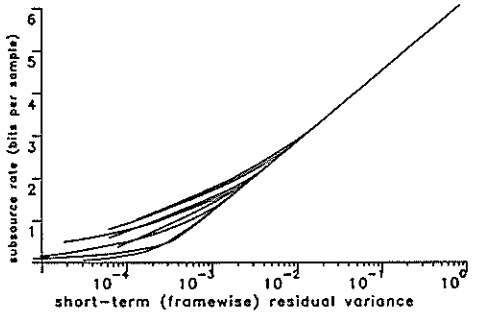


Fig. 4: Subsource rates for an average rate of 2 bits per sample (32 kbits/sec).

coding scheme takes advantage from this computation by encoding a speech frame y_t using the transmission rate that is found in Fig. 4 for s_t and v_t. This rule for adjusting a time-varying transmission rate is a combination of the methods for a CS model presented in [4] and for a spherically invariant random process [8], which can be regarded as a gain-adaptive CS with a fixed predictor ($K = 1$). An analog bit assignment technique for transform coding has been proposed in [9].

5 CS model with fixed gain

Practical considerations often require a fixed transmission rate even in the case of an adaptive source coding scheme. For applications like this, we propose a CS model, that is restricted to the case of equal residual variances v_t. Maximizing (5) with respect to s and Θ_a is now equivalent to the minimization of the average residual variance

$$\overline{Q} = \frac{1}{T}\sum_{t=1}^{T} Q(y_t; a^{(s_t)}), \qquad (18)$$

the solution being independent of the specific value $v = v_t$. The decision rule (8) remains valid and the normal equations (9) are simplified, because the normalization by v_t is obsolete now. Using these equations alternately, $\overline{Q}$ will be minimized. This algorithm may also be seen as vector quantization using directly $Q(y_t; a)$ as distortion measure and (9) with $v_t = v$ as the centroid computation. (As far as we know, this simple method has not been proposed up to now.)

Setting the derivative of (5) to 0, the ML-estimate of v ($v = v_t$) is simply $v = \overline{Q}$. Rate distortion functions for this model are plotted in Fig. 2 for $K = 8$ and $K = T$, i.e. using the optimal predictor of each frame. In both cases, the curves are significantly above the ones of CS models which use separate residual variances for the subsources. In the asymptotic range, all subsource RDFs are identical, so that a constant transmission rate is supposed by this model. In this sense, the reduction in transmission rate achievable by fixed rate adaptive coding is depicted in Fig. 3, clearly recommending the use of a time-varying transmission rate for speech signals.

6 Conclusions

Gain-adaptive CS models are capable of representing time-varying spectral shape and gain in an efficient manner. They provide usefull rate distortion bounds for waveform speech coding. Since the side information of forward adaptive coding schemes is explicitly modeled in the CS, different adaptation techniques may be compared on the basis of computed bounds. A relationship between the gain-optimized version of the Itakura-Saito distance measure and rate distortion theory for composite source models is presented, which states the optimality of this distance for LPC-parameter quantization in a class of coding schemes. Furthermore, a rule for adjusting the transmission rate according to the side information is developed.

References

[1] T. Berger, *Rate Distortion Theory*, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[2] P. Meissner, "Composite source models for speech: State estimation and Rate-Distortion functions", *ITG-Fachtagung Stoch. Modelle und Meth.*, Nürnberg, April 1989, 73–78.

[3] H. Kalveram, P. Meissner, "Rate distortion bounds for speech waveforms based on Itakura-Saito segmentation", *EUSIPCO*, Grenoble 1988, in J. L. Lacoume et al. (eds.), *Signal Processing 4: Theories and Applications*, 137–140.

[4] H. Kalveram, P. Meissner, "Itakura-Saito clustering and rate distortion functions for a composite source model of speech", *Signal Processing*, Vol. 18 (1989), 195–216.

[5] D. Chahabadi, P. Meissner, "Differentielle Vektorquantisierung mit variabler Codewortlänge auf der Grundlage eines Composite Source Modells", *Proc. ITG-Fachtagung Digitale Sprachverarbeitung*, Bad Nauheim, 1988.

[6] M. J. Sabin, R. M. Gray, "Product code vector quantizers for waveform and voice coding", *IEEE Trans. Acoust., Speech and Signal Process.*, Vol. ASSP-32, 1984, 474–488.

[7] R. M. Gray, A. Buzo, A. H. Gray, Y. Matsuyama, "Distortion measures for speech processing", *IEEE Trans. Acoust., Speech and Sign. Process.*, Vol. ASSP-28, 1980, 367–376.

[8] H. Brehm, K. Trottler, "Adaptive vector quantization of band-limited speech and speech-like waveforms" *EUSIPCO*, Grenoble 1988, in J. L. Lacoume et al. (eds.), *Signal Processing 4: Theories and Applications*, 887-890.

[9] N. Farvardin, Y. Hussain, "Adaptive block transform coding of speech based on the Hidden Markov model" *EUSIPCO*, Grenoble 1988, in J. L. Lacoume et al. (eds.), *Signal Processing 4: Theories and Applications*, 883-886.

This work has been supported by the Deutsche Forschungsgemeinschaft.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

ENHANCED ADPCM TREE CODEC AT 16 AND 9.6 KBIT/S

Flávia Martinho Ferreira; José Sindi Yamamoto; Fábio Violaro(*)

CPqD-TELEBRÁS
P.O. Box 1579
Campinas-SP-13085
Brazil

(*)UNICAMP
FEE/DECOM
P.O. Box 6001
Campinas-SP-13801
Brazil

This paper deals with an ADPCM codec at 16 kbit/s based on G.721 algorithm, where the scalar quantizer was substituted by a tree quantizer (*ADPCM Tree Codecs*) and spectral shaping was introduced to enhance the speech quality [1]. The algorithm delay is about 2 msec (16 samples at 8 kHz) and informal subjective tests have shown that the encoder produces high speech quality. This high performance achieved has conducted to an investigation at a bit rate of 9.6 kbit/s.

1. Introduction

The G.721 ADPCM algorithm achieves high speech quality at 32 Kbit/s. However, at 16 Kbit/s it presents a significant degradation. In order to avoid such degradation, the scalar quantizer of G.721 ADPCM was substituted by a tree quantizer.

The main characteristic of tree quantizers [2,3,4] is that, at a given sampling instant, the quantizer does not make a final decision on the quantized amplitude of the difference signal, but examines other M paths over the tree for L sampling instants, in order to obtain a better reconstructed signal. In other words, in the ADPCM G.721 Codec, for an input sample at sampling instant k, only data at sampling instants $j \leq k$ are considered in the encoding process, while in the ADPCM Tree Codec, the encoding decision is delayed of L samples, in order to examine all possible M encoding sequences through time $k+L$ for a best fit to the input signal.

Even using the tree quantizer, the quantization noise is perceptible at 16 Kbit/s and lower rates. In order to reduce such noise perceptibility, two different alternatives of spectral shaping have been introduced : postfiltering [5] and perceptual weighting filter [6].

At the receiver, the decoder is the same of the G.721, except the inclusion of a postfiltering.

First it is described the ADPCM Tree Codecs operating at 16 Kbit/s with the two alternatives of spectral shaping and a combination of them. Then an explanation concerning the way a bit rate of 9.6 Kbit/s was achieved is given. Finally both objetive and subjetive results are presented.

2. 16 Kbit/s ADPCM Tree Codec with postfilter

2.1. Encoder

The block diagram of the encoder is shown in figure 1. The encoder consists basically of a distortion measure, a tree search algorithm and two adaptive structures: an ARMA predictor (pole-zero predictor) and an inverse quantizer.

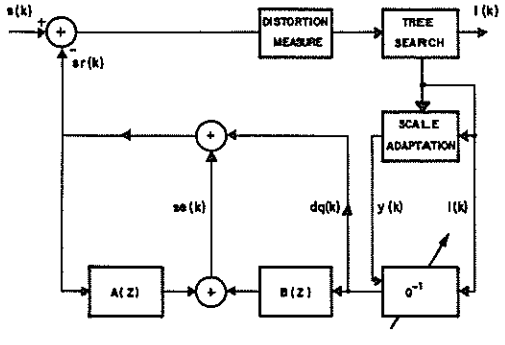


Figure 1: ADPCM Tree Encoder.

- Adaptive Predictor

As mentioned, in this work the scalar quantizer of G.721 was substituted by a tree quantizer. However, all the other features of G.721 encoder such as synthesis configuration and updating of filter coefficients are maintained. Thus, if $dq(k)$ is the quantized difference signal, the basic equations describing the synthesis operation are

$$sr(k) = se(k) + dq(k)$$

and

$$se(k) = \sum_{i=1}^{2} a_i(k-1)sr(k-i) + \sum_{i=1}^{6} b_i(k-1)dq(k-i).$$

The signal estimate $se(k)$ is obtained from two adaptive structures: a second order section that models poles and a sixth order section that models zeroes in the input signal. The coefficients a_i and b_i are updated using a simplified gradient algorithm.

• Adaptive Quantizer

In order to keep a wide dynamic range and minimize complexity, the quantization and its adaptation are performed in the base 2 logarithmic domain. The normalized output inverse quantizer characteristic is shown in table 1.

Table 1: 2 Bits-Normalized Output Inverse-quantizer

$\|I(k)\|$	Normalized Inverse-quantizer Output
1	2.49
0	0.75

The scale factor adaptation is given by

$$y(k) = (1 - 2^{-5})y(k-1) + 2^{-5}W(|I(k)|),$$

where $1.06 \leq y(k) \leq 10.0$ and $W(|I(k)|)$ is an empirical discrete function defined in table 2.

Table 2: Values of $W(|I(k)|)$

$\|I(k)\|$	$W(\|I(k)\|)$
1	22.00
0	-0.25

• Tree Search Algorithm

The tree search block performs the multipath tree search procedure known as *(M,L) Algorithm* [7,8,9,10,11], where M is the maximum number of paths over the tree that are examined and L is the length of these paths.

A distortion between the reference input signal and each possible reconstructed sequence to a depth L in the tree is calculated. Then, the encoding decision is taken with a delay of L samples in order to examine all M possible encoding sequences through L samples for the best fit to the reference input signal in terms of the calculated distortion for each of the M paths.

• Distortion Measure

The distortion measure used in this work is the squared-error given by

$$d(s, sr) = \sum_{i=1}^{L} [s(i) - sr(i)]^2,$$

where $s(i)$ and $sr(i)$ are referred to current values at a depth i in the tree.

2.2. Decoder

The subjective quality of the reconstructed speech may be enhanced by using a spectral postfilter as the final processing step as shown in figure 2. The structure of this tree decoder is the same of the G.721, except the inclusion of the postfilter.

The postfilter is a pole-zero structure whose transfer function is determined from the synthesis filter and it is given by

$$F(z) = \frac{1 + B(z/\beta)}{1 - A(z/\alpha)},$$

where

$$A(z/\alpha) = \sum_{i=1}^{2} a_i \alpha^i z^{-i}, \quad 0 \leq \alpha \leq 1,$$

and

$$B(z/\beta) = \sum_{j=1}^{6} b_j \beta^j z^{-j}, \quad 0 \leq \beta \leq 1.$$

The best performance was achieved empirically, considering the tradeoff between noise reduction and signal distortion, by choosing $\alpha = 0.07$ and $\beta = 1.0$. The effect of increasing the perceived loudness due to the postfiltering is compensated by a factor of about 2.2.

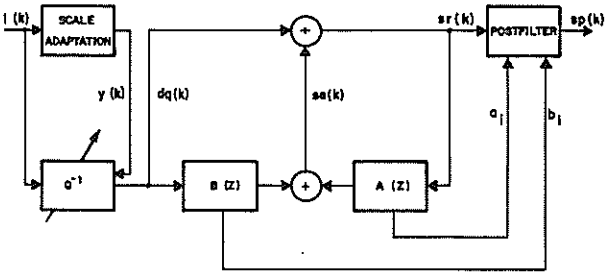


Figure 2: ADPCM Tree Decoder with postfilter.

3. 16 Kbit/s ADPCM Tree Codec with a perceptual weighting filter

3.1. Encoder

It is well known that the error signal, simply calculated as $s(k) - sr(k)$, is not a valid measure of the perceptual difference between the original and the reconstructed speech signals. To obtain a better measure of this difference, a perceptual weighting filter is introduced in the encoder before performing the distortion measure as shown in figure 3.

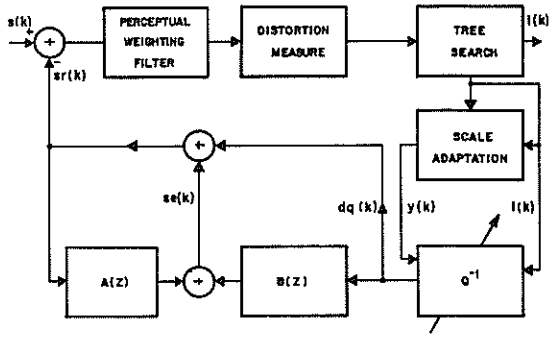


Figure 3: ADPCM Tree Encoder with perceptual weighting filter.

The weighting filter is a pole-zero structure whose transfer function is determined from the synthesis filter and it is given by

$$F_w(z) = \frac{1 - A(z/\alpha_1)}{1 - A(z/\alpha_2)} \frac{1 + B(z/\alpha_2)}{1 + B(z/\alpha_1)}, \quad 0 < \alpha_2 < \alpha_1 < 1,$$

where

$$A(z/\alpha_1) = \sum_{i=1}^{2} a_i \alpha_1^i z^{-i}, \quad A(z/\alpha_2) = \sum_{i=1}^{2} a_i \alpha_2^i z^{-i},$$

and

$$B(z/\alpha_1) = \sum_{j=1}^{6} b_j \alpha_1^j z^{-j}, \quad B(z/\alpha_2) = \sum_{j=1}^{6} b_j \alpha_2^j z^{-j}.$$

The best performance was achieved empirically by choosing $\alpha_1 = 0.94$ and $\alpha_2 = 0.70$.

3.2. Decoder

The structure of this decoder is the same as the one of G.721, i.e., it is iqual to figure 2 with the postfilter removed.

4. 16 Kbit/s ADPCM Tree Codec with postfilter and perceptual weighting filter

Although a significant improvement is achieved by using either the postfilter or perceptual weighting filter, a further speech quality improvement may be obtained with a combination of them. In this case, the codec consists of a coder and decoder shown in figures 3 and 2, respectively.

5. 9.6 Kbit/s ADPCM Tree Codec

The tree coding delay allows all possible M encoding sequences through time $k+L$ to be examined for a best fit of the input signal. Each different encoding sequence is called a path and therefore the tree coding is a multipath search procedure.

In order to achieve a bit rate of 9.6 kbit/s, the concept of *multi-tree source code* [12] is used. It consists of different rate trees interleaved with each other.

In figure 4, the tree structure for 9.6 Kbit/s is shown. Defining as supernodes the nodes of the tree where the multi-tree structure repeats, the rate of a tree code in bits/symbol is given by

$$R = \frac{1}{\beta_t} \log_2 \alpha_t,$$

where α_t is the number of paths between supernodes and β_t is the number of symbols per path between supernodes.

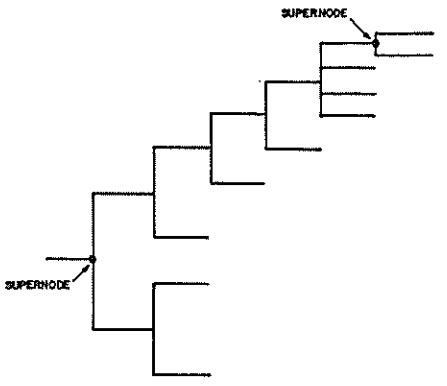


Figure 4: Tree structure for 9.6 Kbit/s.

For 9.6 Kbit/s, $\alpha_t = 64$ and $\beta_t = 5$, so $R = 6/5$ bits/symbol. Since the sampling frequency is 8000 Hz, a bit rate of 9.6 Kbit/s is achieved.

Concerning the codec structure, it is maintained the same as the 16 kbit/s ADPCM Tree Codec with a inclusion of an additional inverse-quantizer and its scale factor adaptation for the paths with 1 bit/symbol.

The 1 bit/symbol normalized output inverse-quantizer characteristic is shown in table 3.

Table 3: 1 Bit-Normalized Output Inverse-quantizer

$I(k)$	Normalized Inverse-quantizer Output
1	1.562
0	-1.562

The scale factor adaptation is given by

$$y(k) = \begin{cases} (1 - 2^{-7})y(k-1) + 2^{-7}W(I(k) = 1), \\ \quad \text{if } sgn(dq(k-1)).sgn(dq(k)) = 1 \\ \\ (1 - 2^{-7})y(k-1) + 2^{-7}W(I(k) = 0), \\ \quad \text{if } sgn(dq(k-1)).sgn(dq(k)) = -1 \end{cases}$$

where $1.06 \leq y(k) \leq 12.0$ and $W(I(k))$ is an empirical discrete function defined in Table 4.

Table 4: Values of $W(I(k))$

$I(k)$	$W(I(k))$
1	19.0
0	-7.0

For paths with 2 bits/symbol, the inverse-quantizer characteristic and the scale adaptation are the same as for 16 kbit/s. Also, at 9.6 kbit/s, the postfilter and the perceptual weighting filter are implemented as at 16 kbit/s.

6. Performance Results

In order to establish the performance attainable with the ADPCM Tree Codec as described in the previous sections, objective measures in terms of Segmental Signal-to-Noise Ratio and informal subjective testing were conducted at three different input speech levels: -12 dB, -22 dB and -32 dB below the overload. The objective results of the ADPCM tree codec in comparison with 32 kbit/s ADPCM (G.721) are presented in table 5 for weighted (W) and unweighted (UW) cases without postfiltering .

Table 5: Values of Segmental SNR (dB)

CODEC	Input Level (dB)		
	-32	-22	-12
32 Kbit/s (G.721)	19.93	17.16	11.75
16 Kbit/s (UW)	10.81	11.04	10.17
16 Kbit/s (W)	10.38	10.69	9.87
9.6 Kbit/s (UW)	5.09	5.27	4.79
9.6 Kbit/s (W)	4.91	4.95	4.53

The correspondent informal subjective testing results in terms of Mean Opinion Score (MOS) are presented in table 6. Additonal MOS results are presented for the cases with postfilter (P) and combination of weighting filter with postfilter (W+P).

From these results, it is evident the high performance of the ADPCM Tree Codec at 16 kbit/s with spectral shaping. On the other hand, there is an audible loss of quality when comparing 9.6 kbit/s and 16 kbit/s, as it is indicated in table 5 as well as in table 6. However, the intelligibility at 9.6 kbit/s continues quite good with a low level of granular noise.

Table 6: Values of MOS

CODEC	Input Level (dB)		
	-32	-22	-12
32 Kbit/s (G.721)	4.5	4.2	3.5
16 Kbit/s (UW)	3.6	4.0	4.0
16 Kbit/s (W)	3.5	3.6	4.1
16 Kbit/s (P)	3.9	4.1	4.0
16 Kbit/s (W + P)	4.0	3.5	4.3
9.6 Kbit/s (UW)	2.3	2.4	2.2
9.6 Kbit/s (W)	2.4	2.2	2.3
9.6 Kbit/s (P)	2.5	2.8	2.5
9.6 Kbit/s (W + P)	2.4	2.4	2.5

7. Conclusions

This work has shown that a toll quality and low delay 16 kbit/s codec can be implemented using the G.721 structure with the substitution of the scalar quantization by a tree quantization and incorporating spectral shaping features. Additionally, simulations at 9.6 kbit/s were conducted in order to have an idea of an attainable bit rate maintaining a speech quality ranging from toll quality to good quality and intelligibility. At 9.6 kbit/s it has shown that a good inteligibility is maintained with a low level of granular noise. However, a significant degradation in speech quality has been perceived.

References

[1] CPqD-Telebrás, "16 kbit/s ADPCM with Tree Quantizer", Contribution on CCITT Group XV- Ad Hoc Group on 16 kbit/s Speech Coding, July, 1989.

[2] J. D. Gibson, G. B. Haschke, "Adaptive Code Generators for Tree Coding of Speech", Proc. IEEE Int. Conf. Comm., June, 1987.

[3] J. D. Gibson, G. B. Haschke, "Backward Adaptive Tree Coding of Speech at 16 kbps", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, April, 1988.

[4] V. Iyengar, P. Kabal, " A Low Delay 16 kbit/sec Speech Coder", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, April, 1988.

[5] N. S. Jayant,V. Ramamoorthy, "Adaptive Postfiltering of 16 kb/s ADPCM Speech", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, April, 1986.

[6] Simon Fraser University, University of California, Voicecraft, Inc., "Description of the Consortium's Low Delay Vector Excitation Coder (LD-VXC), Version 2", CCITT Group XV- Ad Hoc Group on 16 kbit/s Speech Coding, June, 1989.

[7] N. S. Jayant, S. A. Christensen, "Tree-Encoding of Speech Using the (M,L)-Algorithm and Adaptive Quantization", IEEE Trans. on Comm., vol.COM-26, September, 1978.

[8] J. B. Anderson, J. B. Bodie, "Tree Encoding of Speech", IEEE Trans. on Inform. Theory, vol. IT-21, July, 1975.

[9] T. Svendsen, "Tree Encoding of LPC Residual", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, April, 1984.

[10] N. S. Jayant , P. Noll, "Digital Coding of Waveforms-Principles and Applications to Speech and Video", pp. 443-444, Englewood Cliffs, Prentice Hall,1984.

[11] H. G. Fehn, P. Noll, "Multipath Search Coding of Stationary Signals with Applications to Speech", IEEE Trans. on Comm., vol.COM-30, April, 1982.

[12] J. D. Gibson, W. W. Chang, "Fractional Rate Multi-Tree Speech Coding", Department of Electrical Engineering, Texas A&M University.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

COMBINED SPEECH AND CHANNEL CODING AT 11.2 KBPS

Ira A. Gerson, Mark A. Jasiuk, Michael J. McLaughlin, and Eric H. Winter
Chicago Corporate Research and Development Center
Motorola Inc.
1301 E. Algonquin Road, Schaumburg, IL 60196

Digital speech coding will be widely used for future land-mobile radio systems. In particular, next generation cellular telephone systems in Europe, North America and Japan will incorporate digital speech transmission. Digital land-mobile radio channels introduce channel errors due to the effects of severe attenuation, co-channel and adjacent channel interference and Rayleigh fading. These channel errors can severely impact the quality of the received speech if the speech coder is not properly designed and without the use of error mitigation techniques.

A combined speech and channel coding system has been developed which offers excellent speech quality and very good robustness to channel errors. The basic speech coder was designed so that error protection schemes could be employed efficiently. Many of the speech coder bits are inherently robust to channel errors. The coder is also designed to limit error propagation of channel errors in the speech decoder.

Several types of error mitigation mechanisms are employed in conjunction with the basic speech coder. They include forward error correcting codes, interleaving and error detection.

A total of 11.2 kbps are used for the combined speech and channel coder. 6.7 kbps are used for the basic speech coder with an additional 4.5 kbps used for error correction and detection.

I. INTRODUCTION

A combined speech/channel coder has been designed for application to land-mobile radio channels. The total rate of the speech/channel coder is 11.2 kbps. The coder has been designed to provide excellent speech quality and robustness to channel errors. This speech/channel coder was submitted as a candidate coder for the Japan Digital Cellular system. The Japan Digital Cellular system will be a TDMA (time division multiple access) system with three TDMA speech channels per radio channel. The TDMA structure defines 20 msec. speech frames which are coded and transmitted over a 6.7 msec burst. The total data rate for each speech channel for speech and error control is 11.2 kbps or 224 bits per 20 msec of speech.

The speech coder employed is a Vector Sum Excited Linear Prediction (VSELP) coder. VSELP falls into the class of speech coders known as Code Excited Linear Prediction (CELP) (also called Vector Excited or Stochastically Excited) [1,4,5]. The VSELP speech coder was designed to accomplish three goals:

1. Highest possible speech quality
2. Reasonable computational complexity
3. Robustness to channel errors

The VSELP speech coder achieves these goals through efficient utilization of a structured excitation codebook. The structured codebook reduces computational complexity and increases robustness to channel errors [2,3]. A novel gain quantizer is also employed which achieves high coding efficiency while providing robustness to channel errors.

The output data stream of the VSELP speech coder needs to be provided some means of protection for transmission over the radio channel. The radio channel may produce errors in the speech coder data. These errors are the result of the radio channel experiencing severe attenuation, co-channel and adjacent channel interference and the effects of Rayleigh fading. The effect of these channel errors is to degrade the quality of the speech at the receiver. Because the overall performance depends on how well the speech coder and channel coder can withstand channel errors, it is important that channel coder make the speech coder as robust as possible.

The channel error control for the speech coder data employs three mechanisms for the mitigation of channel errors. The first is to use a forward error correcting (FEC) code to protect the more vulnerable bits of the speech coder data stream. The second technique interleaves the transmitted data for each speech coder frame over two time slots to mitigate the effects of Rayleigh fading. The third technique employs the use of a cyclic redundancy check (CRC) to ensure detection of bit errors over the most perceptually significant bits.

II. BASIC SPEECH CODER STRUCTURE

Figure 1 is a block diagram of the VSELP speech decoder. The 6.7 kbps VSELP coder/decoder utilizes two excitation sources. The first is from the long term ("pitch") predictor state, or adaptive codebook [4]. The second source is from the VSELP excitation codebook. The VSELP codebook contains the equivalent of 512 vectors. These two excitation sources are multiplied by their corresponding gain terms and summed. This becomes the combined excitation sequence ex(n). After each subframe, ex(n) is used to update the long term filter state (adaptive codebook). The synthesis filter is a direct form 10th order LPC all-pole filter. The LPC coefficients are coded once per 20 msec frame and updated in each 5 msec subframe through interpolation. The excitation parameters are also updated in each 5 msec subframe. The number of samples in a subframe, N, is 40 at an 8 kHz sampling rate. The "pitch" prefilter and spectral postfilter are adaptive filters which are used to enhance the quality of the reconstructed speech [2].

Table 1 shows the bit allocations for the 6.7 kbps VSELP coder. The 10 LPC coefficients are coded using scalar quantization of the reflection coefficients. An energy term, $R_q(0)$, which represents the average speech energy per frame is also coded once per frame. The two excitation gains are vector quantized to 7 bits (GS-P0 code) per subframe. The gain quantizer will be described in Section IV.

III. VSELP CODEBOOK STRUCTURE

The coder uses a VSELP excitation codebook containing 2^M codevectors. These are constructed from a set of M basis vectors, where M = 9. Defining $v_{k,m}(n)$ as the m^{th} basis vector of the k^{th} codebook and $u_{k,i}(n)$ as the i^{th} codevector in the k^{th} codebook, then:

$$u_{k,i}(n) = \sum_{m=1}^{M} \theta_{im} v_{k,m}(n) \qquad (1)$$

where k = 1 or 2 for the first or second VSELP codebook, $0 \le i \le 2^M-1$, and $0 \le n \le N-1$.

In other words, each codevector in the codebook is constructed as a linear combination of the M basis vectors. The linear combinations are defined by the θ parameters. θ_{im} is defined as:

$$\theta_{im} = +1 \text{ if bit m of codeword i} = 1$$

$$\theta_{im} = -1 \text{ if bit m of codeword i} = 0$$

Note that if we complement all the bits in codeword i, the corresponding codevector is the negative of codevector i. Therefore, for every codevector, its negative is also a codevector in the codebook. These pairs are called complementary codevectors since the corresponding codewords are complements of each other.

The excitation codewords for the VSELP coder are more robust to bit errors than the excitation codewords for random codebooks. A single bit error in a VSELP codeword changes the sign of only one of the basis vectors. The resulting codevector is still similar to the desired codevector.

The VSELP codebook structure also allows for a very fast search procedure [2,3] in addition to other advantages [2].

IV. QUANTIZATION OF EXCITATION GAINS

The quantization of the excitation gains consists of two stages. The first stage codes the average speech energy once per frame. The quantized value of this energy, $R_q(0)$, is specified with five bits, using 2 dB quantization steps for 64 dB of dynamic range. In the second stage, a GS-P0 code is selected every subframe. This code, when taken in conjunction with $R_q(0)$ and the state of the speech decoder, determines the excitation gains for the subframe. The selection of the GS-P0 code takes place after the two excitation vectors, L and I, have been chosen.

The following definitions are used to determine the GS-P0 code. The combined excitation function, ex(n), is given by:

$$ex(n) = \beta\, c_0(n) + \gamma\, c_1(n) \qquad 0 \le n \le N-1 \qquad (2)$$

where:

$c_0(n)$ is the long term prediction vector, $b_L(n)$

$c_1(n)$ is the codevector selected from the VSELP codebook, $u_I(n)$

The energy in each excitation vector is given by:

$$R_x(k) = \sum_{n=0}^{N-1} c^2_k(n) \qquad k = 0,1 \qquad (3)$$

Let RS be the approximate residual energy at a given subframe. RS is a function of N, $R_q(0)$, and the normalized prediction gain of the LPC filter. It is defined by:

$$RS = N\, R_q(0) \prod_{i=1}^{N_p} (1-r_i^2) \qquad (4)$$

where r_i is the i^{th} reflection coefficient corresponding to the set of direct form filter coefficients (α_i's) for the subframe. GS, the energy offset , is a coded parameter which refines the estimated value of RS. R, the approximate total subframe excitation energy, is defined as:

$$R = GS\, RS \qquad (5)$$

P0, the approximate energy contribution of the long term prediction vector as a fraction of the total excitation energy at a subframe, is defined to be:

$$P0 = \frac{\beta^2 R_x(0)}{R} \qquad \text{where } 0 \le P0 \le 1 \qquad (6)$$

Thus β and γ are replaced by two new parameters: GS and P0. The transformations relating β and γ to GS and P0 are given by:

$$\beta = \sqrt{\frac{RS\, GS\, P0}{R_x(0)}} \qquad (7)$$

$$\gamma = \sqrt{\frac{RS\, GS\, (1-P0)}{R_x(1)}} \qquad (8)$$

The GS-P0 pair is vector quantized using a codebook of 128 vectors. The codebook was designed using the LBG algorithm [6], with the normalized weighted error as the distortion criterion.

The vector, which minimizes the total weighted error energy for the subframe, is chosen from the GS-P0 codebook. The codebook search procedure requires only five multiply-accumulates per vector evaluation.

This technique of quantizing the gains has a number of advantages. First, the coding is efficient. The coding of the energy once per frame solves the dynamic range issue. The gain quantization performs equally well at all signal levels within the range of the $R_q(0)$ quantizer. With the average energy factored out, the two gains can be vector quantized efficiently. In minimizing the weighted error, the vector quantizer takes into account the correlation between the two weighted excitation vectors. Second, the values of GS and P0 are well behaved. Whereas the optimal value for β, the adaptive codebook gain, can occasionally get very large, P0 is bounded by 0 and 1. Error propagation effects are also greatly reduced by this quantization scheme. Since the energies in the excitation vectors are used to normalize the excitation gains, previous channel errors affecting the energy in the adaptive codebook vector have very little effect on the decoded speech energy. Channel errors in the LPC coefficients are also automatically compensated for at the decoder in calculating the excitation gains. In fact as long as the code for the average frame energy, $R_q(0)$, is received correctly, the speech energy at the decoder will not be much greater than the desired energy and no "blasting" will occur.

V. ERROR CORRECTION

Informal listening tests demonstrate that a bit error's impact is related to which bit of the speech coder data is corrupted. Some bits are inherently robust to bit errors while others are highly sensitive. The 134 speech coder bits are partitioned into two classes. Class 1 which is protected by a FEC code and class 2 which remains unprotected. The process is depicted in Figure 2.

The first step in the error protection process is the separation of the 134 bit speech coder frame's information into class 1 and class 2 bits. The 75 bits which make up class 1 are as follows: 15 of the 37 LPC bits, 4 of the 5 $R_q(0)$ bits, all of the lag bits and all of the GS-P0 bits. The remaining 59 bits fall into class 2 namely: all of the codevectors, 22 of the LPC bits and 1 of the $R_q(0)$ bits.

A 7 bit CRC is used for error detection purposes and is computed over the 44 most perceptually significant bits of the class 1 bits for each frame. These 44 bits are a subset of the class 1 bits and include: 9 LPC bits, 3 $R_q(0)$ bits, all of the lag bits and the most significant bit of GS-P0.

The type of error correction code used is rate 9/17 convolutional code. The code is derived from a rate 1/2 code using the puncturing matrix a [10].

$$a = \begin{bmatrix} 1111\ 1111\ 1 \\ 1111\ 1111\ 0 \end{bmatrix}$$

The generators of the memory order 5 code are $g_0 = 65$ and $g_1 = 57$ following the notation in [11]. Using this matrix results in a total of 9 punctures of the rate 1/2 code. There are 87 bits input into the encoder: 75 class 1 bits, 7 CRC bits and 5 tail bits. The tail bits are needed to drive the final state of the encoder back to the initial state. There are 165 bits output by the encoder. The decoding of the convolutional code is done using a Viterbi algorithm (VA). Paths are truncated at 29 bits. The output bit decision is reached using a majority vote [11].

The ordering of bits to be encoded places the most perceptually significant bits at the two extremes of the trellis. The ends of the trellis are better protected than the middle section, so the most perceptually significant bits are provided the best protection. Figure 3 illustrates this effect. 10,000 frames each with 82 information bits and 5 tail bits per frame

were encoded using the rate 9/17 code. The encoded bits where then corrupted with random errors at an average BER of .079. The corrupted data was then decoded using the Viterbi algorithm with a decoding lag of 53. Figure 3 shows the resulting number of errors after decoding for each information bit as a function of position in the trellis (order into the encoder).

VI. INTERLEAVING

Interleaving provides significant improvement in error control for fading or bursty channels. By time multiplexing one frame of data over more than one TDMA time slot, time diversity is achieved. The effect of the time diversity is to spread a burst of errors over more than one frame, significantly reducing the peak bit error rate.

The implementation that is used is a variation on traditional diagonal interleaving. Interleaving is done over two frames so that every TDMA time slot contains information from two consecutive frames (x and y). See Figure 2.

The type of error correcting code must be taken into account when designing the interleaving. Convolutional codes have the property of being self synchronizing. That is, if the metrics and states of the convolutional code are arbitrary, after 3 to 4 constraint lengths of valid data the VA will converge and produce valid output. This is significant since a burst will cause a similar situation. Thus the 87 data bits may be viewed as being made up of 3 to 4 independent sections. Stated alternately a bit is not correlated to errors occurring about 24 (4*6) bits back.

Simulations were performed on various types of interleaving schemes. Conventionally interleaving is based on a matrix which is filled row wise with bits and read column wise. This works well enough, but does not exploit the limited range of bit dependencies that convolutional codes have. Errors in the trellis should be separated by as many stages as possible because the decoder is independent of errors which occur far enough back in the trellis. An interleaving scheme was designed which takes advantage of these characteristics.

VII. ERROR DETECTION

Error detection is an essential element of any channel coding scheme. At poor signal conditions the FEC will be unable to correct all the errors introduced. It is vital that errors in certain parameters are caught and not passed to the speech coder.

A CRC code is used for error detection. The CRC is performed over the 44 most perceptually significant bits as mentioned above. The detection is therefore frame based rather than parameter or subframe based. The generator polynomial for the CRC is:

$$g_{crc}(X) = 1+X+X^2+X^4+X^5+X^7.$$

The CRC or parity polynomial b(X) is the remainder of the division of the input polynomial and the parity polynomial.

$$\frac{a(X)*X^7}{g_{crc}(X)} = q(X) + \frac{b(X)}{g_{crc}(X)}$$

Where a(X) is the 44 bit input expressed as a polynomial, q(X) is the quotient of the division, b(X) the remainder. At the receiver b'(X) is generated from the received information and compared to b(X), with any difference flagging an error.

If a frame is found to be in error as indicated by its CRC, the erroneous frame's LPC, $R_q(0)$, GS-P0, and lag values are replaced with those from the previous frame. A single repeated frame is often not noticeable to the listener. When successive frames are found to be in error, the following technique is used. The first two "bad" frames are overwritten with the substituted $R_q(0)$ at its full value. The subsequent third, fourth and fifth consecutive "bad" frames have their energy reduced by 2, 4 and 6 dB respectively from the initial value. Finally, the sixth and subsequent contiguous bad frames are totally muted. Once in the mute state, two consecutive "good" frames (no CRC errors) are required to restore the speech coder's output level. This scheme results in a very reliable error detection circuit which can exist at arbitrarily high error rates with a very low falsing rate.

VIII. REFERENCES

[1] M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates", **Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing**, pp. 937-940, March 1985.

[2] I. Gerson and M. Jasiuk "VSELP Speech Coding at 8kbps", **Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing**, April 1990.

[3] I. Gerson and M. Jasiuk, "Vector Sum Excited Linear Prediction (VSELP)", **IEEE Workshop on Speech Coding for Telecommunications**, pp. 66-68, September 1989.

[4] W. B. Kleijn, D. J. Krasinski and R. H. Ketchun, "Improved Speech Quality and Efficient Vector Quantization in SELP", **Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing**, pp. 155-158, April 1988.

[5] G. Davidson and A. Gersho, "Complexity Reduction Methods for Vector Excitation Coding", **Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing**, pp. 3055-3058, May 1986.

[6] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design", **IEEE Trans. Comm.**, vol. COM-28, pp. 84-95, Jan. 1980.

[7] W. Bastiaan Kleijn, "Source Dependent Channel Coding for Celp", **Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing**, April 1990, S1.1, pp. 1-4.

[8] R. Cox, W. B. Kleijn, P. Kroon, "Robust Celp Coders for Noisy Backgrounds and Noisy Channels", **Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing**, April 1989, S14.2, pp. 739-742.

[9] D. Rahikka, T. Tremain,V. Welch, J. Campbell, "CELP Coding for Land Mobile Radio Applications", **Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing**, April 1990, S9.4, pp. 465-468.

[10]Joachim Hagenauer, "Rate Compatible Punctured Convolutional Codes and their Applications", **IEEE Trans. on Comm.**, Vol 36, No. 4, April 1988, pp. 389-400.

[11]Shu Lin and Daniel Costello, **Error Control Coding: Fundamentals and Applications**, Prentice Hall, 1983

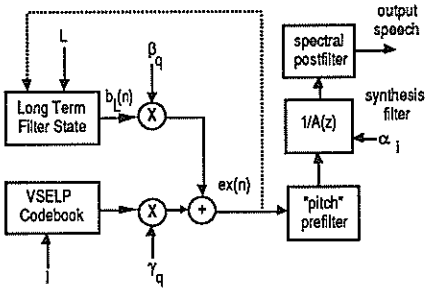
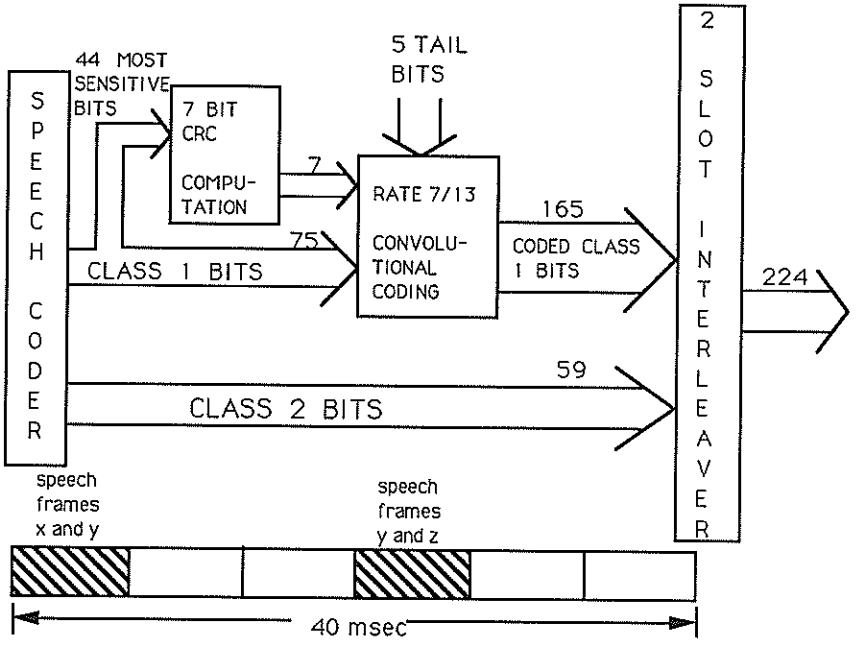


Figure 1 – VSELP Speech Decoder

PARAMETER	BITS/SUBFRAME	BITS/FRAME
LPC coefficients	37	
energy - $R_q(0)$	5	
excitation code - I	9	36
lag - L	7	28
GS-P0 code	7	28
TOTAL	23	134

Table 1 – Bit Allocations for 6.7 kbps coder



Class 1 bits are convolutionally coded and interleaved with the class 2 bits and transmitted over two time slots.

Figure 2 - Channel Coding

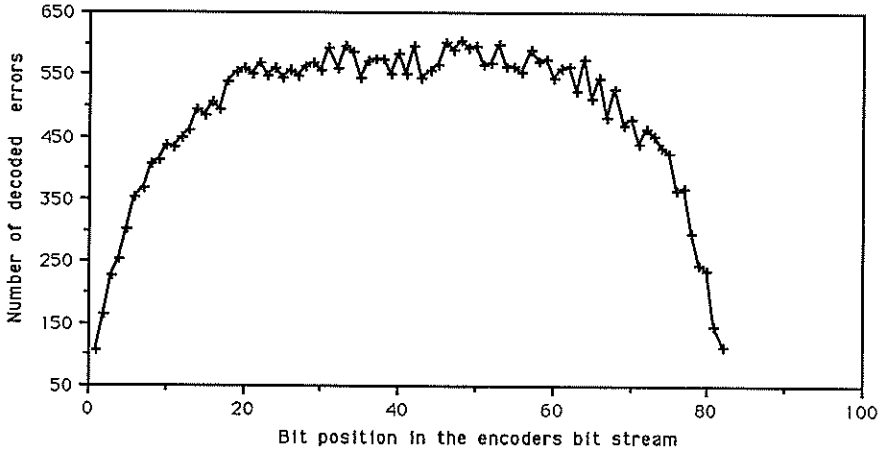


Figure 3 - Error Protection versus Bit Position

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

FILTER-BANK APPROACH TO TIME SCALING OF SPEECH

M.K. Asi and B.E.A. Saleh

Department of Electrical and Computer Engineering
University of Wisconsin-Madison
Madison, WI 53706

A method is introduced for the time scaling of speech, based on scaling the short-time Fourier transform in time, but not in frequency. Since this method is suitable for periodic signals, a bank of filters is used to decompose speech into quasi-periodic components that are scaled separately. The overall process is a periodically time-varying linear filter.

1. INTRODUCTION

The time scaling of speech with minimum loss of intelligibility is a difficult but important problem. Recently, we have developed a method for time scaling of speech based on scaling the short-time Fourier transform (STFT) of the original signal along the time axis leaving it unchanged along the frequency axis [1]. An optimal signal $\hat{f}(t)$ was then found with an STFT that matches the time scaled function with the least mean-square error. In this paper we consider extensions of this technique based on the use of a bank of filters.

2. THE OPTIMAL TIME-SCALING FILTER

If a signal f(t) is to be time scaled by a factor β, the STFT $C(t,\omega)$ must be modified to $C(\beta t, w)$. A new signal $\hat{f}(t)$ is found whose STFT $\hat{C}(t,\omega)$ matches $C(\beta t,\omega)$ in the least mean-square error sense. This optimization problem led to the relation [1]

$$\hat{f}(n) = \sum_{r=-\infty}^{\infty} f(n-rN)g(r,n), \qquad (1)$$

where

$$g(r,n) = \sum_{m=0}^{N-1} h(m)h\left(\lfloor \beta m+(\beta-1)n \rfloor + rN\right), \qquad (2)$$

h(n) is the STFT window function of length N, $\beta>0$ is the scaling factor, and $\lfloor x \rfloor$ denotes the largest integer less than or equal to x.

The filter defined by (1) and (2), which relates the time-scaled signal $\hat{f}(t)$ to the original signal f(t) is a linear periodically time varying filter with system function g(r,n). It has the interesting property that if the signal f(n) is periodic with period P, then perfect time scaling is accomplished if the window length is equal to an integer multiple of the period.

3. THE FILTER BANK APPROACH

Speech signals are clearly not periodic. Nevertheless, they contain vowels that are approximately periodic. We have applied this method to speech by setting the window length equal to an average value of the pitch period of the vowels. The resulting speech was of high quality and intelligibility, but the scaling of the consonants was not of the same quality. This problem can be alleviated by using a filter bank.

The original signal f(n) is filtered by a set of M narrow-band contiguous filters with center frequencies $\omega_k=2\pi k/M$, with outputs

$$f_k(n) = \sum_m h_k(n-m)f(m)\cos(2\pi km/N),\ k=1,2,\ldots,M. \qquad (3)$$

The function $h_k(n)$ is the impulse response function of filter k. The components $f_k(n)$ are then time scaled individually using our technique; the corresponding components $\hat{f}_k(n)$ are then used to construct the scaled signal $\hat{f}(t)$. If the filters are sufficiently narrow, then each of the components $f_k(n)$ are approximately periodic with period N/k, so that they can be scaled using a window of size N=M/k. The overall relation between the original and the scaled signals is then

$$\hat{f}(n) = \sum_u G(n,u)f(u), \qquad (4)$$

where

$$G(n,u) = \sum_{k=1}^{M} \sum_{r=-\infty}^{\infty} \sum_{m=0}^{\frac{M}{k}-1} h(n-u-rM/k)$$
$$h_k(m)h_k\left(\lfloor \beta m+(\beta-1)n \rfloor + r\frac{M}{k}\right)\cos(2\pi k(n-m)/k). \qquad (5)$$

Again, this is a linear periodically time-varying filter. The number of filters in the bank, M, and the shapes of the window functions, h(n) and $h_k(n)$, may then be varied to optimize the performance.

4. DELETION/REPETITION METHOD

An alternative and simpler method is to scale each of the components by use of the conventional periodic deletion/repetition method [2]. The introduction of the filter bank solves the problem of discontinuities that usually arises in that method, since each of the components is approximately periodic. The transformation

$$f_t(n) = f\left(n + (s-q)\left\lfloor \frac{n}{qP} \right\rfloor P\right) \qquad (6)$$

scales any periodic signal of period P by a factor $\beta = s/q$, where s and q are integers. If $s > q$ this transformation achieves time compression by deleting s periods and leaving q periods. When $s>1$, this transformation pastes qP periods and advances sP periods. The overall relation in this case is described by the linear time-varying filter

$$\hat{f}(n) = \sum_m H(n,m)f(m),$$

where

$$H(n,m) = \sum_{k=1}^{M} h_k\left(n+(s-q)\left\lfloor \frac{n}{qM/k} \right\rfloor \frac{M}{k} - m\right)\cos(2\pi k(n-m)/M).$$

5. RESULTS

We have applied these two methods to speech sentences spoken by male and female speakers using time scaling factors β=1.5, 2, 2.5, and 3. The scaled speech was judged to be of high quality. Examples of the STFT's obtained is shown in Figure 1. The STFT of f(t) is shown in Figure 1a with the abscissa representing time and the ordinate representing frequency. The STFT's of $\hat{f}(t)$ are shown in Figures 1b, 1c, and 1d, for β = 1.5, 2, and 2.5, respectively. A bank of six filters was used in all cases.

As a quantitative measure of performance, we have scaled the scaled signal by an inverse factor $1/\beta$ and determined the fidelity between the resultant, twice-scaled, signal and the original signal using the cross-correlation coefficient as a measure. The fidelity was found to be high. This method can be easily adapted to performing frequency scaling, instead of time scaling.

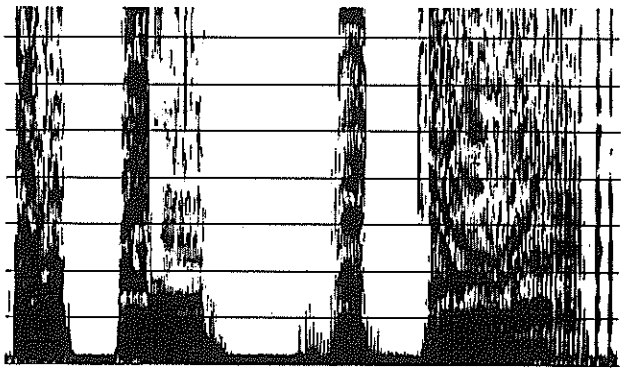
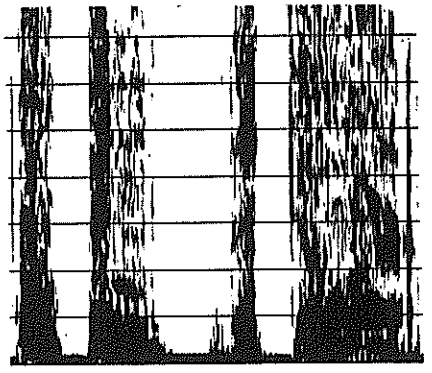
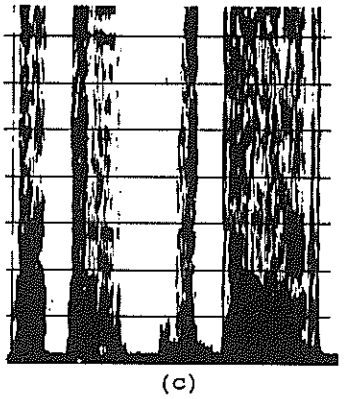


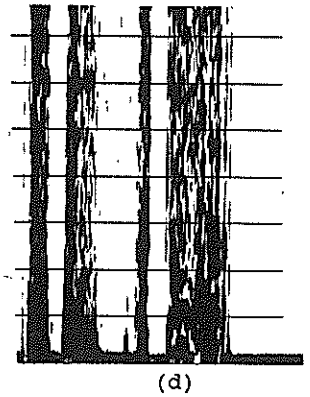
Fig. 1a Original STFT



(b)



(c)



(d)

Fig. 1b,c,d STFTs of scaled signals with scaling factors $\beta = 1.5$, 2, and 2.5.

REFERENCES

[1] M.K. Asi and B.E.A. Saleh, "A Linear Time-varying Filter for Time Scaling of Speech," Proc. IEEE Int. Conf. ASSP, New York, NY., Apr., 1988.

[2] F.F. Lee, "Time Compression and Expansion of Speech by the Sampling Method," Jour. of Audio Eng. Soc., vol. 20, pp. 738-742, Nov. 1972.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition

Xavier Serra and Julius O. Smith
Center for Computer Research in Music and Acoustics
Department of Music, Stanford University
Stanford, CA 94305, USA

A technique is described for modeling time-varying spectra as (1) a collection of sinusoids controlled through time by piecewise linear amplitude and frequency envelopes (the "deterministic" part), and (2) a time-varying filtered noise component (the "stochastic" part). The analysis procedure first extracts the sinusoidal trajectories by tracking peaks in a sequence of short-time Fourier transforms. These peaks are then removed by spectral subtraction. The remaining "noise floor" is then modeled as white noise through a time-varying filter: a piecewise linear approximation to the upper spectral envelope of the noise is computed for each successive spectrum, and the stochastic part is synthesized by means of the overlap-add technique. The technique has proved to give general, high quality transformations for a wide variety of musical signals.

1. Introduction

When generating musical sound on a digital computer, it is important to have a good model whose parameters provide a rich source of meaningful sound transformations. Three basic model types are in prevalent use today for musical sound generation: instrument models, spectrum models, and abstract models. Instrument models attempt to parametrize a sound at its source, such as a violin, clarinet, or vocal tract. Spectrum models attempt to parametrize a sound at the basilar membrane of the ear, discarding whatever information the ear seems to discard in the spectrum. Abstract models, such as FM, attempt to provide musically useful parameters in an abstract formula. This paper addresses the second category of synthesis technique: spectrum modeling.

2. The Deterministic plus Stochastic Model

A sound model assumes certain characteristics of the sound waveform or the sound generation mechanism. In general, every analysis/synthesis technique has an underlying model. The system presented in this article assumes the input sound to be composed of a deterministic plus a stochastic component.

A deterministic signal is traditionally defined as anything that is not noise (i.e., a perfectly predictable part, predictable from measurements over any continuous interval). However in the present discussion the class of deterministic signals considered is restricted to sums of quasi-sinusoidal components (sines with piecewise linear amplitude and frequency variation). Each sinusoid models an actual sinusoidal component of the original sound and is described by an amplitude and a frequency function.

A stochastic, or noise, signal is fully described by its amplitude and its general frequency characteristics. When a signal is assumed stochastic it is not necessary to preserve either the instantaneous phase or the exact frequency information.

Therefore, the input sound $s(t)$ is the sum of a series of sinusoids plus a noise signal,

$$s(t) = \sum_{r=1}^{R} A_r(t)\cos[\theta_r(t)] + e(t) \qquad (1)$$

where $A_r(t)$ and $\theta_r(t)$ are the instantaneous amplitude and phase of each sinusoid and $e(t)$ is the noise component.

The model assumes that the sinusoids are stable partials of the sound and that each one can be characterized by its amplitude and frequency. The instantaneous phase is then taken to be the integral of the instantaneous frequency $\omega_r(t)$, and therefore satisfies

$$\theta_r(t) = \int_0^t \omega_r(\tau)d\tau \qquad (2)$$

where $\omega(t)$ is the frequency in radians, and r is the sinusoid number.

By assuming that $e(t)$ is a stochastic signal, it can be described as filtered white noise,

$$e(t) = \int_0^t h(t,\tau)u(\tau)d\tau \qquad (3)$$

where $u(t)$ is white noise and $h(t,\tau)$ is the impulse response of a slowly time varying filter.

The filtering of a noise signal can be implemented by taking the inverse Fourier transform of the filter frequency response times a random phase term. This last approach is the one taken to synthesize the stochastic signal.

3. Description of the System

Fig. 1 and 2 show the block diagrams for the analysis and synthesis parts of the system. The first step is the derivation of a series of magnitude spectra of the waveform by computing the FFT of every windowed portion of the input signal, i.e., computation of the Sort-Time Fourier Transform (STFT). From the series of magnitude spectra the priminent peaks are detected in each spectrum. These peaks are then organized into frequency trajectories by means of a peak continuation algorithm.

The stochastic part of the waveform is calculated by first computing the STFT of the deterministic component, in the same way that the STFT of the original waveform was obtained, and then subtracting each magnitude spectrum from the corresponding spectrum of the original waveform. The envelope of each "residual" spectrum is then derived by performing a line-segment approximation. These envelopes represent the stochastic signal.

The deterministic signal, i.e., the sinusoidal component, results from the magnitude and frequency trajectories, or their transformation, by generating a sine wave for each trajectory (i.e., additive synthesis).

The stochastic signal is the result of creating a complex spectrum (i.e., magnitude and phase spectra) for every spectral envelope of the residual, or its modification, and performing an inverse-STFT (using the overlap-add method to form the final output).

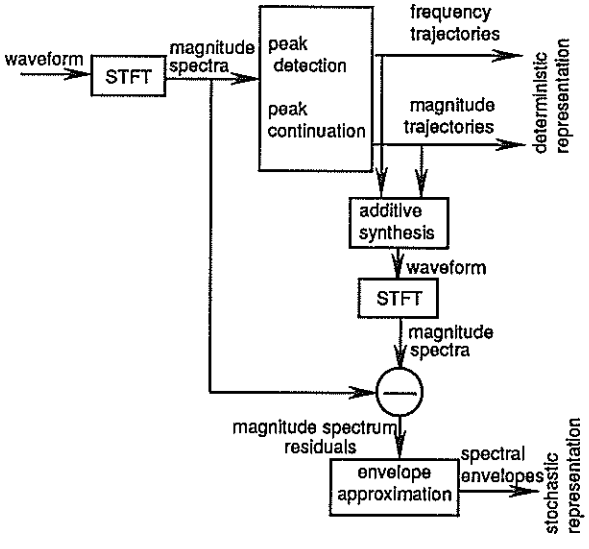


Figure 1.

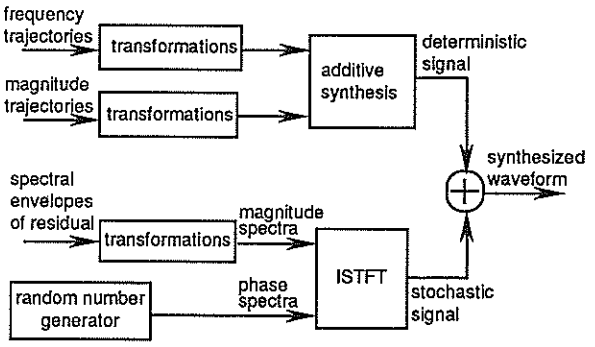


Figure 2.

4. Computation of the Magnitude Spectra

The analysis part of the system starts by computing a set of magnitude spectra using the STFT. This transform is defined as

$$X_l(k) \triangleq \sum_{n=0}^{N-1} w(n)x(n+lH)e^{-j\omega_k n}, \qquad l = 0,1,\ldots \quad (4)$$

where $w(n)$ is a real "window" that determines the portion of the input signal $x(n)$ that receives emphasis at a particular frame l. H is the hop-size, or time advance, of the window.

The choice of the analysis window is important. It determines the trade-off of time versus frequency resolution which affects the smoothness of the spectrum and the detectability of different sinusoidal components.

All the standard windows are real and symmetric and have a frequency spectrum with a sinc-like shape. For the purpose of our system, and in general for any sound analysis/synthesis application, the choice of window is mainly determined by two of its spectral characteristics: (1) the width of the main lobe, and (2) the highest side-lobe level. Ideally, we want a narrow main lobe. The choice of window determines this trade-off. For example, the rectangular window has the narrowest main lobe, 2 bins, but the first side-lobe is very high, -13dB relative to the main-lobe peak. The Hamming window has a wider main lobe, 4 bins, and the highest side-lobe is 43dB down. A very different window, the Kaiser, allows control of the trade-off between the main-lobe width and the highest side-lobe level. Since control of this trade-off is valuable, the Kaiser window is a good general-purpose choice.

The size of the FFT, N, is normally chosen to be the first power of two that is at least twice the window length M, with the difference $N - M$ filled with zeros ("zero-padded"), since zero-padding in the time domain corresponds to interpolation in the frequency domain.

5. Peak Detection

Once the set of spectra of a sound is computed, the system extracts the prominent peaks of each spectrum.

A peak is defined as a local maximum in the magnitude spectrum. However not all the peaks are equally prominent in the spectrum, and it is important to have control over their selection. This is done by measuring the height of each peak in relation to the neighboring valleys. Where the neighboring valleys are the closest local minima on both sides of the peak. At the same time not all the peaks of the same height are equally relevant perceptually, their amplitude and frequency is very important. Thus, it is useful to specify frequency and magnitude ranges where the search for peaks takes place.

Due to the sampled nature of the spectra returned by the STFT, each peak–a spectral bin that is a local maximum–is accurate only to within half a sample. A bin represents a frequency interval of f_s/NHz, where N is the FFT-size, and f_s is the sampling rate. As we saw in the previous section, zero-padding in the time domain increases the number of DFT bins per Hz and thus increases the accuracy of the simple peak detection. However, to obtain good frequency accuracy the zero-padding factor required is very big. A more efficient spectral interpolation scheme is to zero-pad only enough so that parabolic spectral interpolation, using only bins immediately surrounding the maximum-magnitude bin, can be used to refine the estimate accuracy [1].

6. Peak Continuation

Once the spectral peaks have been detected, a subset of them is organized by the peak continuation algorithm into peak trajectories, where each trajectory represents a stable sinusoid.

The algorithm is intended for a variety of sounds. Whether it is speech, an instrumental sound with a harmonic spectrum, a sound of a gong, a sound of an animal, or any other, the time evolution of the component partials will vary. Thus, the algorithm, apart from being general, requires some knowledge about the characteristics of the sound that is being analyzed. In the current algorithm there is no attempt to make the process completely automatic. The user is expected to know some of the characteristics of the sound beforehand, specifying them through a set of parameters.

The basic idea of the algorithm is that a set of *frequency guides* advances in time through the spectral peaks, looking for the appropriate ones (according to the specified constraints) and forming trajectories out of them. The instantaneous state of the guides, their frequency, is continuously updated as the guides are turned on, advanced, and finally turned off.

The output of the peak continuation algorithm is a set of peak trajectories. These represent the deterministic component, i.e., the partials of the analyzed sound. Each peak is a pair of numbers of the form $(\hat{A}_r(l),\hat{\omega}_r(l))$ where $\hat{A}$ and $\hat{\omega}$ are the amplitude and frequency, respectively, for each frame l and each trajectory r. The pairs corresponding to a trajectory r are interpreted as breakpoints for amplitude and frequency functions, one breakpoint for each frame l. From these functions a series of sinusoids can be synthesized which reproduce the deterministic part of the sound.

7. Deterministic Synthesis

Given the representation of the deterministic part of the sound, the generation of the time domain waveform is done with an additive synthesis technique. From the amplitude and frequency functions, $\hat{A}_r(l)$ and $\hat{\omega}_r(l)$, a frame of the deterministic sound is obtained by

$$d^l(m) = \sum_{r=1}^{R^l} \hat{A}_r^l \cos[m\hat{\omega}_r^l], \qquad m = 0,1,2,\ldots,H-1 \quad (5)$$

where R^l is the number of trajectories present at frame l and H is the length of the synthesis frame (without any time scaling H is the analysis hop-size). The final sound $d(n)$ results from the juxtaposition of all the synthesis frames. To avoid "clicks" at the frame boundaries, the parameters $(\hat{A}_r^l,\hat{\omega}_r^l)$ are smoothly interpolated from frame to frame.

The instantaneous amplitude $\hat{A}(m)$ is obtained by linear interpolation,

$$\hat{A}(m) = \hat{A}^{l-1} + \frac{(\hat{A}^l - \hat{A}^{l-1})}{H} m \quad (6)$$

where $m = 0,1,\ldots,H-1$ is the time sample in the lth frame.

The instantaneous phase is taken to be the integral of the instantaneous frequency, where the instantaneous radian frequency $\hat{\omega}(m)$ is also obtained by linear interpolation,

$$\hat{\omega}(m) = \hat{\omega}^{l-1} + \frac{(\hat{\omega}^l - \hat{\omega}^{l-1})}{H} m \quad (7)$$

and the instantaneous phase for the rth trajectory is

$$\hat{\theta}_r(m) = \hat{\theta}_r(l-1) + \hat{\omega}_r(m)m \quad (8)$$

Finally, the synthesis equation becomes

$$d^l(m) = \sum_{r=1}^{R^l} \hat{A}_r^l(m) \cos[\hat{\theta}_r^l(m)] \quad (9)$$

where $\hat{A}(m)$ and $\hat{\theta}(m)$ are the calculated instantaneous amplitude and phase.

8. Computation of the Stochastic Part

Once the deterministic component of the sound has been detected, the next step is to obtain the residual, which in a simplified form becomes the stochastic component.

Since the deterministic component does not preserve the phases of the original sound, a time domain subtraction cannot be performed (for a method that allows a time domain subtration see [1]). However since the magnitude and frequency of each sinusoid are preserved, the magnitude spectrum of both signals is comparable. Accordingly it is possible to perform a frequency domain subtraction from the magnitude spectra of both signals. The result is a set of magnitude-spectrum residuals.

Assuming that the residual signal is quasi-stochastic, each magnitude-spectrum residual can be approximated by its envelope, since only its shape contributes to the sound characteristics. This type of problem is generally solved by performing some sort of curve fitting, i.e., finding a function which matches the general contour of a given curve, which in our case is a magnitude spectrum. Standard techniques are: spline interpolation, the method of least squares, or straight line approximations. For the purpose of our system a simple line-segment approximation is accurate enough and gives the desired flexibility.

9. Stochastic Synthesis

The synthesis of the stochastic component can be understood as the generation of a noise signal that has the frequency and amplitude characteristics described by the spectral envelopes of the stochastic representation. The intuitive operation is to filter white noise with these frequency envelopes, that is, performing a time-varying filtering of white noise. But in practice we generate the stochastic signal by an overlap-add synthesis technique from the spectral envelopes. The inverse Fourier transform of each envelope is computed and the resulting waveforms are overlaped and added.

Before the inverse-STFT is performed, a complex spectrum (i.e., magnitude and phase spectra), is obtained from each frequency envelope. The magnitude spectrum is the envelope itself and the phase spectrum is a random signal. To avoid a periodicity at the frame rate different values are generated at every frame.

The inverse Fourier transform of the complex spectrum gives one frame of the noise waveform,

$$\hat{e}'_l(m) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} \hat{E}_l(k) e^{j\omega_k m}, \quad m = 0, 1, \ldots, N-1 \quad (10)$$

where $\hat{E}_l(k)$ is the complex spectrum, $\hat{e}'_l(m)$ is a constant-amplitude waveform of size N, and N is the FFT-size.

Since the phase spectrum used is not the result of an analysis process (with windowing of a waveform, zero-padding, and FFT computation) the resulting signal does not taper to zero at the boundaries. This is because a phase spectrum with random values corresponds to a phase spectrum of a rectangular-windowed noise waveform of size N. But in order to succeed in the overlap-add, i.e., obtain smooth transitions between frames, we need a smoothly windowed waveform of size M, where M is the synthesis-window length. Therefore the resulting waveform $\hat{e}'_l(m)$ is multiplied by a length M window,

$$\hat{e}_l(m) = \hat{e}'_l(m) w(m), \qquad m = 0, 1, \ldots, M-1 \qquad (11)$$

Then the stochastic signal results from the overlap and add of these windowed waveforms,

$$\hat{e}(n) = \sum_{l=0}^{L-1} \hat{e}_l(n - lH) \qquad (12)$$

where H is the analysis hop-size and l is the frame number.

10. Representation Modifications

The deterministic analysis results in a set of amplitude and frequency functions, $\hat{A}_r(l)$ and $\hat{\omega}_r(l)$, where r is the function number, and l the breakpoint number in each function. The stochastic analysis results in a set of spectral envelopes, $\hat{E}_l(q)$, where q is the breakpoint number in the envelope. From these representations a great number of sound transformations are possible.

Time-scale modifications are accomplished in both representations by resampling the analysis points in time. Due to the stochastic and deterministic separation, this representation is more successful in time-scale modifications than traditional additive synthesis techniques. The noise part of the sound remains "noise" no matter how much the sound is stretched (which is not the case in sinusoids-only representations).

In the deterministic representation each function pair, amplitude and frequency, accounts for a partial of the original sound. The manipulation of these functions is easy and musically intuitive. All kinds of frequency and magnitude transformations are possible. For example, the partials can be transposed in frequency, with different values for every partial and varying along the sound. It is also possible to decouple the sinusoidal frequencies from their amplitude, obtaining effects such as changing pitch while maintaining formant structure.

The stochastic representation is modified by changing the shape of each of the envelopes. Changing the envelope shape corresponds to filtering the stochastic signal further. Their manipulation is much simpler and more intuitive than the manipulation of a set of all-pole filters, such as those resulting from an LPC-analysis.

11. Conclusion

An analysis-based synthesis technique has been presented capable of capturing the perceptual characteristics of a wide variety of sounds. The representation that results from the analysis is intuitive and it is easily mapped to useful musical parameters.

The analysis part is central to the system. It is a complex algorithm that requires the manual setting of a few control parameters. Further work may automate the analysis process, particularly if there is a specialization for a group of sounds. Also, some aspects of the analysis are open to further research, in particular the peak-continuation algorithm.

The synthesis from the deterministic plus stochastic representation is simple and can be performed in real-time with current technology. A real-time implementation of this system would allow the use of this technique in performance. The representation would be precomputed and stored, and the sound transformations would be done interactively.

12. References

[1] Serra, X. 1989. *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition.* Ph.D. Dissertation, Stanford University.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

AUTOMATIC SELECTION OF SUBLEXIC TEMPLATES BY USING DYNAMIC TIME WARPING TECHNIQUES*

M.J. Castro†, P. Aibar, F. Casacuberta, E. Vidal

Dpto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Spain

The aim of this work is to develop automatic learning procedures to obtain the model or models for a certain type of linguistic unit, under the framework of a distance-based approach. The chosen unit is the phoneme and the model is the template. The models are obtained in an iterative process, by using Dynamic Time Warping and Word Spotting based techniques. Some experimental results are also reported for multi-speaker and single-speaker tasks.

1. INTRODUCTION

The aim of Acoustic-Phonetic Decoding (APD) is to obtain an interpretation of the speech signal in the form of certain linguistic units. All techniques used for APD associate one or more models to each linguistic unit. These models try to represent the variability of the acoustic events that characterize the linguistic units.

Some difficulties arise when an Acoustic Phonetic Decoder is designed. Two of these problems are: 1) election of the concrete linguistic units, and 2) use of the technique and consequently, the type of model.

Phonemes or phoneme-like units seem to be more popular in recent years [1] [2] [3]. However, other units such as diphonemes [4], syllables [5] [6], or demisyllables [7], are used by speech researchers. Other chosen units are not strictly linguistic and they are defined from automatic procedures based on acoustic properties of the speech signal [8] [9].

The techniques for APD can be classified in four broad categories [2]:

1. Rule-based approaches
2. Distance-based approaches
3. Probabilistic approaches
4. Discriminant approaches

Rule-based approaches use Expert System technology, and their popularity seems to have decreased in recent years. The distance-based approach uses techniques for template matching and is one of the less explored techniques for APD. The probabilistic approaches are mainly based on Hidden Markov Modeling, which appears to be one of the most promising techniques. Finally, the discriminant approach is based on Artificial Neural Nets, and is an open field of research.

The most popular technique used with the distance-based approach is Dynamic Time Warping (DTW). DTW was introduced originally for Isolated Word Recognition [10]. It was based on a Dynamic Programming scheme and allows comparison of two parameter vector strings, obtaining a time warping path and a dissimilarity or distance between the two strings. This technique was extended to Connected Word Recognition, with "Level building" and "One-stage" algorithms being the most widely used extensions [11] [12] [13]. These techniques attempt to compare a set of templates with an unknown sequence of words. Other extensions of DTW were Word Spotting techniques, that allow us to find key words in a sentence [14] [15] [16].

The aim of this work is to develop automatic learning procedures to obtain the model or models for each linguistic unit, under the framework of the distance-based approach. The chosen unit is the phoneme, the model is the template, and the techniques are two: one is based on simple DTW, and the other is based on Word Spotting techniques, within an iterative process. The basic

*Supported in part by the Spanish CICYT, under grant TIC 89/0448.
†Supported by a postgraduate grant from the "Conselleria de Cultura, Educació i Ciència de la Generalitat Valenciana".

idea of obtaining templates from a training set by using an iterative process was proposed in [17] to train the models for Connected Word Recognition and in [18] to train diphonemes. On the other hand, a similar idea has been developed for variable-length segment quantization in [19].

2.TWO APPROACHES FOR BUILDING A PROTOTYPE PER PHONEME

In this section, we will propose two ways for the utilization of DTW techniques to search templates as phoneme-representations from a set of training utterances. These approaches need a set of training samples and their corresponding phonetic transcriptions.

Both techniques consist in an iterative process, in which the prototypes of the phonemes are improved from one iteration to another. The procedure is iterated until a stable set of prototypes is obtained. The initialisation, in both cases, is based on a bootstrap phase, in which initial prototypes are built from a set of training samples that are manually segmented and labelled.

The difference between both proposed approaches is the distinct use of DTW techniques in the iteration phase.

The first approach builds a set of templates for each phoneme (M_i) in each iteration. These templates are obtained by using a conventional DTW algorithm. Each training sample is DTW compared with a reference sample that is formed by the concatenation of the prototypes of each phoneme given by the phonetic transcription (DTW procedure). The result of the comparison is a time warping path which relates each frame of the training sample with the reference sample obtained by concatenation. From this path and the known limits of the concatenated prototypes, a segmentation and labelling of the training sample can be performed (segmentation procedure) (Figure 1). These new segments are incorporated to the set of templates of each phoneme. Once all the training samples are segmented and labelled, the centroid of each set of templates, with respect to the DTW distance between them, is obtained and it is considered as the new prototype of the phoneme. This process is iterated until the prototypes do not change from an iteration to another.

The second approach is similar to the first one, except in the application of DTW and the segmentation procedures. In this case the training sample and each phoneme prototype corresponding to the phonetic transcription are individually compared, as for "word spotting" (WSDTW). The aim is to output all possible optimum

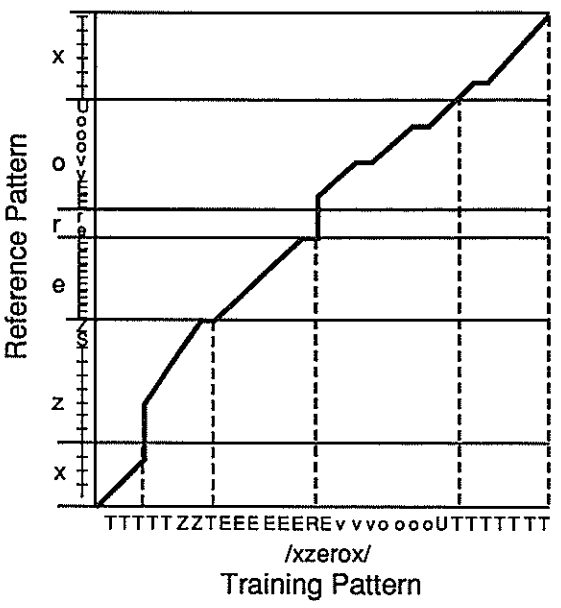


Figure 1: An example of segmentation of the training sample /xzerox/ by using the CDTW procedure (approach 1).

paths between the phoneme prototype and substrings of the training sample. In order to carry out this process, the contour conditions of the conventional DTW procedure are relaxed, in the sense that a warping path can start at any frame of the training sample and can reach any successive frame of this sample. On the other hand, the segmentation procedure is carried out through the search of the best sequence of segments, according to the phonetic tanscription and the output of the WSDTW procedure for each phoneme (Figure 2). The output scores of each path are normalized by their path lengths which are obtained by the WSDTW procedure, along with the scores.

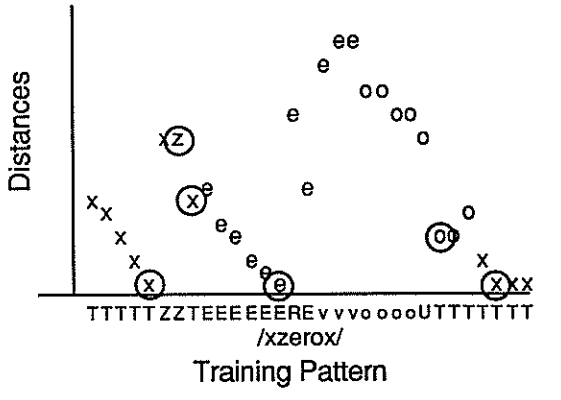


Figure 2: An example of segmentation of the training sample /xzerox/ by using the WSDTW procedure (approach 2).

The computational cost of the first algorithm is mainly due to the number of times that the DTW procedure is performed and the computation of the centroid. Let us assume N phonemes and M training samples with an average of n frames and m phonemes by sample. The average computational complexity is $O(I \cdot M \cdot n^2)$ for the computation of DTW, where I is the number of iterations until convergence, and $O(I \cdot M^2 \cdot n^2/N)$ for the centroid computation. Given that M>>N, the overall complexity is $O(I \cdot M^2 \cdot n^2/N)$. It is easily seen that the computational cost of the second algorithm is the same as that of the first one.

The actual Acoustic-Phonetic Decoder used in the recognition phase is based on the One-Stage algorithm [12], with some penalty in the inter-word productions [1].

3.EXPERIMENTAL RESULTS AND DISCUSSION

Some preliminary experiments were made to test the feasibility of the two approaches described in section 2. Two speech corpora have been used. The first one is the multi-speaker Spanish Digits. Ten repetitions of the ten digits from ten speakers were available. The second corpus consisted of a set of 209 aromatic and medicinal herbs. This corpus comprised 4 repetitions from a single speaker. Now, a new corpus of phonetically balanced Spanish sentences is being acquired to perform more thorough experiments.

The acquisition and parametrization of these utterances of the two corpora were rather standard, resulting in strings of 10-dimensional parameter vectors of Cepstrum coefficients, which were obtained at the rate of 66.66 vectors per second. Each coefficient vector (frame) was then given a label from a set of 15 broad microphonetic categories for the digits corpus and 32 for the herbs corpus. The labelling procedure consisted simply of a Nearest-Neighbour Classification of each frame with respect to 15 (32) sets of prototypes. These sets amounted to 255 overall prototypes and were derived by a two-level clustering procedure which was applied to a training set of frames.

The results of each experiment are reported through some rather standard figures that have been obtained through a Dynamic Programming algorithm which, for each test sentence, compares the correct phoneme sequence with that provided by the APD algorithms. These figures are "percent total" (pt), "insertion rate" (ir), "deletion rate" (dr) and "substitution rate" (sr):

$$pt = 100\,\frac{c}{c+s+i+d} \qquad ir = 100\,\frac{i}{c+s+i+d}$$

$$dr = 100\,\frac{d}{c+s+i+d} \qquad sr = 100\,\frac{s}{c+s+i+d}$$

where c is the number of correctly recognized phonemes, i is the number of inserted phonemes, s is the number of confused phonemes and d is the number of deleted phonemes.

The experiments with the Spanish digits were carried out by using all the utterances of 6 speakers for training and the rest, 4 speakers, for testing. The results are presented in table 1. For comparison purposes, we include similar experiments with Hidden Markov Modeling.

Table 1: Experimental results for the Spanish digits. N is the number of different phonemes in the dictionary.

	N	pt	ir	dr	sr
CDTW	15	58	15	10	17
WSDTW	15	56	23	4	17
HMM	14	67	6	11	16

It should be taken into account that this is a very easy task, since the number of phonemes and possible contexts are very small.

With the corpus of the herbs, we have made two series of different experiments. In the first one, one repetition of the 209 words was used for training, and the rest for testing (E1). In the second one, one repetition of only 100 words was used for training, and the rest of the words for testing (E2). This last, rather challenging experiment is aimed to approach the real situation of an APD which, in general, will have to do its job for speech data with different linguistic contents than that of the training data.

The results of these experiments are presented in table 2. In the recognition phase, segments of less than four frame lengths, are not permitted.

Table 2: Experimental results for the aromatic herbs. N is the number of different phonemes in the dictionary. E1 and E2 are two series of experiments described in the text.

	N	pt	ir	dr	sr
E1 CDTW	25	60	4	14	20
E2 CDTW	25	58	5	12	23
E1 WSDTW	25	58	8	8	24
E2 WSDTW	25	58	8	8	24

4.CONCLUSIONS AND FUTURE WORK

In this work we have presented two algorithms for automatically learning template-based models of phonemes from a training set of utterances.

The obtained results are encouraging, if we take into account the very coarse representation of each phoneme. In fact, in the actual implementation all the acoustic variability of each phoneme is represented by one template that is the centroid of a set of possible manifestations (contexts) of the phoneme in the training set.

Now, we are developing new versions of the presented algorithms in which the representation of each phoneme will consist of a set of prototypes. These prototypes are obtained by application of a clustering algorithm to each set of samples of each phoneme. The number of prototypes per phoneme depends on the corresponding output distortion of the clustering procedure.

It is expected that, with this improvement, the algorithms will be able to capture the most representative contexts of each phoneme.

Finally, new continuous speech corpora are now available to test the methods proposed in this work in a more appropiate way.

REFERENCES

[1] K.F. Lee: "Large-Vocabulary Speaker Independent Continuos Speech Recognition: the SPHINX System", PH. Thesis, Tec. Rep. CMU-CS 88-148. Carnegie Mellon Univ.

[2] R.M. Schwartz et. al.: "Acoustic-Phonetic Decoding of Speech" in "Recent Advances in Speech Understanding and Dialog Systems", H. Nieman, M. Lang, G. Saguer (eds.), Springer-Verlag, pp. 25-50, 1988.

[3] S. Nakagawa: "Speaker-independent continuos-speech recognition by phoneme-based word spotting and time-syncronuous context-free parsing", Computer Speech and Language, vol.3, pp 277-299, 1989.

[4] A.M. Colla, A.E. Rosenberg: "Unsupervised bootstrapping of diphone-like templates for Connected Speech Recognition", Proc. ICASSP 87, pp 1281-1284, 1987.

[5] T. Watarabe: "Syllable recognition for continuous Japanese speech recognition", Proc. ICASSP 86, pp 2295-2298, 1986.

[6] M. Wagner: "A speech recognition experiment with the entire syllable inventory of standar Chinese", Speech Communication, no. 6, pp 363-369, 1987.

[7] W. Weigel: "Recognition of demisyllables based on Dynamic Programming methods", Speech Communication, vol. 7, pp 297-304, 1988.

[8] J.G. Wilpon, B.H. Juang, L.R. Rabiner: "An Investigation on the use of acoustic sub-word units for automatic speech recognition". ICASSP 87, pp 821-824, 1987.

[9] C.H. Lee, B.H. Juang, F.K. Soong, L.R. Rabiner: "Word recognition using whole word and subword models". Proc. ICASSP 89, pp 683-686, 1989.

[10] H. Sakoe, S.Chiba: "Dynamic Programming Algorithm Optimization for Spoken Words Recognition". IEEE Trans. ASSP, vol. 26, pp. 43-49, Feb. 1978.

[11] J.B. Kruskal, D. Sankoff: "An Anthology of algorithms and concepts for sequence comparison" in "Time warper, string edits and macromolecules: The theory and practice of sequence comparison", D. Sankoff and J.B. Kruskal (eds.), Addison Wesley, chp. 10, 1983.

[12] H. Ney: "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition", IEEE Trans. ASSP, vol. 32, no. 2, pp 263-271, Apr. 1984.

[13] C. Godin, P. Lockwood: "DTW schemes for continuous speech recognition: a unified view", Computer Speech and Language, vol. 3, pp. 169-198, 1989.

[14] A. Komatsu, A. Ichikawa, K. Nakata, Y. Asakawa, H. Matsuzaka: "Phoneme recognition in continuous speech", Proc. ICASSP 82, pp. 883-886, 1982.

[15] S. Nakagawa, A. Hauptmann, M. Tomita: "On Quick Word Spotting Techniques", Proc. ICASSP 86, pp 2311-2314, 1986.

[16] F. Casacuberta, E. Vidal: "Reconocimiento Automático del Habla", Marcombo, 1987.

[17] L.R. Rabiner, J.G. Wilpon, B.H. Juang: "A Continuous Training Procedure for Connected Digit Recognition", Proc. ICASSP 86, pp 1065-1068, 1986.

[18] A.M. Colla: "On training a large vocabulary speech recognition system", Proc. International Conference on Speech Technologies, pp 107-116, Verba 90.

[19] Y. Shiraki, M. Honda: "LPC Speech Coding Based on Variable-length Segment Quantization", IEEE Trans. ASSP, vol. 36, no. 9, pp 1437-1444, Sep. 1988.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

AUTOMATIC SEGMENTATION OF CONTINUOUS JAPANESE SPEECH INTO PHONEMIC UNITS

Satoshi IMAI and Chieko FURUICHI

Research Laboratory of Precision Machinery and Electronics
Tokyo Institute of Technology
Nagatsuta, Midori-ku, Yokohama 227 Japan

Abstract This paper presents a high performance automatic phonemic segmentation system for speaker and context independent continuous Japanese speech recognition. The algorithm is implemented as the hierachical segmentation and broad category classification, using selected segmentation parameters and acoustic phonetic knowledge concerning continuous Japanese speech. The segmentation of continuous, reading-rate speech utterances and phonetically balanced word utterances with various phonetic environments into phonemic units is successfully performed. In the evaluation of the segmentation for 6 continuous, reading-rate Japanese speech utterances produced by 3 female and 3 male speakers, the segmentation error was 3.8 %, consisting of 2.2 % missed and 1.6 % extra. For the 492 polysyllabic word utterances in the phonetically ballanced word set, the score of the phonemic segmentation was found to be 94.7 %.

1. INTRODUCTION

The automatic phonemic segmentation and labeling is a key technique for continuous speech recognition. This paper presents an effective method for the automatic phonemic segmentation and broad category labeling of continuous Japanese speech.

Our ultimate objective is the realization of a speaker and context independent continuous speech recognition system. We are attempting to reach this goal by representing utterances by sequences of phonemic units. Every word in the arbitrary lexicon can be easily coded into a phoneme string. Besides, the total number of phonemes is very small. Gived these considerations, a natural choice for a recognition unit to be used in an automatic continuous speech recognition with unlimited vocabulary is the phoneme.

However, it is not easy to achieve effective and reliable automatic segmentation into phonemic units. It is considered to be largely due to the diversity in the acoustic properties of speech sounds arising from different interphonemic contexts, speaking rate, and variety of speakers.

To solve the difficult problem, we used six selected segmentation parameters, and a hierachical segmentation and broad category labeling algorithm based on acoustic phonetic knowledge concerning continuous Japanese speech. Four parameters among the six ones are computed from the spectral envelopes of speech obtained by the unbiased logarithmic spectral estimation technique [1]. The spectral estimation technique yields sufficient quality spectral envelopes for obtaining high performance segmentation parameters.

The effectiveness of the segmentation system has been substantiated for continuous, reading-rate speech utterances and phonetically balanced word utterances.

2. SYSTEM DESCRIPTION

The proposed segmentation and labeling system is sketched in Figure 1. The algorithm is implemented as the hierachical segmentation and broad category classification.

2.1. Speech signal analysis

The speech wave is sampled at 10 kHz. A 25.6 ms Blackman window is applied to the speech signal every 10 ms. Four segmentation parameters are computed from spectral envelopes of speech. The spectral envelope is represented in a form of the mel log spectrum.

The mel-log spectral envelope $G_i(\tilde{\Omega})$ is defined by

$$G_i(\tilde{\Omega}) = \sum_{m=0}^{M} g_i[m] \cos(m\tilde{\Omega}) \qquad (1)$$

where $g_i[m]$ $(0 \leq m \leq 12)$ is the m-th mel cepstral coefficient at i-th analysis frame for the minimum phase system representing the mel-log spectral envelope obtained by the

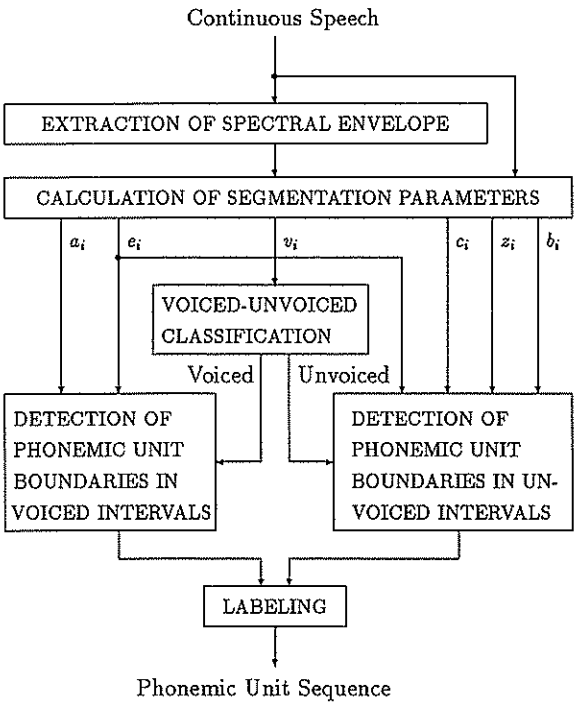


Figure 1: Block diagram of the automatic segmentation and broad category labeling system using three static and three dynamic segmetation parameters.

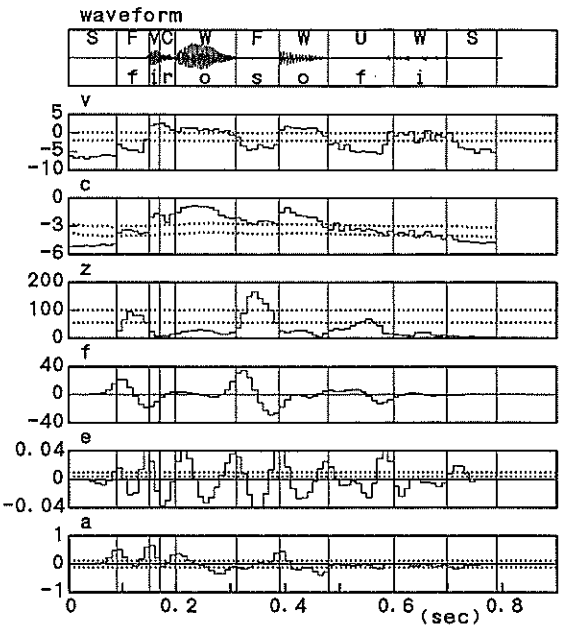


Figure 2: Waveform of Japanese utterance /firosofi/, which means philosophy, of a female speaker and the segmentation parameters $v_i, c_i, z_i, f_i, e_i, a_i$ and example of automatic segmentation into phonemic units and broad category labeling.

unbiased logarithmic spectral estimation technique [1].

The relation of the linear frequency scale Ω and the mel frequency scale $\tilde{\Omega}$ is represented as

$$\tilde{\Omega} = \Omega + 2\tan^{-1}\frac{\alpha\sin\Omega}{1-\alpha\cos\Omega} \qquad \left(\Omega = \frac{\omega}{f_s}\right) \tag{2}$$

where α is the frequency warping parameter set as 0.35, and ω and f_s are radian frequency and sampling frequency, respectively.

The mel-log spectral envelopes represented by the trigonometric polynomials are of pole-zero type, and they can exactly express the acoustic phonetic properties of speech. Furthermore, the spectral envelopes are sufficiently smooth to take time-derivative.

The remaining two segmentation parameters are the zero crossing number and its time-derivative, respectively. The zero crossing number is the number of zero crossing of the speech signal in a 25.6 ms interval.

2.2. Static segmentation parameters

The upper three parameters below the speech waveform in Figure 2 are:

- voiced sound detection parameter v_i defined by the averaged value of the log spectral envelope components between 78 and 312 Hz, of pitch frequency band,
- zero-th order mel-cepstral coefficient c_i (= $g_i[0]$) corresponding to the averaged value of the mel-log spectral envelope over whole frequency band,
- zero crossing number z_i defined as the number of zero crossing of the speech signal in a 25.6 ms interval.

2.3. Dynamic segmentation parameters

The lower three parameters below the speech waveform in Figure 2 are:

- time-derivative f_i of the zero crossing number, obtained by quasi-derivative filtering of z_i,
- time-derivative e_i of the mel log spectral envelope, obtined by quasi-derivative filtering of $G_i(\tilde{\Omega})$,
- time-derivative a_i of the zero-th order mel cepstral coefficient, obtained by quasi-derivative filtering of c_i.

The time-derivative of the zero-th order mel-cepstral coefficient a_i is represented by

$$a_i = K_M \sum_{n=-M}^{M} w_n n c_{i+n} \qquad \left(k_M = \Big(\sum_{n=-M}^{M} w_n n^2\Big)^{-1}\right) \tag{3}$$

where w_n is a Blackman window and the system of impulse response $w_n n$ operates as a quasi-derivative filter [2]. The length of the window 2M+1 was set to seven frames (70 ms), a priori, based on the effectiveness indicated by preliminary experiments. The pass band frequency of the time-derivative filter is approximately between 5 and 30 Hz.

3. SEGMENTATION AND LABELING

Vertical lines in Figure 2 denote phonemic boundaries automatically extracted by the algorithm applied to the six segmentation parameters. The doted lines denote the threshold. The segmentation procedure is hierachically performed evaluating the segmentation parameters.

3.1. Voiced-unvoiced classification

At first, the speech signal is classified into two classes: voiced intervals and unvoiced intervals including silence. The voiced-unvoiced boundaries are detected using three-valued logical parameter V_i^T represented as

$$\begin{aligned} V_i^T &= U(v_i - T_{VL}) + U(v_i - T_{VH}) \\ &= \begin{cases} 2 & (v_i > T_{VH} = 0.0) \\ 1 & (T_{VL} < v_i \leq T_{VH}) \\ 0 & (v_i \leq T_{VL} = -2.0) \end{cases} \end{aligned} \tag{4}$$

where the variable v_i is the voiced sound detection parameter and $U(\cdot)$ is a unit step function.

3.2. Segmentation in voiced intervals

Plosives (/b/, /d/, /g/) and fricatives (/z/, /h/, /f/) are phonemes that may carry either voiced or unvoiced intervals by a difference in intensity of voicing occurred in interphonemic context.

Voiced intervals include five Japanese vowels (/a/, /i/, /u/, /e/, /o/), semivowels (/y/, /w/), nasals (/m/, /n/, /N/), liquid (/r/), voiced plosives (/b/, /d/, /g/) and voiced fricatives (/z/, /h/, /f/). The candidates of phonemic boundaries within the voiced interval are detected by picking of the peaks exceeding a threshold, of two dynamic segmentation parameters: the time-derivative a_i of the zero-th order mel-cepstral coefficient and the time-derivative e_i of the mel-log spectral envelope.

3.3. Segmentation in unvoiced intervals

Unvoiced intervals include unvoiced plosives (/p/, /t/, /k/), unvoiced fricatives (/s/, /c/, /ts/), devocalized vowels (/i̥/, /u̥/) and silence. The silence and unvoiced sound segmentation within the unvoiced intervals is performed using a dynamic multi-level threshold logic of the zero-th order mel-cepstral coefficient c_i and the zero crossing number derivative of the zero-th order mel-cepstral coefficient a_i and the time- derivative of the mel-log spectral envelope e_i.

The phonemic boundaries within the unvoiced sound are detected using peak-picking rules applied to the two dynamic segmentation parameters: they are, the time-derivative of the zero crossing number of the waveform f_i and the time derivative of the mel-log spectral envelope e_i.

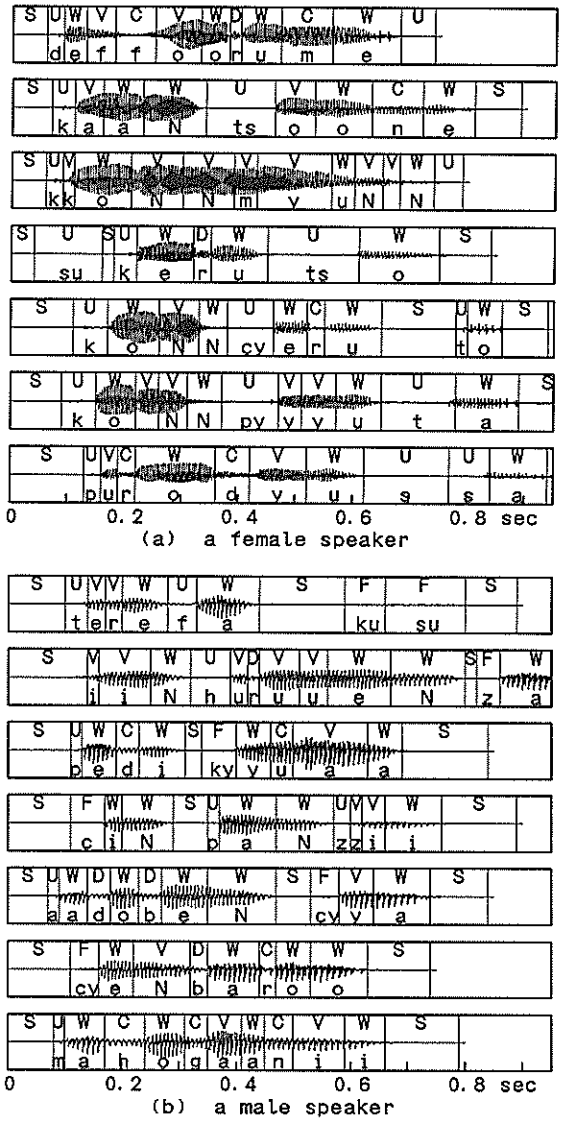


Figure 3: Examples of automatic segmentation into phonemic units and broad category labeling for Japanese loan word utterances selected from phoneticaly balanced word set produced by a female and a male speakers. Words of (a) mean "déforme", "canzone", "commune", "scherzo", "concerto", "computer" and "producer". Words of (b) mean "telefax", "influenza", "pedicure", "chimpanzee", "adventure", "cembalo" and "mahogany".

3.4. Labeling of the resulting segments

The resulting phonemic boundaries are selected from these candidates using a Japanese acoustic phonetic knowledge-based algorithm.

The labeling argorithm applied to the six segmentation patameters assigns broad category labels to the resulting segments. Each segments is assigned one of seven broad category labels: vowel(W), vowel-like(V), voiced consonant(C, D), unvoiced plosive(U), unvoiced fricative(F) and silence(S). In Figure 2, the labels above the waveform denote broad category automatically assigned by the argorithm and lower labels denote hand labeled Japanese phonemes.

4. EVALUATION

The evaluation of the segmentaion system was carried out using 6 continuous, reading-rate speech utterances and phonetically balanced word utterances.

4.1. Voiced-unvoiced decision

The error rate of voiced sound detection for 60,358 frames utterances with various phonetic environments produced by a female and a male speakers was 0.26 % [3].

4.2. Segmentation of continuous speech utterances

To evaluate the automatic segmentation system for continuous, reading-rate speech utterances, segmentation experiments were carried out using news and weather forecasts broadcasted over the FM radio. The utterances were produced by 3 female and 3 male announcers, and the sentences contained 1012 phonemic units. In this evaluation, the segmentation error rate was 3.8 %, consisting of 2.2 % missed and 1.6 % exta.

4.3. Segmentation of balanced word set utterances

Moreover, for various phonetic environments, the effectiveness of the system was substantiated. The experiments were carried out to a phonetically balanced word set in three phoneme sequences consisted of 492 selected polysyllabic words which had almost all vowel-consonant-vowel (VCV) environments in the Japanese dictionary. Among the 492 words, there were 1006 types of VCVs and 1905 types of VCVs and CVCs [4].

A phonetically balanced word set utterances were produced by two speakers: a female and a male. The 984 polysyllabic word utterances contained 9928 phonemic units. Examples of automatic segmentation into phonemic units and broad category labeling for loan words utterances in Japanese selected from phonetically balanced word set produced by a female and a male speakers are shown in Figure 3.

The total score of the segmentation for the 9928 phonemic units was found to be 94.7 %. The missing error rate for five Japanese vowels (/a/, /i/, /u/, /e/, /o/), consonants (/w/, /f/, /k/, /s/, /c/, /ts/) was less than 3 %. Approximately 50 % of all missed segnents involved nasals(/n/, /m/), voiceless plosive(/p/) and devocalized vowels (/hi/, /hu/). Closer examination of the errors reveals that the miss mostly involves acoustic transitions that are not always distinct, such as those between initial nasals of words and vowels, between plosives and vowels, and those loudness is below the noise threshold.

5. CONCLUSION

We presented a high performance automatic phonemic segmentation system for speaker independent and context independent continuous Japanese speech recognition. In this system, we used the high quality segmentation parameters derived from the spectral envelopes obtained by the unbiased logarithmic spectral estimation technique. The spectral envelopes are of pole-zero type based on the trigonometric polynomial log spectral representation, and they can exactly express the acoustic phonetic properties of speech. The hierachcal segmentation and broad category classification algorithm was implemented, based on acoustic phonetic knowledge concerning Japanese speech.

It has been substantiated that this automatic segmentation and broad category labeling algorithm is very useful for phonemic recognition of context independent continuous Japanese speech.

REFERENCES

[1] Imai, S. and Furuichi, C.,"Unbiased estimatior of log spectrum and its applications to speech signal processing," in Proc. EUSIPCO, pp.203-206, Sep. 1988.

[2] Imai, S. and Furuichi, C.,"Segmentation of continuous speech into phonemic units," (in Japanese), Trans. Inst. Electron. Inform. Commun. Eng. Japan, vol.J72-D-II, 1, pp.11-21, Jan. 1989.

[3] Furuichi, C. and Imai, S.,"Phonemic units segmentation in various phonetic environments," (in Japanese), vol.J72-D-II, 8, pp.1221-1227, Aug. 1989.

[4] Hayamizu, S., Tanaka, K., Yokoyama, S., and Ohta, K., 'MGeneration of VCV/CVC balanced word sets for speech data base," (in Japanese), Bul. Electrotechnical Lab., vol.49, 10, pp.803-834, Oct. 1985.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

A SPEAKER-ADAPTIVE SPEECH RECOGNITION SYSTEM FOR A LARGE, EXTENDABLE VOCABULARY

Heidi HACKBARTH, Peter FESSELER, Michael TROMPF, Manfred IMMENDÖRFER, Harald ECKHARDT

SEL Alcatel Research Center, Lorenzstr. 10, D-7000 Stuttgart 40, F.R. Germany

The speech recognition system presented here performs speaker-adaptive recognition of speech from large, extendable vocabulary. It is based on segmentation and classification of a particular type of subword units which are aligned to words. The reference vocabulary can be generated and extended automatically from written text input, thereby accounting for probable variations in pronunciation by so-called word models for each item in the lexicon. The hypothesized word sequences can be submitted to further linguistic analysis for phrase recognition. In the article, the modular system and the methodology are described and recognition results are given.

1. INTRODUCTION

The aim of the activity presented here was to develop a complete and comprehensive speaker-adaptive recognition system for speech from unrestricted vocabulary. A modular system structure, as sketched in figure 1, has been chosen to allow for individual optimization of each component, as performed here for the German language.

For signal pre-processing, loudness vectors are extracted as features and segmented into subword units. These are classified according to a previously trained inventory and aligned to sequences for word recognition.

A vocabulary of practically unlimited size can be generated and arbitrarily extended from written text input. The internal structure of the lexicon items also accounts for probable pronunciation variations. The search process in large vocabularies is accelerated by the implementation of a preselection method. To finally arrive at phrase or sentence recognition, a dedicated syntax is to be derived from the particular application area.

The system is speaker-adaptive, which means that it adapts to the particular pronunciation of a novel speaker within a brief enrollment phase without having the new user utter the whole vocabulary. Here, adaptation is following the approach to affect individual feature vectors or entire subword units.

In the course of the article, first the type of the subword units and their extraction from the speech signal are described. Next, focus is on the word recognition module, including the lexicon with its particular format and the vocabulary generation from text. Several approaches to speaker adaptation are illustrated and finally a summary is given.

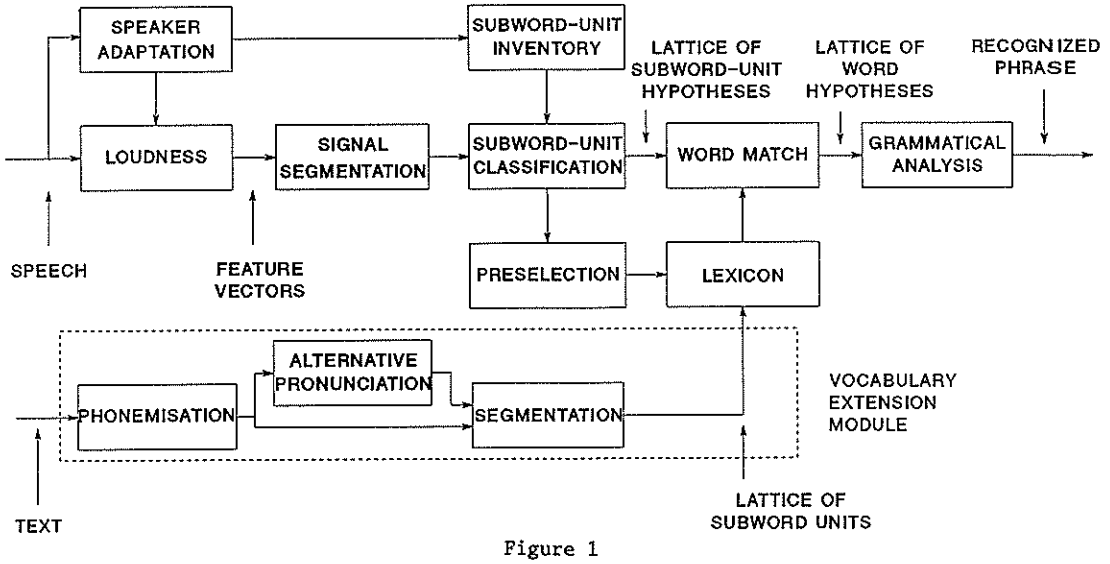


Figure 1

2. CVC SUBWORD UNITS

In this paragraph, the particular type of subword units is described, together with its segmentation in the speech signal and the classification procedure.

2.1. Definition

As subword entities, CVC units [1] were chosen, which divide each syllable into a vocalic kernel (V) with a syllable-initial and syllable-final consonant cluster (C) surrounding it. Consonant clusters may be empty.

This type of subword units has been selected, because they suggest themselves for automatic extraction from the speech signal in so far as no phonetic expert knowledge is requested. It further bears the advantage of a limited inventory size of a few hundred units, variations according to coarticulation effects already being considered, as compared to larger units, e.g. several thousand demisyllables. Third, the strict CVC pattern per syllable facilitates several other procedures within the system, for example word match and vocabulary extension.

2.2. Pre-Processing

Feature extraction from speech signals is based on models of the human auditory system as nonlinear filter bank [2]. To gain feature vectors, 8 ms intervals of digitized speech (sampled at 16 kHz) are passed through a 64 channel filter bank and transformed into 20 loudness channels, covering the frequency range from 50 Hz to 6.4 kHz with increasing bandwidth per channel. Apart from the 20 vector coefficients, further loudness functions [3] are derived and stored for each frame for use with subsequent modules.

2.3. Segmentation

Segmentation by the total loudness function is conducted in three steps; syllable detection, syllable and vowel boundary determination. The method is illustrated by figure 2 for the German word "Erd-bee-ren" (strawberries), here pronounced as "Erd-beer'n" with two syllables. The initial consonant cluster is lacking.

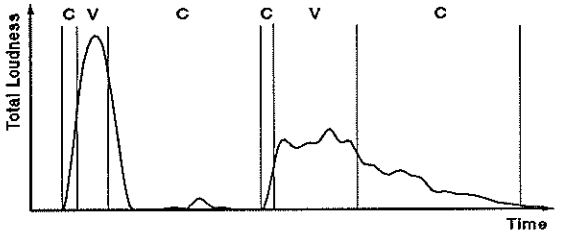


Figure 2

Syllables can be detected reliably from the prominent maxima of the smoothed summed loudness function. The error rates lie at about 3%. Misses occur mainly with faint German "i" or the schwa, insertions at long and stressed vowels or vowels being followed by nasals or liquids. Subsequently, the boundaries between successive syllables are determined from the minima of the summed loudness. Finally, the vocalic kernel duration is specified within a threshold-dependent range around the maximum of that function. The boundaries are indicated by vertical lines in the figure.

2.4. Classification

For identification of the segmented feature vector sequences, it is distinguished between vowels and initial and final consonant clusters, the locations of which are known from the strict CVC pattern. The prevailantly stationary vowel parts are compressed to four vectors, whereas the consonant clusters are normalized by dynamic interpolation [1] to 18 vectors minus four towards the boundary next to the vowel.

During training, the CVC units are identified, acoustically checked according to the pre-scribed text and stored in the specific sub-inventories according to the different CVC types, usually with several representations per unit. About 450 different CVC units appear to sufficiently cover the German language. They are extracted from 100 German sentences plus 1000 words in two or three repetitions each. Inventory generation has to be performed only once for each speaker, independent of the size of the lexicon, and might even be derived for a new speaker from an existing inventory by adaptation.

In the recognition mode, the segmented and normalized vector sequences are compared with the items of the corresponding CVC sub-inventory by means of the city-block distance. This results in multiple hypotheses per unit, the number of which can be threshold-dependent or fixed. The highest recognition accuracy is obtained for vowels. In the mean, however, the correct CVC unit is among the top two candidates to more than 70%.

3. WORD RECOGNITION

The module dedicated to the word match has a lattice of multiple subword units at its disposal, from which it generates overlapping word hypotheses, thereby referencing the lexicon with its particular internal structure. For use with larger vocabularies, a preselection method is introduced.

3.1. Lexicon Generation and Extension

The structure of the lexicon entries has to

meet several requirements in order to be compatible with other system modules. For example, a search procedure according to various keys such as for preselection must be possible. A lexicon item for a word thus consists of several entries. Apart from orthography, the CVC transcription of the standard as well as alternative pronunciations are noted in the so-called word model. These transcriptions as carried out by the vocabulary extension module in figure 1, are discussed in detail in a recent paper [4], the most important issue being the compatibility between the generated CVC transcription and the CVC hypotheses lattice of a test utterance. The number of syllables, the word class and stress marks are also included by interactive editing for use by successive modules.

3.2. Pronunciation Variations

The module for automatic vocabulary extension produces a subword unit string from phonemic transcription for the standard pronunciation of a word, as well as alternatives which cover the most common variations in pronunciation [5]. This leads to a lattice of subword units for each lexicon item, as shown in the left part of figure 3.

3.3. Preselection

For large vocabulary, it is suggested to insert a preselection stage before performing fine classification. The goal of preselection is to considerably reduce the amount of words to be verified during fine classification while realizing a high hit rate and short computing time. Several methods for fast lexicon search are under investigation, among them one is key-oriented and one based on sequences of "coarse" vowels, gained from clustering the 25 different vowel qualities into 3 to 8 coarse classes by a method called EDGE [7].

Two differented key-oriented strategies for lexical access have hitherto been tested, one referring to the number of syllables, the second to the first coarse vowel in a reference word. The latter method also applies to continuous speech recognition. A combined access by these two parameters obtained a reduction of a 1000-word lexicon by a factor of 7.4 for subsequent fine classification. The correct word was always contained in the preselected sub-vocabulary if the best two hypotheses out of 4 coarse vowel classes were considered.

Relying on a sequence of coarse vowels for preselection, satisfactory recognition rates were obtained with 4 coarse classes by 99% correctly classified vowels in the top 2 candidates. To gain 100% hits within the preselected word sub-vocabulary, coarse-vowel sequences with two hypotheses each proved to be sufficient. A reduction rate of at least 3.9 could be achieved.

A combination of the listed search methods resulted in a reduction by a factor of 9.3 compared to the original vocabulary size.

3.4. Recognition Algorithm

The algorithm dedicated to word recognition deals with the comparison of the network of the multiple CVC hypotheses with the word models of the (preselected) reference words in the lexicon. Two aspects have to be considered during this process, time-dynamic alignment and three-dimensional comparison.

To neutralize insertion or omission errors in syllable detection or elisions in pronunciation variations, a dynamic pattern alignment is required. Due to the rigid CVC structure per syllable, only syllable-sized "jumps" may occur. For example, a vowel can only be followed by the syllable-final consonant cluster of the next syllable to maintain the CVC pattern.

Due to the incorporation of pronunciation variations for each lexicon item and the multiple hypotheses of the actual CVC sequence, respectively, a third dimension is accessed in the comparison space, in addition to the time axes of the test and reference words. This is illustrated in figure 3, together with the temporal dynamics, for the German word "Erd-bee-ren". A "1" is indicating correct, "0" wrong CVC match, and "-" prohibited local warping path. Again, the rigid CVC structure can be exploited to economize the search in three dimensions, as compared to the phonemic approach by Kobayashi and Niimi [6], where overlap may frequently occur.

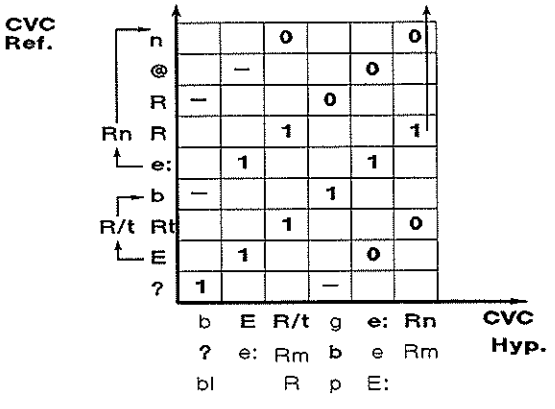


Figure 3

Recognition results for a 1000-word vocabulary and two male speakers amount to about 75%. This accuracy can be further increased by the incorporation of pronunciation variants and/or an optimized distance measure between CVC input lattice and reference model. In continuous speech, the error rate may also be reduced by an application-specific language model.

4. ADAPTATION

Speaker adaptation represents an essential feature of a high-performance speech recognition system. The goal here is to adapt the system to a novel speaker within a brief enrollment phase, thus avoiding the user to pronounce each word of a possibly large vocabulary.

The task of speaker adaptation has been tackled under three different aspects, frequency-dynamic, linear, and non-parametric transformation of feature vectors or their sequences. Considered the relevance of vowels for the recognition process, their adaptation was realized first. Training material for the first method were 20 to 25 s of speech, or was extracted from about 300 words. Test material always contained about 70 vowels.

In one procedure, the 64 channels of the signal FFT - thus a pre-version of the 20-channel feature vectors - were submitted to frequency dynamic warping [8]. The according warping path is derived from the alignment of the long-term spectra (over the 20 to 25 s) of the new and a reference speaker. As for data hitherto analyzed, no significant improvement could be gained.

The method of linear vector transformation from the reference speaker's pattern space into that of the new speaker via a common pattern space is known from Choukri and Chollet [9]. A transformation matrix was calculated for the vowel inventory. Improvements of up to 20% as compared to cross-speaker recognition were obtained for data from different pairs of speakers, thereby ranging about 10% below the speaker-dependent rates.

The third alternative comprises feature vector adaptation by means of a neural network. A multi-layer perceptron was trained with speech vectors from the reference speaker as input and the corresponding vectors from the new speaker as desired output by error back propagation [10]. The so-trained network is used to modify the vowel inventory of a reference speaker according to the particular pronunciation of the new speaker.

The network structure performing best among those investigated up to now was one with two hidden layers, resulting in a more than 20% higher recognition rate for speaker-adaptive vowel recognition in contrast to cross-speaker rates. It should be noted though that such achievements are strongly dependent on the speakers, the training material and the adjustment of network training parameters.

In current investigations, the effectiveness of the listed methods for consonant clusters is quantified.

5. SUMMARY

The individual methods and algorithms implemented in the modules of a comprehensive speaker-adaptive system for speech recognition from large vocabulary have been illustrated. Several advantages of the particular type of subword units for several system components were elucidated. The vocabulary can be generated and arbitrarily enlarged from written text input to practically unlimited size. Pronunciation variations are accounted for in the automatically produced lexicon entries.

For use with large vocabularies, a preselection process can be activated in the system to accelerate the search process. Fine classification is performed by a dedicated three-dimensional dynamic matching procedure. It was further demonstrated that recognition performance for new speakers can be increased by appropriate adaptation algorithms. Results for individual system components were given.

ACKNOWLEDGEMENTS

The project has benefited from software encoding by G. Thierer and from a discussion on CVC units by G. Ruske from the Technical University of Munich, FRG. The work described was partly supported by the German Ministery for Research and Technology (BMFT) under contract no. 413-5839-ITM880101. Only the authors are responsible for the contents of this publication.

REFERENCES

[1] Ruske, G. and Schotola, T., IEEE ICASSP (1978) 722-725.
[2] Zwicker, E., Terhardt, E. and Paulus, E., J. Acoust. Soc. Amer. 65(2) (1979) 487-498.
[3] Schotola, T., Speech Communication 3 (1984) 63-87.
[4] Fesseler, P., Hackbarth, H., Kugler, M. and Boehm, A., Proc. Eurospeech (1989) 75-78.
[5] Jekosch, U. and Becker, T., Informationstechnik it 31(6) (1989) 400-406 (German).
[6] Kobayashi, Y. and Niimi, Y., IEEE ICASSP (1985) 41.14.1-4.
[7] Schuhmacher, K., NTG-Fachberichte 94, Sprachkomm. (1986) 15-20.
[8] Ainsworth, W.A. and Foster, H.M., The use of dynamic frequency warping in a speaker-independent vowel classifier, in: DeMori, R. and Suen, C.J. (eds.), New Systems and Architectures for Automatic Speech Recognition and Synthesis (NATO ASI Series, F16, Springer, Berlin, 1985), 389-403
[9] Choukri, K. and Chollet, G., Computer Speech and Language 1 (1986) 95-107.
[10] Rumelhart, D.E. and McClelland, J.L., Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1 (MIT Press, Cambridge, MA, 1986).

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

RECOGNITION OF NUMBERS BY USING DEMISYLLABLES AND HIDDEN MARKOV MODELS

J.B. Mariño, A. Bonafonte, A. Moreno, E. Lleida, C. Nadeu, E. Monte

Department of Signal Theory and Communications
Politechnic University of Catalonia. Spain.

Abstract

A continous speech recognition system (called RAMSES) has been built based on the demisyllable as phonetic unit and tools from connected speech recognition. Speech is parameterized by band-pass lifted LPC-cepstra and demisyllables are represented by hidden Markov models (HMM). In this paper, the application of this system to recognize integer numbers from zero to one thousand is described. The paper contains a general overview of the system, a description of the HMM training procedure and an assessment on the recognition performance in a speaker independent experiment.

1. INTRODUCTION

During the last two years, a continous speech recognition system based on demisyllables and discrete hidden Markov models (HMM) has been built in our laboratory. Demisyllables afford a convenient phonetic coding of Spanish utterances, according to the syllabic character of this language. Hidden Markov models have been shown to be a successful tool for describing in a probabilistic way the acoustic features of speech. Our system has been called RAMSES, Spanish acronym for "automatic recognition by means of semisyllables (demisyllables)". In this paper we provide a general overview of RAMSES and report its application to recognize the Spanish integer numbers from zero to one thousand, in both multispeaker and speaker independent tasks.

The paper is organized in the following way: in Section 2 the block-diagram of RAMSES is described, Section 3 addresses the task oriented aspects, in Section 4 the HMM training procedure is outlined, Section 5 is dedicated to report the recognition experiment results, and finally Section 6 contains the main conclusions.

2. RAMSES' OVERVIEW

Figure 1 shows a general block-diagram of the system architecture. The speech signal is band-pass (100 Hz - 3400 Hz) filtered by an antialiasing filter and sampled at 8 kHz. The utterance is isolated by an end-point detection algorithm and pre-emphasized. A linear prediction (LP) based parameterization follows: the signal is segmented into frames of 30 milliseconds by a Hamming window at a rate of 15 milliseconds, and every frame is characterized by a LP-filter with 8 coefficients. Afterwards, 12 band-pass lifted cepstrum coefficients are computed [1]; the energy of the frame completes the parameterization. Before entering the recognition algorithm, the system evaluates the spectral difference d(t) corresponding to the frame t by using [2]:

$$d(t) = \sum_{k=-2}^{2} k\, s(t+k)$$

where s(t) is the ceptstral vector in frame t. This difference implies a time-average along 90 milliseconds. In a similar way, the energy difference e(t) is calculated. The spectral vector and the spectral and energy differences are vector-quantized separately; in that way, every frame of speech signal is represented by three symbols.

According to the most recent proposals, RAMSES considers energy and time evolution information. However, in our system, the energy is not used directly as a parameter of the signal. This is because the energy depends on the prosody of the sentence and the intensity of the utterance, two very fluctuant features of speech. On the contrary, if the energy is expressed by a logarithmic measure, its difference does not vary with a change in the intensity of the overall sentence, and the variation due to prosodic effects is greatly alleviated.

This work was supported by the PRONTIC grant number 105/88

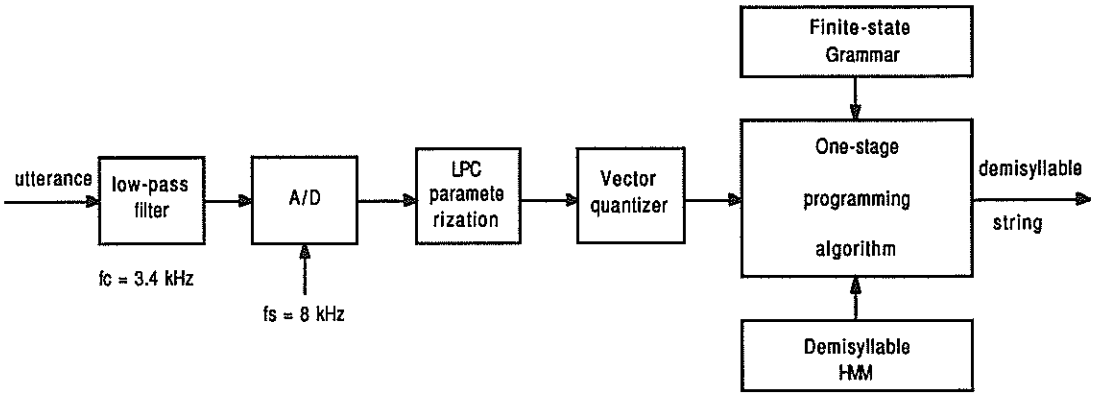


Figure 1.- Recognition System architecture

The recognition algorithm performs an one-stage dynamic programming (described for connected word recognition [3]) driven by a finite state grammar. So, the algorithm computes the string of demisyllable models that provides the most likely path of states throughout the utterance and, at the same time, satisfies the grammar constraints. If necessary, a dictionary provides the semantic meaning of the issued sequence of demisyllables.

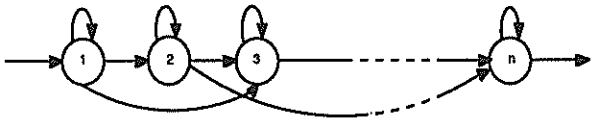


Figure 2.- HMM structure

Figure 2 shows the structure used for the hidden Markov models. It is a typical left-to- right structure, that allows to skip one state when the model makes a transition between states. The emission of symbols is associated to the states, that issue three independent symbols (spectrum, spectrum difference and energy difference) when they are visited. The number n of states is a parameter to be determined. During the recognition task, the transition probability between the final state of a model M_i to the first state of the following model M_{i+1} is determined by the duration probability of the demisyllable modeled by M_i. The length probability of a demisyllable is parameterized by the mean and the variance of a gaussian distribution.

This general architecture can be oriented to a specific application by designing the regular grammar and training the demisyllable models.

3. NUMBER RECOGNITION TASK

The recognition of numbers is an interesting task to try a recognition system. Besides its practical interest, this application exhibits an inherent difficulty due to the subtle acoustic differences that can separate very distinct semantic meanings.

In a previous experiment on number recognition in a speaker dependent environment, the set of necessary demisyllables for this application was stablished. This set, including 67 demisyllables, was designed in order to cope the most usual realizations for numbers issued by Spanish speakers. In that same experiment, the finite state grammar describing the numbers in terms of demisyllable strings was built; the number of states of this grammar was 118. Details on demisyllable definition and grammar inference can be found in [4] and [5].

From ten speakers (6 male and 4 female) a speech data base was acquired in our laboratory. Every speaker uttered one realization of a set of 44 numbers, designed in such a way that: a) every set included at least two samples of each necessary demisyllable in the application, and b) the 44 numbers performed a suitable sampling of the integers from zero to one thousand. The articulation rate of speech spanned from 5 to 7 syllables per second. This data base was segmented by hand into demisyllables and labeled.

4. DISCRETE HIDDEN MARKOV MODEL TRAINING

Demisyllable models were trained following the procedure outlined in Figure 3. Each model was trained independently of the others. Let D_i be the demisyllable which model has to be trained. Every

```
for every demisyllable Di
   - for every speaker
      - collect the samples of Di
      - if the number of samples > 5
         - perform a k-means clustering
         - select 5 representants of Di
      - end if
   - end for
   - train HMM by Baum-Welch algorithm
   - smooth HMM
end for
```

Figure 3.- HMM training algorithm

sample of D_i was collected from the utterances recorded by the first speaker; if the number of samples surpass 5, the 5 most representative samples were selected by a k-means clustering procedure [6]. This strategy aimed to prevent a very dissimilar training for demisyllables with a number of representants very different. Once the samples from every speaker were obtained, the Baum-Welch estimation algorithm was applied. At the same time, the mean and the variance of demisyllable length was computed. Finally, the demisyllable models were smoothed according to the co-occurrence probability method introduced in [7].

Previously to apply this procedure, the values for some important parameters of the models had to be determined, i.e., the size of the three codebooks and the number of states. In order to assess the choice, some training and recognition experiments were accomplished. Specifically, the value for those parameters were fixed and the models were trained with the ten speakers; afterwards, the signals in the data base were recognized. Then the parameter values were modified, and the training and recognition procedures were carried out; and so on. As a result of these trials, we drew the following conclusions:

a) as far as the size of the codebooks is concerned, the most suitable choices are: 64 for the two codebooks dedicated to spectral information and 32 for the codebook devoted to energy differences. Although similar performance can be got with other parameter selections, this option requires the minimum codebook sizes.

b) the recognition performance is noticeably dependent on the number of states of the hidden Markow models. Several criteria to determine the most suitable number of states for every model were tested: equal number of states, number of states according to the number of sounds included in the demisyllable, and number of states as a function of the average length of the demisyllable. This third criterium yielded the best performance for almost every experiment carried out, and when it did not lead to the best choice, it afforded a performance near the optimum. As a consequence, we used this criterion in our finaldesign. In Table 1 the definition of the average length criterion is provided.

average length in frames	number of states
≤4	2
5,6	3
7,8	4
9,10	5
>10	6

Table 1.- Criterion to select the number of states of HMM as a function of the average length of demisyllables

5. SPEAKER INDEPENDENT EXPERIMENT

Although we acknowledge that our data base is rather reduced, we were interested in carrying out some experiments that allowed to ascertain the ability of RAMSES to cope with speaker independent tasks. To this aim, we made six different training and recognition trials. In each experiment we trained the system with 8 speakers, and then we recognized the speech signals of the other two; in every case, the outside training speakers were taken with different sex. Table 2 shows the six couples of outside training speakers.

M1 - F1
M2 - F2
M3 - F3
M4 - F4
M5 - F1
M6 - F2

Table 2.- The couples of outside training speakers

Table 3 provides the recognition error percentage achieved for every speaker, when he or she was inside and outside the training set. The results before and after the smoothing of the HMM output probabilities is also shown. We count as one error every number recognized incorrectly, indepently of the number of demisyllables misrecognized. It is worth mentioning that, in most of the cases, the errors affected only one digit in the number (corresponding either the hundreds, or the tenths or the units); for instance, 677 was recognized as 637, or 721 as 621.

	before smoothing		after smoothing	
	trained	not trained	trained	not trained
M1*	0.4	0.7	0.0	0.0
M2	2.3	6.8	2.7	6.8
M3	1.4	4.5	3.2	4.5
M4**	0.0	12.1	2.7	6.9
M5	0.0	27.3	6.4	11.4
M6	0.0	2.3	0.5	0.0
F1	0.0	9.0	0.0	3.4
F2	0.0	3.4	0.0	0.0
F3	0.0	4.5	2.3	2.3
F4	0.0	0.0	0.0	0.0
M	0.6	7.0	1.9	3.8
F	0.0	4.9	0.6	1.5
Total	0.4	6.1	1.5	2.8

The total number of utterances of this speaker is: * 136, ** 58

Table 3. Recognition error percentage

From Table 3 we can observe the following facts:

a) the smoothing afforded a remarkable decreasing of recognition errors outside the traininig set; however, the prize to be paid was an increasing of recognition errors inside the training set.

b) For some speakers (for instance, M4, M5 and F1) the RAMSES recognition ability was very different when the speaker was either inside or outside the training set.

c) The recognition performance fluctuated greatly from some speakers to others. This behaviour is much more evident for the speakers outside the training set.

d) The average performance (2.8% of error percentage) was satisfactory.

6. CONCLUSION

Our interpretation of the enunciated observations is twofold. Firstly, RAMSES seems suitable for recognizing number in continous speech, either in a multispeaker task (every speaker inside the training set) or in a speaker independent application (all of speakers outside the training set); and secondly, the data base required for training this latter case must be increased.

Currently, we are recording a new data base, involving 20 new speakers and utterances with strings of integer numbers from zero to one million.

REFERENCES

[1].- B. H. Juang et al.,"On the use of bandpass liftering in speech recognition", IEEE Trans ASSP-35, pp. 947-954: July, 1987

[2].- S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans ASSP-34, pp. 52-59: February, 1986

[3].- H. Ney, "The use of an one-stage dynamic programming algorithm for connected word recognition", IEEE Trans ASSP-32, pp. 263-271: April, 1984

[4].- J. B. Mariño et al., "Finite state grammar inference for connected word recognition", Proc. EUSIPCO'88, pp. 1035-1038: September, 1988

[5].- J. B. Mariño et al., "Recognition of numbers and strings of numbers by using demisyllables: one speaker experiment", Proc. EUROSPEECH'89 vol. 1, pp. 102-105: September, 1989

[6].- J. Wilpon, L. R. Rabiner, "A modified k-means clustering algorithm for use in isolated word recognition", IEEE Trans ASSP-23, pp. 587-594: June, 1985

[7].- K. -F. Lee and H. -W. Hon, "Speaker-independent phone recognition using hidden Markov models", IEEE Trans ASSP-37, pp. 1641-1648: November, 1989

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Word Verification in Continuous Speech by means of Demisyllable Synthesis

Jorge Romano-Rodríguez

Corporate Research and Development, ZFE IS KOM 3, Siemens AG
Otto-Hahn-Ring 6, D-8000 München 83, West Germany

A top-down driven module verifies word hypotheses in a bottom-up recognition system by matching synthetic word patterns against suitable intervals of the unknown utterance. The word patterns are synthesized from a reduced inventory including 600 german demisyllables. An additional module predicts long words after parts of them have been hypothesized. The verification system duplicates the word recognition rate and allows also a reduction of the hypotheses amount.

1. Introduction

This paper describes a word verification module within the SPICOS system (Siemens Philips IPO Continuous Speech Recognition and Understanding), a German language man-machine dialogue interface [1]. As in most speech recognizers for continuous speech, the incoming speech signal is first segmented and classified into subword units. The resulting segment hypotheses are then combined step by step to valid word hypotheses and finally to linguistically correct phrases and sentence hypotheses.

A top-down driven word verification module is added to the above mentioned bottom-up recognizer. Word patterns are synthesized from demisyllables for every bottom-up word hypothesis and then matched as a whole against the corresponding parts of the unknown utterance. The rescoring of the hypotheses with the verification distances yields higher recognition rates than the original bottom-up scores. This allows also an effective reduction of the large amount of word hypotheses.

The motivation for this additional 'verification-by-synthesis' method results from the fact that we expect to be able to better model the context of the speech units than the bottom-up module for three reasons. First, most of the the local context (coarticulation) of the different phones is stored implicitly in the reference demisyllables, since a german demisyllable includes generally 2 or 3 phones with their transitions. Second, the global context, e.g. syllable position in the word, neighbouring syllables, prosody, etc., is taken into account explicitly by applying synthesis rules, that modify the demisyllables before they are concatenated to larger patterns, e.g. words. Third, different word pronunciations can be taken into account in an appropiate pronunciation lexicon, allowing the synthesis of alternative word patterns. These knowledge sources and the architecture of the verification module are depicted in fig. 1. It shows also the long-word prediction module described in section 4.

2. Synthesis of word patterns with a reduced demisyllable inventory

The word patterns used for the verification match are synthesized with demisyllables (DSs) as concatenation units. If a syllable is splitted 60 ms after the vowel onset, we get an **initial demisyllable (IDS)**, including the initial consonant(s) and the vowel onset, and a **final demisyllable (FDS)**, including the remaining major part of the vowel an the final consonant(s). If a FDS ends with /t/, /s/, /f/, /ʃ/ or a combination thereof, this part is split off and treated as a **suffix** while the remainder of the FDS is defined to be a **rudiment** [2]. The number of IDSs needed results from the number of initial consonant clusters multiplied by the number of different vowel onsets; the number of FDSs is the number of vowels multiplied by the number of final consonant clusters.

In order to reduce the number of DSs needed to synthesize the patterns, we start with an inventory proposed for a German speech synthesis-by-rule system [3], that originally included 1286 DSs. Phonotactic restrictions, that had not yet been considered there, allow the elimination of 240 FDSs without any loss of quality, because these FDSs do not appear in the German language.

A further reduction of the inventory is based on the replacement of acoustically similar DSs. The parametric similarities between DS-pairs are measured with a DP-algorithm, the same that is used for the verification. On the one hand, the shortness of the vowel onset in the IDSs allows the replacement of 200 IDSs, leaving only 350 IDSs in the reduced inventory. On the other hand, the 248 rudiments can be replaced by their counterpart DSs only cutting their last 90 ms. In summary, the **reduced demisyllable inventory** for pattern synthesis includes merely 598 DSs and 17 suffixes [4].

The DSs concatenation rules were proposed in [2,3] for a natural and intelligible speech output. Naturalness is not a goal in the verification system. Hence, a part of the concatenation rules can be simplified for a faster verification[4].

3. Word hypotheses verification

In order to test the efficiency of a verification with synthetic patterns we tried to begin with words as verification units. The reasons for that are twofold: on the one hand, words are the smallest units that allow to meaningfully apply the DS concatenation rules; on the other hand, the word hypotheses level is a well defined interface in most recognition and understanding systems. This allows an easy and direct comparison between the recognition rates of the bottom-up module and the top-down verification.

The comparison method is based in a verification of all word hypotheses generated by the bottom-up module, rescoring them with the verification distances obtained at the parametric level (18 LPC-cepstrum-coefficients) by means of a classic DP-match between the synthetic word patterns and the corresponding intervals of the unknown utterance. The lengths of the word patterns are kept constant during the match, because their DSs are segmented by hand. But the boundaries of those utterance intervals, which are proposed by the bottom-up module, can be relaxed during the match, because the bottom-up segmentations need not agree exactly with word boundaries. Hence, a jitter around all the interval boundaries is convenient (fig. 2a).

Due to alternative segmentation and classification in the acoustic-phonetic module [4], multiple hypotheses of the same word can appear at neighbouring segment borders. In order to accelerate the verification, these related multiple hypotheses are grouped and only verified once, but with an increased DP-region, which overlapps the single regions (fig. 2b). The resulting variable jitter of the DP end points depends on the extreme boundaries of the hypotheses in a group.

In order to increase the number of correctly hypothesized words, alternative pronunciations of some words are also considered during the word hypotheses generation. In the experiments reported here, only frequent intraword variants are regarded, but not interword assimilations. The hypotheses that belong to alternative variants can be generally verified with word patterns synthesized from the standard phonetic transcription, because the deviations between most of the different pronunciations which we regarded are small. Alternative word patterns are synthesized only if strong deviations from the standard variant occur, e.g. elisions of syllables.

4. Long word prediction and verification

Synthetic patterns can be more useful for the verification of larger speech units, e.g. phrases or other word chains, than for single words, because a larger context can be taken into account during the concatenation of the DSs. In this section we propose the use of lexical knowledge for predicting long words.

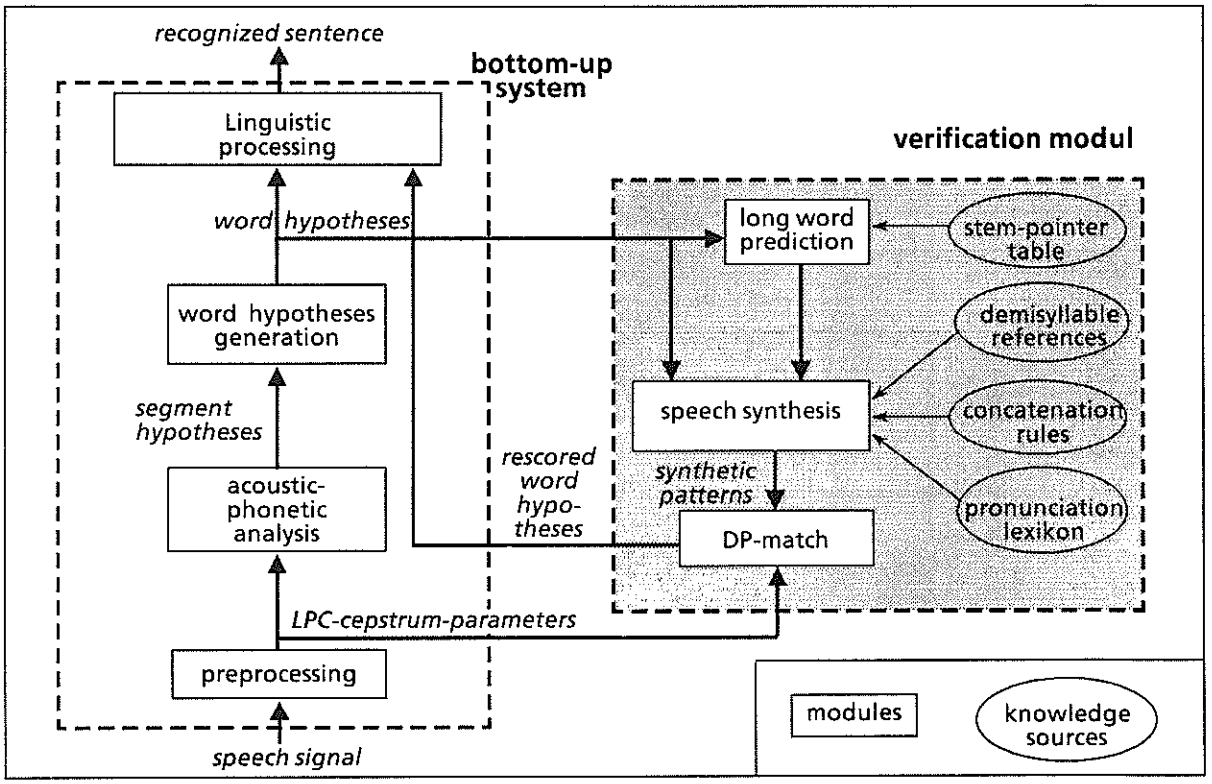


Fig. 1: Architecture of the recognition system with the verification module and its knowledge sources

The reason for this is twofold: firstly, reducing the computational effort in the bottom-up module, and secondly, increasing the word recognition rate.

The German language presents a large number of long words, mainly due to the frequent use of compounds. Because they can have many different pronunciations, it would be costly to model all the possible variants during the word generation. Nevertheless, parts of these long words are always realized in standard pronunciation. For example, the word "1984", which has 8 syllables in German standard pronunciation, can also be pronounced with 7, 6 or 5 syllables. However, the *parts* "19" and "80" are generally present in all of these pronunciations. Thus, we would be able to predict the whole long word if any one of these two stems had been hypothesized. Hence, a prediction module is added into the verification system with the aim of catching long words. If certain syllable sequences of long words are hypothesized with good scores, the pertinent long words are predicted and afterwards verified as described in section 3.

4.1 Preparation of the lexical knowledge

All of the words of the vocabulary with more than 4 syllables are defined as **long-words** and the *stable* syllable sequences, which are parts of long-words, as **acoustic stems**. These stems must have at least two syllables. We tried defining as few stems as possible, but enough to predict all of the long-words. First we have chosen as stems words already in the lexicon; then additional new stems were defined, including a few semantically meaningless ones [4].

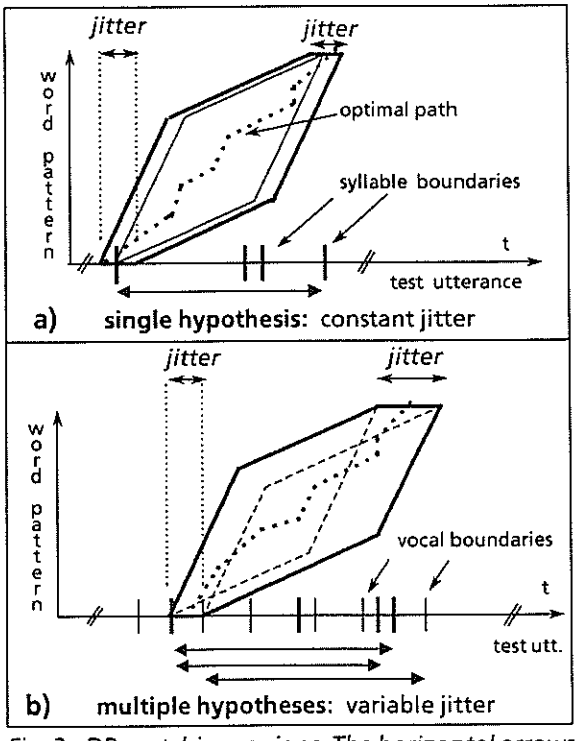


Fig. 2: DP-matching regions. The horizontal arrows represent the extension of the word hypotheses.

In order to reduce the computations, the word generation module works now with a **short word lexicon**, created by deleting the long-words from the original lexicon and including the newly defined stems. As crucial element for the prediction, a **stem-pointer table** was also created. Each acoustic stem is entered in this table with pointers to all the long-words that contain it. Several long-words have more than one pointer from different stems and are entered with two or more pronunciations with different numbers of syllables. The table also indicates for each word variant its permissible positions relative to the syllabic position of the corresponding stem (e.g. "1984" begins in the same syllable boundary as its stem "19" and can end between 3 and 6 syllables after, depending on the variant).

4.2 The prediction and verification steps

The prediction and verification module uses as input all the word hypotheses generated by the bottom-up word module with the short word lexicon. It carries out the following steps:

* parametric verification of all hypothesized stems and selection of those with a distance below a given threshold;
* prediction of the corresponding long-words by looking at the stem-pointer table;
* DS based synthesis of the patterns for the long-word variants that fit into permitted positions of the utterance;
* parametric verification of all the synthesized long-word patterns with the DP-algorithm;
* selection of the best long-word hypotheses by means of the following heuristic strategy. First, in the case that various pronunciation variants from the same word fit, only the best variant is taken. Second, only those hypotheses are considered which have a verification distance below a certain threshold. Third, if several long-words from the same stem are left after step 1 and 2, only the two bests are kept.

The remaining long-word hypotheses together with their distances and their begin and end points are then passed to the linguistic module for processing in combination with the short word hypotheses.

5. Results

The described verification system has been tested on two recording sessions of a set of 200 sentences (including 1391 words). Both sessions were spoken fluently by the same male speaker.

Fig. 3 shows the word recognition rates for the second session. The bottom-up recognition rate increa-

ses nearly linearly with the logarithm of the number of hypotheses considered at each time instant, while the word rate for the verification shows a certain saturation for n>4. The DP-jitter allows a monoton increment of the recognition performance for values up to 120 ms, where the recognition rate for the best hypotheses (top1) is about two times better than that for the bottom-up approach. It must be kept in mind, that the verification without prediction can not find new words; hence, the rates for "top all" must be equal for both the bottom-up and top-down modules.

The upper line in fig. 3 corresponds to the case of verification of short words and additional long-word prediction. The underlying stem-pointer table comprises 88 stems and about 250 long-word entries. Because the 200 sentences include only 111 long-words, the maximum achievable increment of the word recognition rate by means of the prediction is about 8% (however this corresponds to 21% of the syllables!). The prediction module catches about 80% of these long-words. It generates about 11 long-word hypotheses with permitted position per sentence; after applying the described strategy only about 2 are kept. The computation time in the word generation module is reduced by about 20% due to use of the short word lexicon.

The verification of whole words allows not only higher word recognition rates but also an effective reduction of the amount of hypothesized words. Setting an adequate distance threshold to separate good (correct) from poor (mostly false) hypotheses and simultaneously considering only the best 8 candidates at each time instant, the average number of word hypotheses per sentence can be reduced from about 370 down to 150 while the percentage of recognized words decreases merely from 76% to 73%.

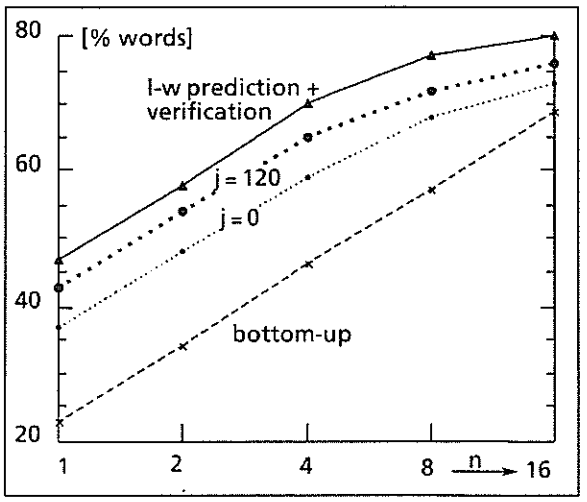


Fig.3: Word recognition rates depending on the top n hypotheses: dashed for the bottom-up module, dotted for the verification (without and with 120-ms-jitter) and upper line for the verification + long-word prediction.

5. Conclusions

The verification of word hypotheses with synthetic patterns allows an noticeable improvement of the system performance. However, interword assimilations must also be taken into consideration. This would require a special verification strategy: assimilated words, e.g. function words, could not be further verified with the standard word patterns, because also their context has to be regarded.

The described approach needs to be further tested with other speakers. Nevertheless, the verification can also be applied to larger speech elements than words, e.g. phrases, allowing the modelling of further phonological effects at word transitions. A preliminary experiment with all pairs of neighbouring monosyllables in the 200 test sentences shows that the verification distances for the pair-patterns are clearly less than those for the isolated monosyllable patterns. Thus, the prediction method could also be employed on word chains.

Acknowlegdements

The author wishes to thank his colleagues from the Siemens speech recognition group as well as Dr. G. Ruske from the Technical University Munich for their support.

References

[1] **Brenner, M.; Höge, H.; Marschall, E.; Romano, J.:** Word Recognition in Continuous Speech using a Phonological Based Two-Network Matching Parser and a Synthesis Based Prediction. Proc. ICASSP 1989, Edinburgh, Vol. S1, S. 457-460.

[2] **Dettweiler, H.; Hess, W.:** Concatenation rules for demisyllable speech synthesis. Proc. ICASSP 1985, S. 752-755.

[3] **Dettweiler, H.:** Automatische Synthese deutscher Wörter mit Hilfe von silbenorientierten Segmenten. Dissertation, Tech. Univ. München 1984.

[4] **Romano, J.:** Verifikation von Worthypothesen durch Halbsilbensynthese bei der automatischen Erkennung fließender Sprache. Dissertation, Tech.Univ. München 1989.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Admissible Strategies for Reducing Search Effort in Real Time Speech Recognition Systems

L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, and D.S. Kanevsky
IBM Research Division
T.J. Watson Research Center, P.O.Box 704, Yorktown Heights N.Y., 10598

Abstract

Given a method for evaluating the acoustic match between a word and a segment of an acoustic signal, the paper describes a general method for attaching acoustic scores to clusters of words. This provides a method for implementing a search tree in the process of selecting a list of candidate words that fit well to a given acoustic signal and therefore produces significant reduction of computation time. The important feature of this method is that it guarantees admissibility, in other words, the number of errors that occur is not larger than the number of errors obtained by doing a standard computation of a list of candidate words from all words in the vocabulary. This method is applied to one specific fast acoustic match method.

1. INTRODUCTION

In a large vocabulary automatic speech recognition system using hidden Markov models, such as the one developed at IBM Research, the output word is chosen to be the one that has the maximum posterior probability given the input acoustic observation. This involves calculating the likelihood of the observed acoustics given the models for each word in the vocabulary. However, when the vocabulary is very large and when the decoding has to be done in real time, we cannot afford to go through a detailed likelihood computation for every word in the vocabulary.

In this case we need a fast method for identifying a short list of candidate words that would end up with high likelihood scores and which contains the correct words almost all the time. Several different methods have been developed for producing short lists of candidate words. For example, a phonetic fast match is described in (Bahl, et al.,(1)) and a polling fast match in (Bahl, et al. (3)). Another fast match algorithm is presented in (Gupta et al. (2)). We will refer to the list of candidate words obtained by such a method as a fast match list.

One can note immediately the essential difference between these methods. In (Bahl, et al., (1)) and (Gupta, et al. (2)) the process of constructing a fast match list is organized in the form of a tree search that considers only a part (on average 1/10) of the vocabulary in order to find all candidate words. On the other hand in (Bahl, et al. (3)) the search for candidate words involves the whole vocabulary . The aim of this paper is to give a general method for organizing the search for candidate words for fast match methods like (Bahl, et.al.(3)) as a tree search procedure that provides further significant reduction of computation time. The important feature of this modification is that it guarantees *admissibility* . In other words, the number of errors that occur is not larger than the number of decoding errors that is obtained by doing a match computation on all the words in the vocabulary.

The rest of the paper is organized as follows. In Section 2 we describe a general principle for obtaining admissible strategies. In Section 3 we describe a specific design of an admissible strategy based on clustering words in the vocabulary and apply it to the polling fast match method (Bahl et.al. (3)). Experimental results are given in Section 4.

2. General Principle

Assume that we have the following standard fast match procedure. Let $F(\bar{f},w)$ be a (fast match) score attached to an acoustic label sequence $\bar{f}$ and a word w from the vocabulary V. Let us enumerate all words in the vocabulary V according to their fast match score:

$w_1, w_2, \ldots, w_n$, such that $F(\bar{f},w_i) \geq F(\bar{f},w_j)$ if $i < j$. Then for a given acoustic label sequence $\bar{f}$ we have the following fast match list

$$(1) \qquad L = \{w_i \in V \mid F(\bar{f},w_i) \geq \varepsilon, i \leq M\}$$

where ε is some threshold and M is the maximum size of a fast match list that is allowed. That is, we include all words in the fast match list whose scores are above a given threshold and whose rank according to the fast match scores does not exceed a certain limit.

Now let us introduce a modified fast match method. Assume that we have some estimator score $E(\bar{f},w)$ such that

$$(2) \qquad E(\bar{f},w) \geq F(\bar{f},w)$$

for any $\bar{f}$ and $w \in V$. Let us now enumerate words in the vocabulary in accordance with the estimator score: $w_{k_1}, w_{k_2}, \ldots, w_{k_n}$, where $E(\bar{f},w_{k_i}) \geq E(\bar{f}, w_{k_j})$ if $i < j$. Further, we construct the following estimator list:

$$(3) \qquad EL = \{w_i \in V \mid E(\bar{f},w_i) \geq \varepsilon\}$$

Let us note the following property of EL. If $w \in L$ then $F(\bar{f}, w) \geq \varepsilon$. By (2) $E(\bar{f}, w) \geq \varepsilon$ which in turn implies that $w \in EL$. Therefore, $L \subseteq EL$. Now, let us enumerate the words in EL according to the scores $F(\bar{f}, w)$ and construct the list

$$L' = \{w_i \in EL \mid F(\bar{f}, w_i) \geq \varepsilon, i \leq M\}$$

Because of the property of EL described above, it is easy to see that $L' = L$.

This procedure is useful if the computation of EL can be done very fast and the size of EL is much smaller than the size of V . In this case computation of the fast match list using estimators will be faster. Now we formulate exact criteria under which the modified method becomes faster.

Let t_1 and t_2 be the average time for computing the fast match and estimator scores respectively. Let d_1 be the size of V and d_2 be the size of EL. Then the execution time for the standard fast match method is $t_1 \times d_1$, whereas the execution time for the modified fast match method (using estimators) is $t_2 \times d_1 + t_1 \times d_2$. Dividing the first expression by the second, we get

$$\frac{t_1 d_1}{t_2 d_1 + t_1 d_2}$$

Defining $a = t_1/t_2$, the ratio of average time for computing the fast match scores to the average time for computing the estimator values for the same set of words, and $b = d_1/d_2$, the ratio of the number of words in the list V to the list EL, the above ratio becomes

$$(4) \qquad \frac{ab}{a+b}$$

Thus it makes sense to use the modifies method if on average we have $\frac{ab}{a+b} > 1$.

3. Clustering Method

Now we give a general method for constructing the estimator E . Given an acoustic label sequence $f_1^T = f_1, \ldots, f_T$ let $F_x(f_1^T, w)$ be a score function attached to word w for that label sequence, where $\mathbf{x} = \{x_{i,w}\}$ is a set of parameters (depending on label i and word w). Let us divide words in clusters C_i, $i = 1, \ldots, k$, in some way, requiring that for different words w_j, w_k in the same cluster C_i their score functions $F_x(f_1^T, w_j)$ and $F_x(f_1^T, w_k)$ are identical (thus acoustic scores of these words will be different only because of different parameters). Assume also that for any w the following inequality holds:

(5) $F_y(f_1^T, w) \geq F_x(f_1^T, w)$ if $y_{i,w} \geq x_{i,w}$ for some i

This in particular holds if F_x is a polynomial of $\mathbf{x}$ with nonnegative coefficients.

We define

$$F_C(f_1^T) = \max_{w \in C} F_x(f_1^T, w)$$

Now we can define estimator score $E(\bar{f}, w) = F_C(\bar{f})$ where $w \in C$ and apply the modified fast match strategy that was described in the last section.

We describe this procedure on the example of the polling fast match (Bahl et al. (3)). The polling fast match is an exceptionally fast method for producing a short list of candidate words. Scores in the polling fast match are given by the formula

$$F(f_1^T, w) = \sum_{i=1}^{T} a_w(f_i)$$

Here $a_w(f_i)$ are votes associated with each label and a word and are precomputed. If M is the total number

of unique labels that can occur in f_1^T and n_j is the number of times the label j occurs in f_1^T then the above score can be written as

$$F(f_1^T, w) = \sum_{j=1}^{M} n_j a_w(j)$$

In order to construct an estimator of this formula we proceed as follows. Let us divide words in the vocabulary into clusters in some way. For example, each cluster might contain k acoustically similar words. Assuming for simplicity that k divides n, the size of the vocabulary,

$$C_1 = \{w_1, \ldots, w_k\}, C_2 = \{w_{k+1}, \ldots, w_{2k}\}, \ldots, C_m$$

Let **C** be the set of all clusters $C_1, \ldots, C_m$. For each cluster $C \in \mathbf{C}$ let us define

$$\bar{a}_C(f_i) = \max_{w \in C} \{a_w(f_i)\}$$

$$E(f_1^T; C) = \sum_{i=1}^{T} \bar{a}_C(f_i)$$

$$E(f_1^T; w) = E(f_1^T, C), \; w \in C \tag{6}$$

Clearly, we have $E(f_1^T;w) \geq F(f_1^T;w)$. The running time required for computing the estimator score for each cluster in **C** is n/k times less than the time that is required for computation of $F(f_1^T,w)$ for all $w \in V$.

From equation (4) it immediately follows that given k the modified method is faster only if the following inequality holds for the number of words in an estimator list:

$$\#EL < \frac{k-1}{k} \times n \tag{7}$$

Varying the number of words in clusters one can organize the tree search scheme as follows.

Let each cluster in one level contain k words and for the next level let each cluster contains k_1 words where $k > k_1$. For the first level we can construct the estimator list EL using the rule (3) with the score (6), then we can construct another estimator list applying rule (3) (with the estimator score corresponding to clustering with k_1 words in each cluster) to EL instead of V and so on (see the example on Fig.1). This gives us a method of organizing the candidate list construction into a tree search scheme. If we discard one cluster then no subdivisions of that cluster need be examined further.

Remarks

The idea of an admissible strategy that is close to the one suggested here was used in (Bahl et al., (4)). But there is an essential difference between the methods used for constructing the estimators here and in (Bahl, et al. (4)). Here we reduce the amount of computation for estimator scores for all words in vocabulary in fact by reducing the number of words to be considered (words in the same cluster have the same score). Contrary to this, in (Bahl et al. (4)) we reduced the amount of computation for the estimator scores for each word in a vocabulary, but all words in a vocabulary have to be considered. It is clear that only the method suggested here provides a general procedure for reducing search effort for any match algorithm.

4. Experimental results

Here we present results of experiments involving application of this technique to the polling fast match. Since the polling fast match by itself is very fast (especially when a vector processor is used), it was interesting to see if this method could speed it up further.

A series of experiments for 2 speakers (good and poor with respect to recognizer performance) were carried out for the 5000 word vocabulary isolated speech recognition task, comparing the the polling fast match described in (Bahl et.al (3)) and the strategy suggested in this paper. Each speaker uttered 100 sentences containing about 1700 words in all. In these experiments words were divided into clusters of two and three words simply by running all words sequentially and putting adjacent pairs or triples of words in one cluster. This way many acoustically dissimilar words were put in the same cluster (which may not be very desirable). The speedup obtained is shown in the table below for the two speakers using two different cluster sizes. As guaranteed, the modified method did not introduce any additional errors.

It was observed that only a few words from L are present in EL with low E values and that choosing thresholds ε in such way that these words may not be in EL (with a slight increase in the error rate as a result) led to essentially shorter EL and resulted in speeding up the decoding procedure by a factor of 3 to 4. Thorough analysis of clusters containing these exceptional words shows that they contain highly acoustically dissimilar words. This gives evidence that more accurate division of words into clusters could

provide further improvement of the decoder. By introducing new clusters with k equal to 4 or 5 one can improve the ratio of the average time to compute the fast match score for all the words to that for computing the estimator score for all the words (which depends on this clustering and tree search procedure). Thus a possibility exists to make trees with more levels than was possible in the current experiments. By choosing acoustically similar words to be in the same cluster it is expected that this procedure will be more beneficial.

References

1. Bahl, L.R, Gopalakrishnan P.S., De Gennaro, S.V., Mercer, R.L., "A Fast Approximate Acoustic Match for Large Vocabulary Speech Recognition",Proc. 1989 Eurospeech Conference, Paris, Sep. 1989.
2. Gupta, V.N., Lennig, M., and Mermelstein, P., "Fast Search Strategy in a Large Vocabulary Word Recognizer", J. Acoust. Soc. Am. 84 (6), December 1988.
3. Bahl, L.R., Bakis, R., de Souza, P.V., Mercer, R.L, "Polling: A Quick Way to obtain a short List of Candidate Words in Speech Recognition", Proc. ICASSP88, New York, April 1988.
4. Bahl, L.R., Gopalakrisnan, P.S., Kanevsky, D., Nahamoo, D., "A Fast Admissible Method for Identifying Short Lists of Words ", preprint.

Speaker	2 words per class	3 words per class
1	1.4	1.7
2	1.35	1.4

Table 1. Average speedup achieved using estimators

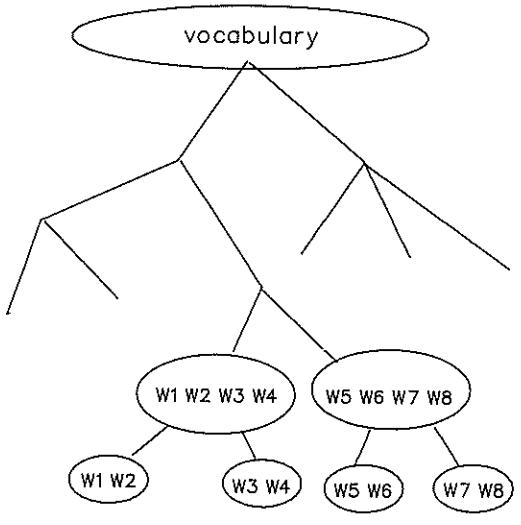


Figure 1. Tree made by clustering words

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

SOME EXPERIMENTS ON HMM STRUCTURE INFERENCE

Alessandro Falaschi °, Piero Pierucci °°

° Univ. di Roma 'La Sapienza', INFOCOM Dpt, Via Eudossiana 18, 00184 Roma, ITALY
°° IBM Italy, Rome Scientific Center, Via del Giorgione, Roma, ITALY

A method for finding a maximally representative graph structure for the acoustic events associated to linguistic units is presented. The method consists in a three-stage Dynamic Programming technique, applied to sets of acoustic label sequences collected for linguistic units, and obtained by Viterbi alignment of the incoming speech against the states of an Ergodic HMM. These sequences are first aligned in between them, and then collapsed into a reduced trellis. The trellis is then repeatedly explored by a second DP procedure which select a list of the more probable sequences. Finally, a third algorithm is utilized for dynamically build an HMM network.

1. INTRODUCTION

HHM are proved to be very effective in speech recognition tasks, due to their intrinsic properties, as the trainability on new speakers and environments, the ability in dealing with time, and the ease of integration in larger syntactic networks. This latter topic allows to implement minimum cost path search techniques as data driven algorithms [1], so giving the opportunity to change the models tophology once the data structure is preserved.

A problem which arises with HHM networks is that the computational complexity theoretically grows with the square of the total number of states. This is generally avoided by application of path pruning euristics, as the beam-search technique [2]; nevertheless, very large vocabulary recognition systems are forced to resort to word class preselection methods [3][4][5] in order to early reduce the number of hypothesis to be evaluated. As a consequence, isolated word pronunciation is recommended for facilitate the lexical class selection; moreover, the interaction among lexical class hypothesizing and word verification processes complicates the overall system architecture.

At a deeper sight, one can realize that the effectiveness of beam search techniques is founded on the ability of the models in accounting for every event which can occur for the modeled linguistic unit. In fact, poorly matched speech segments results in a great increase of the path hypothesis number to be evaluated; as a results, complex recognition tasks can be handled by one-stage [6] DP techniques only if very accurate HMM are available.

Model accuracy generally results from the compromise between the increase of the number of parameters for finer models and the amount of training data needed to obtain a reliable estimate of so many parameters. Many progresses and achievements have been obtained by utilization of proper observation density function definition, as continuous mixture densities [7], explicit state dwell probabilities [8], inclusion of dynamic features [9], perceptual signal processing techniques [10], and state-dependent feature transformation [11].

In particular, the use of continuous density mixtures allows multiple allophonic realization for each of the model states, but at the same time it weaks the model coherency, allowing, for example, to multiplex different allophonic realization without affecting the likelihood score assigned by the model. Moreover, the increase in the number of parameters to be estimated needs larger and larger databases to be used for training.

An alternative to mixture of continuous densities can be the utilization of several states in parallel, with a single pdf associated to each of them; in this way the resulting model allow to attach physical meaning to each state as local models of acoustical sub-events. Two main guidelines exists for approaching this goal, both based on the analysis of multiple realization of the same linguistic unit. The first is a "grow" one, because states are added to the actual model as a consequence of poorly matched phonetic occurrences [12][13]. The second choice is to start from several alternative models of the same unit, and to merge states if some similarity criterion is met [14][15].

The approach here proposed tries to build the HMM of linguistic units by a three-steps Dynamic Programming alignment of all the learning tokens available for the given unit. For this reason, it has some conceptual similarity with the approach described in [16], but its actual mechanism and input data are quite different.

The acoustic label sequences given in input are obtained by means of an Ergodic HMM Viterbi decoding of speech, as described in sect. 2. These sequences are firstly aligned to the same length by a nearest neighbor Dynamic Programming algorithm, as illustrated in Sect. 3. After having collapsed the aligned sequences in a reduced trellis, this is repeatedly explored by a DP minimum cost path search algorithm, as described in Sect. 4. At each iteration, the lowest probability decision on which the best path is based is marked, in order to avoid its choice in the following steps, thus obtaining different paths. Finally, another iterative Dynamic Programming scheme is utilized in order to build the HMM structure starting from the selected acoustic label sequences analysis.

An very important issue of the derived HMM is that the spectra associated to the states of the models can be shared among all the HMM which contain such label, achieving a substantial robustness improvement for the subsequent HMM parameters reestimation process.

2. SIGNAL ANALYSIS AND ACOUSTICAL LABELING

The experiments which follow are based on acoustical labels obtained by Viterbi alignment of speech against the states of an Ergodic HMM (EHMM). The adopted ergodic model is a 256 states, fully connected HMM [17][18], utilizing continuous autoregressive gaussian densities as observation densities [7]. Its parameters are estimated by the classical Baum algorithm on 440 isolated words belonging to a phonologically compact list [19], uttered by a single speaker. These words was yet automatically segmented in phonemic units, so that all the EHMM state label sequences related to the same phoneme can be easily collected.

The utilization of an EHMM for acoustical labeling produces two main advantages with respect to other techniques based on spectral changes criteria, temporal decomposition methods, or Vector Quantization techniques. The first is a greater stability of the labels with respect to the ones derived by VQ [18]. The second is the finite cardinality of the labels alphabet, opposite to the infinite one of spectral changes or temporal composition methods. As an example, Fig. 1 gives the reconstructed sonogram for an EHMM Viterbi decoded speech signal. Finally, it should be remarked that the EHMM derived, acoustical labels, represent yet fully trained HMM states, thus allowing the direct building of HMM topology with spectral characteristics shared among HMM of different units.

3. NEAREST NEIGHBOUR DP ALIGMENT

The acoustic label sequences $S_k(n)$, $n = 1, 2, ..., L_k$ collected for each linguistic units have to be warped to the same length before to further proceed in the HMM structure inference. This task is performed by a Nearest Neighbor version of the classical Dynamic Programming algorithm, aligning the n^{th} sequence against all the previously warped ones. As the sequence order of precedence may affect the results, sequences are ranked on the basis of a criterion which accounts jointly for the difference $|L_k - L^m|$ between the actual sequence length L_k and the average sequence length L^m, and the sequence entropy, obtained on the basis of the probability P(k) of each label k which occur among the sequences set. All the sequences will thus be aligned to the same length L_0.

The NN-DP alignment of sequence $S_k(n)$, $n = 1, 2, ..., L_k$ against the aligned sequences $S_h(l)$, $l = 1, 2, ..., L_0$, defined as

$$S_h(l) = S_h(w_h(m)); \quad h = 1, 2, ..., k-1; \quad m = 1, 2, ..., L_h$$

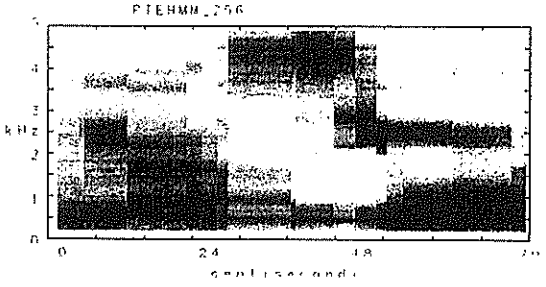


Fig. 1 - Synthetic sonogram after EHMM based Viterbi decoding of actual speech

is done by computation for each (n, l) indexes pair of a local best distance $D(S_k(n), S_h(l))$ as

$$D(S_k(n), S_h(l)) = d(S_k(n), l) + \min \{ D(S_k(n-1), S_h(l-1)), D(S_k(n), S_h(l-1)) + C_r(S_k(n)), D(S_k(n-1), S_h(l)) + C_o(S_k(n)) \}$$

where $C_r(S_k(n))$ and $C_o(S_k(n))$ are the repetition and omission costs for the label $S_k(n)$ defined as

$$C_r(S_k(n)) = \alpha |n - L_k|, \qquad C_o(S_k(n)) = \beta P(S_k(n))$$

and $d(S_k(n), l)$ is the nearest-neighbor spectral distance between the label $S_k(n)$ and the alignment index L defined as

$$d(S_k(n), l) = \min_h \{ d_{LR}(S_k(n), S_h(l)) \}$$

in which d_{LR} is the Likelihood Ratio [20] distortion measure between the spectra related to labels $S_k(n)$ and $S_h(l)$. For each (n, l) index pair, the index of the minimization term which has been chosen is stored, so that the retrieval of the warping path

$$w_k(n) \qquad \text{for} \quad n = 1, 2, ..., L_k$$

for the sequence $S_k(n)$ is possible, thus allowing to decode

$$S_k(l) = S_k(w_k(m)) \qquad \text{for} \quad m = L_k, L_k-1, ..., 1.$$

The positions above outlined address the alignment of $S_k(n)$ in such a way that it optimally matches the spectra of some of the previously aligned sequences, with the additional constraints of trying to avoid label omissions (specially for very frequent ones) as well repetitions (specially for the ones near to the edges of the sequence).

The aligned acoustical label sequences matrix is then collapsed in a reduced trellis, RT (n,l), where each label occurs only once for each column l. This trellis is homomorphic to the Count Trellis CT (n,l) whose elements give the number of labels which have been collapsed in position (n, l). Finally, NL (l) gives the number of labels which appear in column l, and NT(l) gives the sum over n of CT (n, l). Fig. 2a reports about the RT values after having processed 153 label sequences belonging to phoneme /m/, and fig 2b gives the relative label counts CT.

4. MODEL STRUCTURE INFERENCE.

4.1 Trellis pruning

All the paths connecting RT (n,0), for any n in (0-NL(0)), to RT (n, L_0), for any n in (0-NL(L_0)), give an allowable label sequence labels for the unit to be inferred; their number must be reduced before to infer a moderately complex HMM topology. This is done by associating a cost to each path, and retaining for topology inference only the best scored ones. The path costs for the reduced trellis are again obtained by DP, defining now as local score function the value

$$D_R(n, l) = I(n, l) + \min_m \{ I(n, l/m, l-1) + D_R(m, l-1) \}$$

to be calculated for $l = 1, 2, ..., L_0$ and $n = 1, 2, ..., NL(l)$. The quantity

```
Reduced Trellis:
*- 20   20   20   20   20   20   20   84  148  148
*-116   68   84  116   84   84   84  148   84   47
*-212  116  116   84  212  148  148   20   20   20
*- 84  212  212  212  116  212  212  212  212   84
*-100  100   68  148  148  116  116  116  116  212
*-148   84  148   68   68   68   68   68  132  168
*-180  148   19  180  256  256  180  132  256  116
*-114   52  228  256  180  180  256  256  100   83
*- 49  196  180  132  100  100  100  100   68   50
*-256  180  256  100  132  132  132  180   50  256
*-200  132  100    4  152  152   68   83  168  132
*-164  256   11  244   52   83   83  244  143  244
*- 68   11  244  152   11   52   52   68   83  143
*- 15  244  152   52  244  228  228   52  244   44
*-244  152  147   11  228   68  152   50  180  180
*-152   71  132  228   68   15   50  168   47   92
*- 99   68   52   68   83   11   11   11   68  159
*- 73   50   88   15   15  244  244  228   52  100
*- 72   39   15  196  196  196  196  196   11   88
*- 50   15  196  177  177  177  177  177  211   68
*- 11  200  177  147  147  147  168  164  196   52
*-236  177   97  131  131  131  147  152  177   24
*-199  114   87   97   97   97  131  147  164   11
*-178   97   72   87   87   72   97  143  152  211
*-161   87   39   72   72   50   72  131  147  196
*-146   72    9   50   50   39   44   97  131  178
*-133    9       39   39   24   24   72   97  177
*- 88         19    4    4   15   44   92  164
*- 39          9              4   24   72  152
*-241                           15   44  147
*-228                            4   24  131
*-188                               19  114
*-177                               15   97
*-143                                4   40
*-132                                    19
*- 97                                     1
*- 87                                     4
*- 67
*- 41
*- 37
*- 31
*- 16
*-  9
```

Fig. 2a - Reduced Trellis labels after examination of 153 label sequences for phoneme /m/

```
Trellis counts:
*- 52   36   28   26   25   25   24   25   26   29
*- 27   33   19   18   21   24   23   22   23   19
*- 24   25   18   17   16   19   22   20   17   13
*- 22   23   17   16   16   13   12   11    9   10
*- 12   14   15   13   13   13   12   10    8    6
*-  8   14   11   12   13   11   10    9    7    6
*-  7   11   10    7    5    5    6    7    5    6
*-  6   11    7    5    5    5    5    5    5    6
*-  6    7    7    4    5    5    5    5    5    6
*-  5    7    5    4    4    4    4    4    5    5
*-  5    7    5    4    3    3    3    4    4    5
*-  5    5    4    3    3    3    3    3    4    4
*-  5    4    3    3    3    3    3    3    4    4
*-  4    3    3    3    2    2    2    3    3    4
*-  3    3    3    3    2    2    2    3    3    3
*-  3    3    3    2    2    2    2    2    3    3
*-  3    2    3    2    2    2    2    2    2    2
*-  3    2    2    2    2    1    1    1    2    2
*-  3    2    2    1    1    1    1    1    2    2
*-  3    2    1    1    1    1    1    1    1    2
*-  3    1    1    1    1    1    1    1    1    2
*-  2    1    1    1    1    1    1    1    1    2
*-  2    1    1    1    1    1    1    1    1    2
*-  2    1    1    1    1    1    1    1    1    1
*-  2    1    1    1    1    1    1    1    1    1
*-  2    1    1    1    1    1    1    1    1    1
*-  2         1    1    1    1    1    1    1    1
*-  2         1              1    1    1    1    1
*-  1                             1    1    1    1
*-  1                             1    1    1    1
*-  1                                  1    1    1
*-  1                                  1    1    1
*-  1                                  1    1    1
*-  1                                  1    1    1
*-  1                                       1    1
*-  1                                       1    1
*-  1                                       1    1
*-  1
*-  1
*-  1
*-  1
```

Fig. 2b - Reduced Trellis counts for the labels shown in fig. 2a

$$I(n, l) = -\text{Log}_2 (CT(n, l) / NT(l))$$

gives the information received when label RT (n, l) occurs in position l and

$$I(n, l / m, l\text{-}1) = -\text{Log}_2 (N(RT(n, l), RT(m, l\text{-}1)) / N(RT(m, l\text{-}1)))$$

is the information borne by label RT (n, l) when occurring after label RT (m, l-1). The quantities N (.) and N (.,.) are the counts for the argument labels among the learning sequence set.
Again, a backpointer

$$B(n, l) = \text{argmin}_m (D_R(m, l\text{-}1) + I(n, l / m, l\text{-}1))$$

must be stored for the retrieval of the best sequence of Reduced Trellis warping indexes RW (l), given by

$$RW(L_0) = B(n^*, L_0); \qquad n^* = \text{argmin}_n \{ D_R(n, L_0) \}$$

$$RW(l) = B(RW(l+1), l+1); \qquad l = L_0\text{-}1, L_0\text{-}2, \ldots, 1$$

thus allowing to decode the best label sequence BS (l) as

$$BS(l) = RT(RW(l), l); \qquad l = L_0, L_0\text{-}1, \ldots, 1$$

During this last process, also the best path maximum cost decision index is spotted, i.e. the index pair $(n, m)^M$ for which

$$(n = RW(l^*), m = RW(l^*\text{-}1)) \qquad \text{with}$$

$$l^* = \text{argmax}_l \{ D_R(RW(l), l) - D_R(RW(l\text{-}1), l\text{-}1)) \}$$

These quantities give the Reduced Trellis transition whose contribution to the overall path cost is maximum, i.e. the label-transition pair whose probability is the smallest among those which appear into the found best path. Deletion of such transition allows to repeatedly apply the method outlined in this section to the Reduced Trellis array, getting an ordered sequence of best crossing paths for it in decreasing value of likelihood.
Fig 3 reports, as an example, the first 24 outcomes for the reduced trellis of Fig. 2, given in reverse temporal order. As can be seen, minimum probability pruning is able to evidence all the fine spectral variations occurring for the various phonemic contexts, exhibiting a very slow likelihood decrease for the selected sequence labels list.

```
Best Scored Paths:
26.54 ->  20  20  20  20  20  20  20  20  20  20  20  (9)
32.76 ->  84  84  84  84  84  84  84  84  84  84  84  (1)
34.40 -> 148 148 148  20  20  20  20  20  20  20  20  (7)
34.41 -> 148 148 148 148 148 212 212 212 212 212 212  (5)
34.53 -> 148 148 148 148  20  20  20  20  20  20  20  (6)
34.54 -> 148 148  20  20  20  20  20  20  20  20  20  (8)
34.60 -> 148 148 148 148 148 148 212 212 212 212 212  (4)
34.79 -> 148 148 148 148 148 148 148 212 212 212 212  (3)
34.89 -> 148 148 148 148 148 148 148 148 148 148 148  (1)
34.93 -> 148 148 148 148 148  20  20  20  20  20  20  (5)
35.02 ->  84  84  84  20  20  20  20  20  20  20  20  (7)
35.07 -> 148 148 148 148 212 212 212 212 212 212 212  (6)
35.10 ->  50  20  20  20  20  20  20  20  20  20  20  (9)
35.14 ->  84  84  84  84  20  20  20  20  20  20  20  (6)
35.15 -> 148  20  20  20  20  20  20  20  20  20  20  (9)
35.24 ->  84  84  84  84  84  20  20  20  20  20  20  (5)
35.30 ->  84  84  20  20  20  20  20  20  20  20  20  (8)
35.30 -> 148 148 148 148 148 148 148 148 212 212 212  (2)
35.54 ->  84  84  84  84  84  84  20  20  20  20  20  (4)
35.64 -> 148 148 148 148 148 116 116 116 116 116 116  (5)
35.68 ->  84  20  20  20  20  20  20  20  20  20  20  (9)
35.88 -> 148 148 148 148 148 148  20  20  20  20  20  (4)
35.90 -> 148 148 148 148 148 148 148 116 116 116 116  (6)
36.06 -> 148 148 148 212 212 212 212 212 212 212 212  (7)
```

Fig. 3 - First 24 outcomes given by the trellis pruning algorithm for data of fig.2. Sequences are written right-to-left, together with their score on the left and the maximum cost index on the right

4.2 Topology building

The selected label sequences list must be further processed in order to get a graph structure for the HMM to be inferred. This is again done by a DP algorithm, which task is now to align the labels of the selected k^{th} sequence $SS_k(n)$ with the ones of the actually build graph topology, an then update the graph by adding arcs and states only for those labels which have not been associated to an equally labeled HMM state.

The final HMM structure inference procedure can thus be resumed as follows:

- *Build an initial graph from $SS_0(n)$ and join to each element of $SS_0(n)$ the relevant HMM state number*
- *For all the sequences $SS_k(n)$ $k = 1, ..., N_S$*
  - *For all the previous $SS_h(m)$ $m = 1, ..., k-1$*
    - *Best align $SS_k(n)$ and $SS_h(m)$*
    - *If the score is minimum retain the warping path*
  - *Add to the best scored $SS_h(n)$ HMM path the states related to $SS_k(n)$ labels not coincident with existing states*
  - *Join to the $SS_k(n)$ labels the state number to which they have been associated*

In this case the DP algorithm is based on alignment constraints chosen in such a way that insertion or deletion of $SS_k(n)$ are forced if this allows to correctly align identical labels. The local score function will thus be defined as

$$D_S(n, m) = d(SS_k(n), SS_h(m)) + \min\{ D_S(n-1, m-1), D_S(n-1, m) + C_{Si}, D_S(n, m-1) + C_{So} \}$$

where $d(SS_k(n), SS_h(m))$ has the value $2 + \varepsilon$, $\varepsilon > 0$, if $SS_k(n) \neq SS_h(m)$ and 0 otherwise, while $C_{Si} = C_{So} = 1$, so that an insertion plus a deletion is advantageous with respect to a substitution.

The procedure above exposed can be halted when a total number of states for the linguistic unit is reached, as well as when the $SS_k(n)$ likelihood falls below a certain threshold.

5. Conclusions

It has been presented a method for the inference of HMM structure which is based on a three stage DP alignment procedure. The input data are constituted by the collection of acoustical label sequences found in actual speech for the linguistic unit to be estimated; the actual experiments are based on an Ergodic HMM based labeling technique.
The first stage of DP has the scope of warping all the sequences to the same length, accounting for label spectral distance and warping dynamic constraints. The second stage deals with the pruning of the trellis which results from the first DP, getting a list of acoustical label sequences ordered in decreasing values of probability. The third and final step is the one on which the final HMM structure is inferred.

The third pass result to be very similar to the Error Correcting Inference Algorithm exposed in [16]. In our case, the pruning of the input data is done before the inference process and not later; moreover, the utilization of an EHMM for the acoustical labeling stage is fundamental in view of the allowable reliability improvements for the inferred model parameter reestimation process to be done before of HMM experimentation for recognition purposes.
The very early stages of development of the exposed method no not permits to dispose, up to to now, of experimental results and assessment about the quality of the inferred structure models. The authors believe to be able of giving further details about the ongoing tests at the conference.

References

[1] - H.Ney, D.Mergel, A.Noll, A.Paeseler: "A data driven organization of the dinamic programming beam search for continuous speech recognition", Proc of ICASSP 1987, Dallas, Texas

[2] - B.Lowerre, R.Reddy, "The harpy speech understanding system", in W.A.Lea, Ed.: Trends in Speech Recognition, Prentice Hall, N.J., 1980

[3] - L.Fissore, P.Laface, G.Micca, R.Pieraccini, "Lexical access to large vocabularies for speech recognition", IEEE Trans on ASSP-37, N.8, Aug 1989

[4] - P.D'Orta, M.Ferretti, S.Scarci, "Phoneme classification for real-time speech recognition of italian", Proc. of ICASSP 1997

[5] - R.Billi, G.Massia, F.Nesti, "Word preselection for large vocabulary speech recognition", Proc. ICASSP 1986

[6] - H.Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition", IEEE Trans. ASSP-32, N.2, April 1984

[7] - B.H.Juang, L.R.Rabiner, "Mixture autoregressive hidden markov models for speech signals", IEEE Trans, on ASSP-33, N.6, Dec 1985

[8] - S.E.Levinson, "Continuously Variable Duration Hidden Markov Models for Speech Analysis", Proc ICASSP-86, p. 1241, Tokyo

[9] - L.R.Rabiner, J.G.Wilpon, F.K.Soong, "High performance connected digit recognition using hidden Markov model", Proc of ICASSP 1988

[10] - H.Hermansky, K.Tsuga, H.Wakita, "Perceptually based processing in automatic speech recognition", Proc of ICASSP 86, Tokyo

[11] - G.R.Doddington, "Phonetically sensitive discriminants for improved speech recognition", Proc of ICASSP 1989

[12] - L.R.Rabiner,C.H.Lee, B.H.Juang, J.G.Wilpon, "HMM clustering for connected word recognition", Proc ICASSP1989

[13] - F.Casacuberta, E.Vidal, H.Rulot, B.Mas, "Use of the grammatical inference algorithm ECGI for finding the topology of HMM", Proc of the Int.Conf. on Speech Tech. VERBA 90, Jan 1990, Rome, ITALY

[14] - Y.Kamp, "State reduction in Hidden Markov models used for speech recognition", Trans IEEE on ASSP-33, n.4, Oct.1985

[15] - R.P.Lippmann, E.A.Martin, "Discriminant Clustering Using an HMM isolated word recognizer", Proc. of ICASSP 1988

[16] - H.Rulot, N.Prieto, E.Vidal, "Learning accurate finite-state structural models of words through the ECGI algorithm", Proc of ICASSP 1989

[17] - A.Falaschi, M.Giustiniani, P.Pierucci, "A Finite states Markov quantizer for speech coding", Proc of ICASSP 1990

[18] - P.Pierucci, A.Falaschi, "Ergodic HMM for synthesis", this volume

[19] A.Falaschi, "Automatic selection of phonologically compact phrases", Proc of the 1st conference of the French Ac.Soc, SFA, April 1990, Lyon, France

[20] - R.M.Gray, A.Buzo, A.H.Gray, Y.Matsuyama, "Distortion measures for speech processing", IEEE Trans. on ASSP-28, N.4

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Adapting a Large Vocabulary Speech Recognition System to Different Tasks

P. Alto, M. Brandetti, M. Ferretti, G. Maltese, F. Mancini, A. Mazza, S. Scarci, G. Vitillaro

IBM Italy Rome Scientific Center via Giorgione 159, 00147 ROME (Italy)

The probabilistic approach to speech recognition has allowed the development of large-vocabulary, high-performances, real-time speech recognizers. Following this approach a speech recognition prototype for the Italian language has been built at the IBM Italy Rome Scientific Center. Many laboratory tests have shown the effectiveness of the prototype as a tool to create texts by voice. To assess the behavior of the recognizer in real environments it is necessary to adapt the vocabulary of the recognizer to different applications. In this paper we present the techniques needed to adapt the acoustic model and the language model, the results obtained for two different applications are also reported.

1. INTRODUCTION

In the last years the probabilistic approach to speech recognition has allowed the development of high-performances large-vocabulary speech recognition systems. At the IBM Rome Scientific Center a speech-recognition prototype for the Italian language, based on this approach, has been built. The prototype is able to recognize in real time natural-language sentences built using a vocabulary containing up to 20000 words. [1]. Once and for all the user has to perform an **acoustic training phase** (about 20 minutes long), during which he is required to utter a predefined text. Words must be uttered inserting small pauses (a few centiseconds), between them. The prototype architecture is based on a personal computer equipped with special hardware. The first system we developed was aimed at a business and finance lexicon. In the following we will refer to it as **EF**. This system was used to perform in-house experiments to assess the acceptance of the recognizer as a tool to create texts. These experiments showed the effectiveness of the prototype [2]. After this phase the necessity arose to perform experiments in real work environments. Two different applications were considered: the dictation of radiological reports and of insurance company documents. They will be indicated as **RR** and **IR** respectively. Due to their characteristics, these applications seemed to be very well suited for our purposes. To develop the systems to be employed during the experiments, we had to adapt the **EF** recognizer to the lexicon required by the new applications. In the probabilistic approach the vocabulary of the recognizer is predefined and no efficient way to adapt the vocabulary of the system exists. The paper describes the techniques we have used to solve the problem of vocabulary adaptation. The results obtained experimenting automatic text dictation during real work are also presented.

2. SYSTEM STRUCTURE

We look for the sequence of words $\hat{W}$ which has the highest probability given the acoustic information $\overline{A}$ extracted from the observed signal [3]. In our case the acoustic signal is a sequence of **acoustic labels** extracted from the signal every centisecond and representing the energy content of the signal in 20 frequency bands.
Applying the Bayes theorem we can write:

$$P(\overline{W}\,|\,\overline{A}) = \frac{P(\overline{A}\,|\,\overline{W})P(\overline{W})}{P(\overline{A})} \qquad (1)$$

where $P(\overline{W}\,|\,\overline{A})$ is the probability that the sequence of words $\overline{W}$ will produce the sequence of acoustic information $\overline{A}$. $P(\overline{W})$ is the *a priori* probability of the sequence of words $\overline{W}$. $P(\overline{A})$ is the probability of the sequence of acoustic information $\overline{A}$. We want to find the maximum of the above expression with respect to $\overline{W}$. We can ignore $P(\overline{A})$ because it does not depend on $\overline{W}$. Therefore we need to maximize the numerator of the expression (1).
The problem can be reduced to the following steps:

1. perform the **signal processing** stage to extract the acoustic information $\overline{A}$ from the speech signal;
2. compute the acoustic probability $P(\overline{A}\,|\,\overline{W})$ (this task is accomplished by the **acoustic model**);
3. compute $P(\overline{W})$ (this is done by the **language model**);
4. look for the most probable sequence of word through an efficient **search strategy**.

While the signal processing stage and the search strategy can be considered independent of the application, the acoustic and the language model must be changed according to the lexicon of the application. In the next paragraphs the techniques employed to adapt both models will be explained.

3. ACOUSTIC MODEL ADAPTATION

The acoustic model task is to compute $P(\overline{A}\,|\,\overline{W})$. In the probabilistic approach the acoustic model is based on **hidden Markov models**. An hidden Markov model is a finite state automata. For every time slice the model takes a transition from the current state to one of the allowed states (the transition can also produce no state changement). For each transition an acoustic label is produced [3]. Both the transitions and the label emission occur according two probability distributions. The distributions depend on the current state only. These models are called *hidden* because it is only possible to observe the sequence of acoustic symbols produced, while the sequence of states remains hidden.
Each word belonging to the vocabulary is represented by a different model.

Two different techniques exist to construct the models. The first one is based on the idea of automatically building the word model starting from several utterances of it produced by several speakers [4]. According to the second technique an alphabet of acoustic units to represent the basic sounds of the language is defined. The word model is built by concatenating the Markov models representing the acoustic units. In our case the latter technique was employed. Examples of acoustic units used for speech recognition are: syllables, diphones, phones. We choose the phone as phonetic unit. The basic sounds of the Italian language were described by a set of 56 phonetic units [1]. For each phonetic unit a Markov model representing its pronunciation has been created. In our system all the phonetic units have the same topological structure. The distinction between different sounds is left entirely to the probability distributions, called *model's parameters*. The computation of the parameters is accomplished during the acoustic training phase employing the predefined text uttered by the user.

According to the technique chosen, the first step that must be performed when adapting the recognizer to a new application, is to make the phonetic transcription of all the words in the vocabulary. To limit the number of the needed phonetic transcriptions, a database was built containing all the words and the phonetic transcriptions used in previous vocabularies. By using the database it is possible to find all the words for which a new phonetic transcription must be supplied.

Usually, the phonetic transcription is a performed manually. It is a very expensive process and for large vocabularies the transcriptions could contain errors. We tried to make the phonetic transcription process as automatic as possible. The systems that have been proposed to solve the problem of automatic phonetic transcription are based on rules [5] [6] or on automatic learning from training data [7]. Actually, these systems cannot provide the accuracy required for automatic speech recognition. This is due both to the complexity of the problem and to the difficulty to describe all the possibilities with a limited set of rules. We employed a different technique from the mentioned ones. We separated phonotactical knowledge (well described by a limited set of rules) from lexical knowledge (based on experience and not suitable for a formal description). Given the string representing the orthographic form of the word our system produces a set of phonetic transcriptions for that word, which are the ones that can be obtained applying our set of rules for the grapheme-to-phoneme translation. The user can choose manually the correct transcription on the basis of his lexical knowledge. Grapheme-to-phoneme translation for the Italian language has a relatively low uncertainty. A set of 78 rules allows to describe all the ambiguities. Each rule consists of a left part and a right part. The left part consists of a grapheme string and its (possibly empty) left and right graphemic contexts; the right part consists of the set of possible phonetic transcriptions for the grapheme string. The set of transcriptions produced applying this set of rules is then pruned by means of a set of global rules (which, for example, reject all the transcriptions which do not have one and just one stressed vowel). The right phonetic transcription always belongs to the resulting set. The average number of phonetic transcriptions per word is 5. Using this method it was possible to adapt rapidly the recognizer to the new application. The quality of the produced transcriptions was at least equal to a completely manual phonetic transcription.

4. IN-HOUSE TEST FOR THE NEW APPLICATION

Before experimenting the speech-recognizer in a real environment we needed to perform in-laboratory tests to assess the recognition rate of the system when used to dictate pre-defined texts. To perform the experiment a text containing phrases peculiar to the application must be created; it must be dictated by several different speakers. To make a meaningful test it is important that the text contains all the phonetic units used to build the acoustic model in phonetic contexts typical of the application.

Usually, the text is built manually trying to represent a large number of different contexts using the smallest number of sentences. To avoid this manual process a procedure has been built to prepare the test automatically. A set of sentences peculiar to the application is used as initial data. Usually the corpus employed to train the language model is used for this purpose. A preliminary analysis is done to eliminate all the sentences that contain words not included in the vocabulary. The first selected sentence is the one containing the greatest number of distinct phones. The sentences are then added incrementally, and at each step the sentence with the highest score among the selected ones, is chosen. The score is computed in the following way:

- the frequency of each phonetic unit in the previously selected sentences is computed;
- for each sentence in the available data and not yet selected a score is computed according to the following formula:

$$C(S_k) = \sum_i f_i \ \exp\ (-h_i) \qquad (2)$$

  where f_i is the frequency of the phone i in the sentence S_k while h_i is the frequency of phone i in the sentences selected so far;
- the summation is extended only to the phones the frequency of which is less than a predefined threshold.

By applying this algorithm it is possible to select efficiently a set of sentences containing phonetic contexts typical for the application and suitable to assess the accuracy of the recognizer.

5. VOCABULARY ADAPTATION

In our system the vocabulary is predefined; this means that one of the most important factors affecting the usability of the speech recognizer is the availability of the largest number of words needed by the user to create the text. The choice of the vocabulary is a key factor for the system performances.

In our first experiment a vocabulary containing more than 20000 words was used. This vocabulary (EF) is aimed at the dictation of economy and finance reports. The 20000 words were chosen as the most frequent in a corpus containing 44 millions of words composed by: articles from the most important Italian economy and finance newspaper (*Il Sole 24 Ore*), articles from an economy and finance newsmagazine (*Il Mondo*), and press agency news. The coverage of this vocabulary computed on a disjoint corpus was 96.5%

The economy and finance lexicon is very different from the lexicon required to dictate radiological reports, while it has

some similarities with respect to the lexicon used in the insurance company reports. In the first case (**RR**) the vocabulary was selected by using a corpus containing only radiological reports; in the second case (**IR**) the vocabulary was built taking **EF** as a starting point.

Radiological Reports Vocabulary

The available corpus contained about 5 million words (we will call this corpus *HO*) collected in four different hospitals (*HO*1, *HO*2 , *HO*3, *HO*4). The hospital were the experiment was held provided us with a corpus containing only 50000 words (*HE*). The first problem was the choice of the vocabulary size. We adopted the criterion of analyzing the variation of the coverage with respect to the vocabulary size (figure 1).

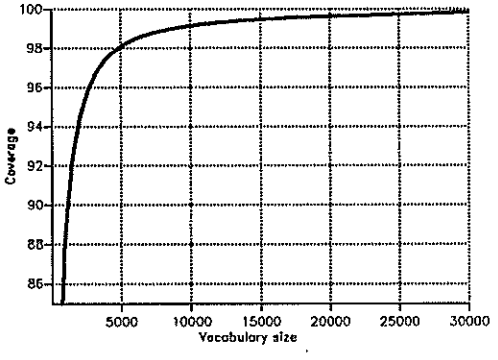


Figure 1. Coverage of corpus HO as function of vocabulary size.

A vocabulary containing 5000 words seemed to us a reasonable trade-off between the need to reach high coverage and the need to have enough data to estimate the language model parameters.

One of the main characteristics found in this kind of lexicon was the presence of a set of words peculiar to the report dictation process at each location. To make the recognizer well suited to the needs of the experimenter all the 3200 different words found in corpus *HE* were included in the vocabulary. The vocabulary was completed adding 1900 words which were the most frequent ones in *HO* corpus not included in the previous list of 3200 words. The *HO* words were ordered according to the average frequency of each word in the various corpora HO_i:

$$\bar{f} = \frac{1}{4} \times \sum_{i=1}^{4} \frac{C_{HOi}(w)}{N_{HOi}} \qquad (3)$$

$C_{HOi}(w)$ is the number of word w occurrences in corpus HO_i while N_{HOi} is the size of corpus HO_i). The number of occurrences was normalized because the four corpora HO_i have different sizes. The resulting vocabulary had a 100% coverage on reports from *HE* and 97.5% coverage on *HO* reports.

Insurance Company Reports Vocabulary

The selection of the words to be used in the **IR** vocabulary was performed by analyzing a corpus (*IC*) containing 1.5 million words provided by the insurance company. There was a certain similarity between the **EF** and **IR** lexicons: the coverage obtained by **EF** vocabulary on *IC* corpus was 95.2%. The 15000 most frequent words in **EF** were selected and the 3100 most frequent words found in *IC* not contained in the previous selected words were added. The resulting vocabulary **IR** showed a 99% coverage on *IC* texts and a 95.7% coverage on economy and finance texts.

6. LANGUAGE MODEL CONSTRUCTION

The task of language model is to compute $P(\overline{W})$. The computation is performed in the following way:

$$P(\overline{W}) = \prod_{i=1}^{N} P(w_i \mid w_{i-1}, w_{i-2}) \qquad (4)$$

This is the so-called **trigram language model**[8] : it contains the approximation of considering equivalent all the sentences ending with the same couple of words $w_{i-1}w_{i-2}$. The number of possible trigrams is so large that is practically impossible to collect the amount of data needed to estimate the probability of each of them. To overcome the problem an interpolation between different probability distributions is performed. Three different distributions are computed for trigrams, bigrams and unigrams. The probability of word w_3 given the words w_1 and w_2 is estimated as follows:

$$P(w_3 \mid w_1 w_2) = \lambda_3 \frac{C(w_1 w_2 w_3)}{C(w_1 w_2)} + \lambda_2 \frac{C(w_2 w_3)}{C(w_2)} + \lambda_1 \frac{C(w_3)}{N} + \lambda_0 \frac{1}{V}$$

where $C(w_1, \ldots, w_n)$ is the number of times the word string $w_1, \ldots, w_n$ was observed in the training data and V is the number of words in the vocabulary. The λ coefficients are estimated using the *Expectation-maximization* algorithm [9]. The trigram language model is an effective tool to represent the linguistic constraints for speech recognition purposes; on the other hand it requires a large amount of training data. The larger is the training corpus, the higher is the recognition rate [10]. The models for the two vocabularies **RR** and **IR** were built following the previously described technique. The **RR**language model was trained using 4.8 million words. In the **IR** case a corpus containing 1.5 million words typical of the application was merged with 40 million words extracted from the **EF** corpus. In the following table the most significant parameters of the language models are reported.

Table 1. Language models characteristics		
Parameter	**RR**	**IR**
Vocabulary size	5100	18100
Training data	4.8	1.5 + 40
Millions of different trigrams	0.62	3.4
Millions of different bigrams	0.19	2.3
Perplexity	38	18

Perplexity is a typical measure of the predictive power of a language model [11]. It estimates the average number of words that are considered equiprobable by the model.

7. RESULTS

In the following paragraph the results obtained experimenting the two prototypes during real work are reported.

Radiological Reports Dictation

Four doctors have used the recognizer during their every-day work to prepare the reports to be delivered to the patients. No one had any difficulty in inserting short pauses between words. The doctors dictated 150 reports containing more than 12000 words. The vocabulary coverage for the dictated reports was 98.6%. Table 2 reports the results of the experiment.

Table 2. Radiological reports dictation			
Speaker	**Number of reports**	**Error rate**	**Speaker's errors**
sp1	25	1.7%	1.4%
sp2	14	2.0%	1.2%
sp3	76	3.5%	0.9%
sp4	35	5.0%	3.2%

The error rate is referred to the number of errors done by the recognizer without taking into account errors due to words not included in the vocabulary. The speaker's error rate is referred to the errors due to misuse of the recognizer (wrong commands, absence of pause between words, etc.). The global error rate can be computed summing the numbers in the third and fourth column of the table. We can see that the recognizer's performances ranges from 90.5% to 95.6% of accuracy.

Insurance Company Reports Dictation

The experimentation was carried on by five different users who dictated more than 8000 words. The vocabulary coverage on the dictated text was about 99%. Table 2 reports the results obtained in this case.

Table 3. Insurance company reports dictation		
Speaker	**Error rate**	**Speaker's error rate**
sp1	1.9%	1.0%
sp2	1.4%	0.5%
sp3	14.0%	2.0%
sp4	7.4%	1.5%
sp5	2.4%	0.8%

REFERENCES

[1] P. D'Orta, M. Ferretti, A. Martelli, S. Melecrinis, S. Scarci, G. Volpi, **Large-Vocabulary Speech Recognition: a System for the Italian Language,** *IBM Journal of Research and Development,* Vol. 32, No. 2, March 1988, pp.217-226.

[2] P. Alto, M. Brandetti, M. Ferretti, G. Maltese, S. Scarci, **Experimenting Natural-Language Dictation with a 20000-Word Speech Recognizer,** *IEEE CompEuro 89,* Hamburg, May 8-12, 1989, pp. 2-78 - 2-81.

[3] L.R. Bahl, F. Jelinek, R.L. Mercer, **A Maximum Likelihood Approach to Continuous Speech Recognition,** *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. PAMI-5, no. 2, 1983, pp. 179-190.

[4] L.R. Bahl, P.F. Brown, P.V. De Souza, R.L. Mercer, M.A. Picheny, **Acoustic Markov Models Used in the Tangora Speech Recognition System,** *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing,*

[5] R. Carlson, B. Granstroem, **A Text-to-Speech System Based Entirely on Rules,** *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing,* Philadelphia, PA, April 1976.

[6] D. H. Klatt, **Structure of a Phonological Rule Component for Synthesis-by-Rule Program** *IEEE Trans. on Acoust., Speech and Sig. Proc.,* vol. ASSP-24, no. 5, 1976, pp. 391-398.

[7] T. J. Sejnowski, C. R. Rosenberg, **Parallel Networks that Learn to Pronounce English Text,** *Complex Systems,* 1 (1987), pp. 145-168.

[8] F. Jelinek, **The development of an experimental discrete dictation recognizer,** *Proceedings IEEE,* vol. 73, no. 11, November 1985, pp. 1616-1624.

[9] F. Jelinek, R.L. Mercer, **Interpolated Estimation of Markov Source Parameters from Sparse Data,** *in "Pattern Recognition in Practice", E.L.Gelsema and L.N. Kanal, Ed.,* North-Holland, New York, 1980, pp. 381-387.

[10] M. Ferretti, G. Maltese, S. Scarci, **Measures of Language Model and Acoustic Information in Probabilistic Speech Recognition,** *Eurospeech 89,* Paris, September 1989, pp. 473-476.

[11] F. Jelinek, R.L. Mercer, L.R. Bahl, J.K. Baker, **Perplexity - a Measure of Difficulty of Speech Recognition Tasks,** *94th Meeting Acoustical Society of America,* Miami Beach, December 1977.

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

REJECTION TECHNIQUES IN CONTINUOUS SPEECH RECOGNITION USING HIDDEN MARKOV MODELS

Pedro J. Moreno*, David B. Roe, and Padma Ramesh

AT&T Bell Laboratories
Speech Research Department
Murray Hill, New Jersey 07974

ABSTRACT

In this paper we deal with the concept of rejection, that is, of electing not to force a decision if there is considerable uncertainty which of several Markov models best matches the speech. We present some recent studies in rejection techniques for connected speech recognition, based on HMMs (hidden Markov models). Several techniques have been studied, with the objective of rejecting as many incorrectly recognized strings as possible while eliminating relatively few correctly recognized strings. The result of this study is that the proposed technique eliminates fully half of the speech recognition errors, while rejecting only 5% of the spoken utterances. Finally, an application is described showing the algorithms proposed here can also be used for rejecting out-of-vocabulary words, or out-of-grammar sentences.

1. Introduction

The most important measure of a speech recognition system is its error rate. Most research efforts are in trying to reduce this variable. But even in the most sophisticated speech recognition system, it is not possible to correctly recognize *all* the speech because the speech signal is inherently variable. (For instance, a low signal to noise ratio, a slip of the tongue, or an unusual accent can cause misrecognitions). In some applications, the resulting errors would be disastrous. In these cases, it is better to reject a doubtful utterance. For instance, telephone-based inquiry systems may request the user to repeat the information.

Our goal is to develop a theoretical framework for the rejection techniques, and describe the various schemes that we have used. Several successful schemes have been proposed for speech recognizers based on template matching, but a corresponding method for Hidden Markov Models is not as straight forward.

The paper is organized as follows. In section 2, we review the basic structure of the recognition system, the algorithms and the database chosen for our study. In section 3 and 4 we describe the proposed rejection algorithms and give the experimental results. Finally in section 5 we give some results of a real application using the proposed techniques.

2. Speech Recognition System

In order to get more valuable results we decided to use a standard speaker-independent database, the Texas Instruments database [1] consisting of 8565 sequences for training and 8578 distinct sequences for recognition. This database contains digit strings, from zero to nine, with the alternative pronunciation "oh" for the digit zero. There are 225 different speakers, both male and female.

In order to get statistically independent results we have divided our database in three different sets:

- A training database with 8565 strings used for the generation of the HMM models of the digits.
- A threshold database with 5065 sentences used for the generation of the threshold parameters values (see next section).
- A testing database with 3513 sentences generated for a complete evaluation of the rejection algorithms.

For these experiments we have used the Bell Laboratories standard speech recognition algorithms [2] developed by Rabiner and others. These algorithms use the connected-word, continuous density, multiple Gaussian mixture, HMM technology [3]. Basically the recognizer has three main parts:

1) Feature analysis: the speech signal is converted to LPC cepstrum and LPC delta cepstrum, plus energy.

2) Level building pattern matching: the input sequence of feature vectors is matched with the set of Hidden Markov Models, getting a list of candidate strings (those with highest string probabilities).

3) Postprocessing: the most likely string is obtained from the list of candidates using duration information.

In our experiments we have used one HMM per digit, 8 states per model, and 9 continuous densities (mixtures) per state.

* TELEFONICA I+D
Emilio Vargas 6
Madrid 28043, SPAIN

3. Rejection Techniques

The key to our rejection techniques is to use auxiliary information from the recognition process, in addition to the most likely word sequence.

Examples of this auxiliary information include:

1) the duration of the Viterbi path in any of the states of the HMM.
2) the sentence likelihood of the best candidate sequence.
3) the word likelihoods in the best candidate.
4) the difference in likelihood between the two best candidate sentences.

In each of the proposed techniques a fixed threshold, or reference value, is determined for one or more of these auxiliary parameters. If a parameter exceeds the threshold, the sentence is accepted and a recognition decision is made. If the parameter does not reach the threshold, the sentence is rejected and no decision is made.

Below we discuss 4 possible techniques for getting thresholds for rejection.

3.1 *Rejection Technique 1: State Duration*

In our recognition algorithms a state duration penalty is built in, such that during the recognition process a high penalty is added to the score of those states which have long durations. It has been observed that incorrectly recognized strings usually have states with long duration, as shown by the second example in Figure 1. In the first example a correctly recognized string is shown. There is an even distribution of the durations over all the states. On the other hand in the second (incorrect) example some states have very long durations, particularly the second word in the sequence.

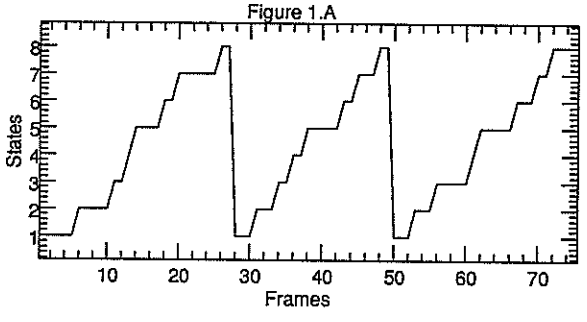


Figure 1.A: HMM state-duration times for a Correctly-recognized 3 digit sequence (1 3 3). The duration (in frames) for each of the states is relatively even.

In this rejection technique we look at the state with the longest duration. In order to evaluate if a state in a word has an unduly long duration, t_v, it has to be compared with an average duration. We compare this longest duration time, t_v, with a local average duration over the word, μ_t, defined as the average number of frames per state in that particular word in that particular string. We choose the state with the highest ratio t_v to μ_t as the "worst" state in the string.

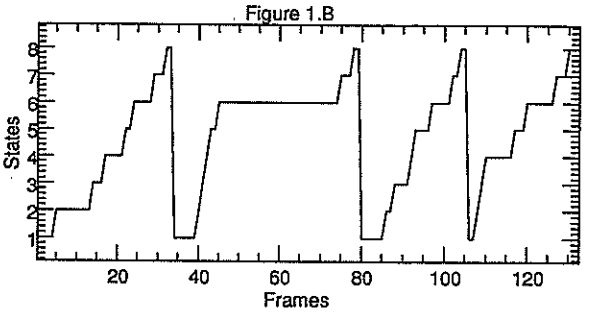


Figure 1.B: State durations for an incorrectly recognized sentence (7 0 3 0). The second word spend too long in state 6 (30 frames).

Our rejection criterion consists of rejecting those strings in which the ratio t_v/μ_t is larger than a certain value. However, to the extend that correct sequences also have large ratios reflecting large durations in one state, they also will be rejected.

3.2 *Rejection Technique 2: Sentence Likelihood*

The overall sentence likelihood is another intrinsic property of the spoken phrase that can be used for rejection. Here we hypothesize that low scores are correlated with incorrectly recognized strings.

For every string, its normalized score ($score_{string}/number_{frames}$) is compared with a threshold, ψ_{string}. The normalized score is used to eliminate any dependence on the length of the spoken utterance. If the normalized likelihood is greater than this threshold, ψ_{string}, the string is accepted. Otherwise it is rejected.

For every value of the threshold ψ_{string}, an error rate

$$E_{string} = f_{string}(\psi_{string}) \quad (1)$$

is obtained, and similarly a rejection rate

$$R_{string} = g_{string}(\psi_{string}) \quad (2)$$

is calculated.

3.3 *Rejection Technique 3: Worst Word Likelihood*

The "worst word likelihood" criterion is based in the hypothesis that in a connected string of words, if the likelihood of one word is very low, the string is incorrect.

For every candidate the score of the worst word, S_W is defined. This score is normalized by the number of frames in the sequence of observations O_W (duration of that word). In order to reject a misrecognized sentence, we compare this score, S_W, with a threshold ψ_W. If S_W is lower than ψ_W the string is rejected, otherwise the string is accepted.

3.4 *Rejection Technique 4: Next-Best Likelihood*

During the recognition process, for every spoken sentence a list of possible strings is obtained, sorted by the string likelihood.

A "distance" (or difference in likelihood) can be defined between the best two sequences of models (or string candidates) λ_1 and λ_2 for a spoken utterance O as:

$$D(\lambda_1,\lambda_2) = \log P(O \mid \lambda_1) - \log P(O \mid \lambda_2) \quad (3)$$

with

$$O = \{O_1 O_2 \cdots O_t\} \quad (4)$$

the sequence of observations characterizing the spoken utterance. This distance is related to the log of the relative probabilities of the top two candidates. The more similar the probabilities, the greater the chance of confusion and the smaller the distance.

To reject a misrecognized sentence, we compare the distance with a threshold ψ_λ. If the difference in likelihood is lower than ψ_λ, the hypothesis is rejected.

An error/rejection plot is generated, for a given threshold ψ_λ an error rate and rejection rate are obtained

$$E_\lambda = f_\lambda(\psi_\lambda) \quad (5)$$

$$R_\lambda = g_\lambda(\psi_\lambda) \quad (6)$$

3.5 *A Rejection Example*

In this section, we give two examples of how the proposed algorithms work, first on a correctly recognized sequence, then on an incorrectly recognized string.

recognized string: *one three three* (correct)

String likelihood: 11.58

Second candidate likelihood: 9.08

String likelihood 11.58			
word	duration	likelihood	$\frac{t_v}{\mu_t}$
1	27	9.52	1.78
3	22	14.53	1.82
3	26	11.21	1.54

Table 1. Score values for a correctly recognized string.

All the thresholds are fixed in order to achieve a 5% rejection rate.

Rejection criterion 1:

$$(\frac{t_v}{\mu_t} = 1.82) \leq 3.2$$

The largest ratio t_v/μ_t in the string is 1.82, and this value is compared with the threshold 3.2. Since this ratio is less than the threshold, the string is accepted.

Rejection criterion 2:

$$(score_{string} = 11.58) \geq (\psi_{string} = 7.42)$$

The score of the string is 11.58, much higher than the threshold limit for rejection. Following this criterion the string is accepted.

Rejection criterion 3:

$$(S_W = 12.1) \geq (\psi_W = 3.32)$$

The worst word likelihood in the string is 9.52. The threshold limit for this criterion is set to 3.32, hence the string is accepted.

Rejection criterion 4:

$$D(\lambda_1,\lambda_2) = (11.58 - 9.08 = 2.50) \geq \psi_\lambda = 0.21$$

The distance between the first candidate string and the second is 2.5. On the other hand the threshold used with this criterion in order to have a 5% rejection rate is 0.21. Hence the string is accepted.

In the next example an incorrectly recognized string is presented. Some of the rejection thresholds reject properly, while others fail to reject.

recognized string: *seven oh three oh* (*incorrect*)

correct string: *seven oh oh three oh*

String score: 9.10

Second candidate score: 9.09

String likelihood 9.10			
word	duration	likelihood	$\frac{t_v}{\mu_t}$
7	33	10.35	2.18
0	46	6.27	5.22
3	26	10.84	1.85
0	27	10.72	2.07

Table 2. Score values for an incorrectly recognized string.

As in the previous example the limit thresholds are set for a 5% rejection rate.

Rejection criterion 1:

$$(\frac{t_v}{\mu_t} = 5.22) \geq 3.2$$

The largest ratio t_v/μ_t in the string is 5.22, much higher than the threshold limit (3.2), hence the string is rejected.

Rejection criterion 2:

$$(score_{string} = 9.10) \geq (\psi_{string} = 7.42)$$

The score of the string is 9.10, higher than the threshold ψ_{string}, so this criterion fails to reject the sentence.

Rejection criterion 3:

$$(S_W = 6.27) \geq (\psi_W = 3.32)$$

The worst word score in this string is 6.27, a larger value than the threshold limit ψ_W (3.32). As with technique 2 fails to reject the incorrect string.

Rejection criterion 4:

$$D(\lambda_1,\lambda_2) = (9.10 - 9.09 = 0.01) \leq \psi_\lambda = 0.21$$

The difference between the first candidate and the second one is 0.01, much less than the threshold limit ψ_λ (0.21), so the fourth rejection technique rejects the error successfully.

4. Results

The procedure for testing the thresholds is as follows: First we train Markov models for each word using the training database. Second, we establish for each of the rejection techniques the appropriate threshold to achieve a given error rate using the threshold database. Third, we run the recognition analysis on the test database using those computed thresholds.

Figure 2 is a comparison among these four different techniques. The vertical axis shows the percent error rate in the recognition test. The horizontal axis shows the percent rejection rate obtained by varying the threshold from zero (no rejection at all, high error rate) to high (many rejections, but virtually no errors.) The ideal behavior is a curve descending steeply as the rejection rate rises. The top left portions of the curves correspond to very small thresholds for which almost all sentences are accepted, while the lower right portions of the curves correspond to high thresholds which eliminate nearly all the errors, but at the expense of a high rejection rate.

It is clear that the best of these techniques is the Next-Best Likelihood criterion. While rejecting only 5% of the spoken utterances, it reduces the error rate by half, going down from 4.6% to 2.1%. In Table 3 the error rates are given for the four different techniques with no rejection, a rejection rate of five percent and a rejection rate of ten percent.

Surprisingly, techniques 1, 2, and 3 don't produce as good results. While it is true that many of the strings that are errors have, say, unusually long state durations, it seems to be the case that many correct strings also have long state durations.

Technique Name	Percent Error Rate		
	0% Rej.	5% Rej.	10% Rej.
1 State Duration	4.6	3.3	2.7
2 Total Sentence Likelihood	4.6	3.9	3.3
3 Worst Word Likelihood	4.6	3.7	2.8
4 Next-Best Distance	4.6	2.3	0.9

Table 3. Comparison among different rejection techniques

5. Out-of-Vocabulary Rejection

Rejection techniques 2 and 4 have also been successfully used in a somewhat different test involving a ninety word, speaker-trained, connected speech recognition task.

The objective in this task is to reject speech utterances containing words not in the vocabulary and to reject sentences

Figure 2

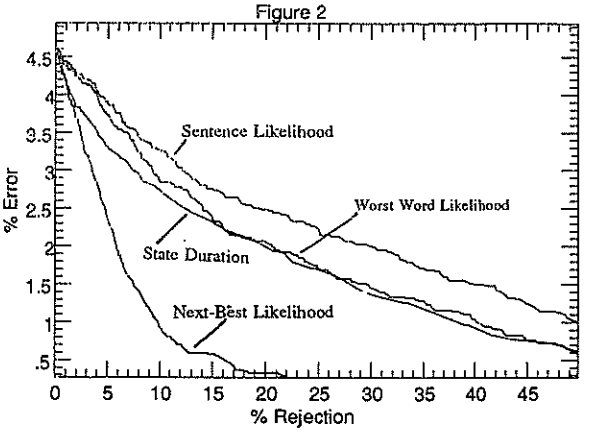


Figure 2: Error/Rejection plot for four different rejection criterions

which, though consisting of proper words, failed to follow the prescribed syntax of the recognition application. This is a considerably easier task than that of identifying a potential misrecognition. Intuitively one would expect that out-of-vocabulary words would have much lower likelihoods than words in the vocabulary.

There were a total of 170 sentences from two speakers in this informal experiment. After training the HMMs, a set of thresholds was fixed using the training database. A combination of techniques 2 and 4 is able to reject *all* 20 of the invalid sentences in the database, while rejecting *none* of the 150 correct sentences. The key ingredient in the grammar is that all sentences begin with a tag phrase like "room controller" or "network controller". The chances of a random utterance matching a sentence beginning with one of these tag phrases is quite small.

6. Conclusion

New rejection techniques for hidden Markov model speech recognition have been proposed. For speaker independent recognition systems, the proposed algorithms work quite well, reducing the error rate by a factor of two (to 2.3 percent) while rejecting only 5% of the spoken sentences. These techniques also are successful at rejecting out-of-vocabulary speech.

References

[1] R. G. Leonard, "A database for Speaker-Independent Digit Recognition", Proc. 1984 ICASSP, vol. 3, 42.11, March 1984

[2] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High Performance Connected Digit Recognition Using Hidden Markov Models", Proc. Int. Conf. ASSP, pp. 119-122, 1988.

[3] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. IEEE, Vol 77, pp. 257-286, Feb 1989

[4] L. R. Rabiner, M. M. Sondhi, and S. E. Levinson, "A Vector Quantizer Incorporating Both LPC Shape and Energy," *Proc. ICASSP 1984*, paper 17.1.1, March 1984

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

PARAMETRIC MODELLING OF STATE TRANSITIONS IN HIDDEN MARKOV MODEL

Lu Chang and M.M. Bayoumi

Department of Electrical Engineering
Queen's University, Kingston, Canada K7L 3N6

Abstract. In this paper we propose to model the state transitions in hidden Markov model (HMM) with one that is based on a parametric probability distribution, especially the binomial distribution. With this extension, the HMM can be fully parametrically represented. This can increase the method's flexibility and capability. The estimation formula for the parametric transition matrix is derived in the paper.

1. Introduction

Hidden Markov model (HMM) has been successfully applied to automatic speech recognition[1,2]. The complicated structure of the model enables it to model the intricate speech signals. The model can be described as follows. A Markov chain is governed by a stochastic transition matrix $A=\{a_{ij}\}_{NxN}$. We observe the model through a family of random variables described by $B=\{b_{jk}\}_{NxM}$ with one variable corresponding to one state, where N is the model state number and M is the dimension of the random variables. The initial state is specified by $\Pi=\{\pi_i\}$. Hence a HMM can be represented by a triplet $\lambda=\{A,B,\Pi\}$.

A significant problem associated with the application of HMM is the large number of parameters which have to be estimated. For speech recognition, this problem is especially serious due to the essentially incomplete nature of speech signals. In practice, simpler types of HMM are usually used such as, left-right model[3], parallel path left-right model[1], instead of fully connected HMMs. It should be mentioned that the number of states used in the model need to be kept small, typically in the order of 5 to 10. Apparently the assigned structure of HMM will limit the representation capacity of the model.

We propose here a new method to formulate the transition matrix. If the transitions from one state to another states can be modeled by a random variable which has a simple form of probability density, we will be able to specify the transitions with a relatively small number of parameters.

The organization of the paper is such that the theory of the hidden Markov model (HMM) is reviewed in section 2. In section 3, it will be shown that the binomial distribution is extremely suitable to the purpose. We will derive the reestimation procedure for the new model in the section 4. Section 5 includes some concluding remarks.

2. Hidden Markov Model

Given the observation sequence $O_1^T=(O_1,...,O_T)$ and a model λ, which has N states $S=(s_1,...,s_N)$, we

can evaluate the probability of the observations produced by the model (which will be called the generating probability). This is calculated by enumerating all the possible paths:

$$P(O|\lambda) = \sum_{Q_i} P(O,Q_i|\lambda) = \sum_{Q_i} P(O|Q_i,\lambda) \quad (1)$$

where $Q_i=q_1,q_2,...q_T$ is the *ith* state sequence. To efficiently evaluate the generating probability, a forward probability $\alpha_t(i)$ is defined as follows

$$\alpha_t(i) = P(O_1^t, q_t=s_i|\lambda) \quad 1\le t\le T \quad 1\le i\le N \quad (2)$$

A recursive equation can be easily established

$$\alpha_1(i) = \pi_i b_i(O_1) \quad 1\le i,j\le N$$
$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i)a_{ij}b_j(O_{t+1}) \quad 1\le t\le T \quad (3)$$

With the forward probability thus defined, the generating probability can be obtained using eq.(4)

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad (4)$$

In a similar way, we can define a backward probability $\beta_t(j)$ as follows

$$\beta_t(j) = P(O^{T_{t+1}}|q_t=s_j,\lambda) \quad 1\le t\le T \quad 1\le j\le N \quad (5)$$

The corresponding recursive equation is given by

$$\beta_T(i) = 1 \quad 1\le i\le N$$
$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}b_j(O_{t+1})\beta_{t+1}(j) \quad 1\le t\le T-1 \quad (6)$$

The generating probability can also be expressed in term of the backward probability using the following equation.

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(O_1)\beta_1(i) \quad (7)$$

Using these two recursive probabilities, we can get other important probabilities, such as, the state transition occurrence probability

$$\xi_t(i,j) = P(q_t=s_i,q_{t+1}=s_j|O,\lambda) \quad 1\le i,j\le N$$
$$= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta(j)}{P(O|\lambda)} \quad 1\le t<T \quad (8)$$

and the state occupancy probability.

$$P(q_t=s_i|O,\lambda) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} \quad (9)$$

Notice that the above probabilities are functions of the observations. Hence they are useful in the estimation of model parameters if training data is given.

3. Binomial Distribution

We are looking for a distribution which satisfy following conditions:

1) Finite discrete distribution
2) Few specifying parameters
3) Distribution controlled effectively by the parameters

The first condition is for the finite Markov chain which is considered in this paper. The second condition is necessary to increase the efficiency of the reestimation. The third condition is the key that helps to keep the structure of the model flexible.

The binomial distribution seems to be the best one for our purpose. This distribution can be expressed as follows

$$P(x=k) = \binom{N}{k} P^k(1-P)^{N-k} \quad 0\le k\le N \quad (10)$$

where $0<P<1$. Fig.1 shows the shape of $P(x=k)$ versus k for N=6 and when P changes from .1 to .9. We can see that the probability distributions have different shapes.

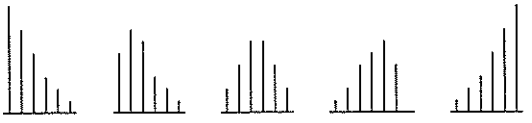


Fig. 1 Binomial Distribution with N=6

Hence when it is used in the HMM, we may be able to achieve a parameter optimization and the structure optimization at the same time. Now the element of transition matrix becomes

$$\begin{aligned} a_{ij} &= P(q_{t+1}=s_j|q_t=s_i) \qquad 1\le i,j\le N \\ &= \binom{N-1}{j-1} P_i^{j-1}(1-P_i)^{(N-j)} \qquad 0\le P_i\le 1 \end{aligned} \tag{11}$$

The number of transition matrix parameters is thus reduced from N^2 to N.

4. Reestimation

Given a finite observation sequence $O=\{O_t,0<t<T\}$, we can compute the probability $P(O|\lambda)$ of the sequence being generated by model $\lambda=(A,B,\pi)$. The reestimation problem aims at estimating the parameters of a new model $\lambda=(A,B,\pi)$. This new model should increase the generating probability, that is, $P(O|\lambda) > P(O|\lambda)$. In the following, the reestimation formula for P_i will be derived in a way which is similar to that of Liporace[4]. Define an auxiliary function $Q(\lambda,\lambda)$ as

$$Q(\lambda,\bar{\lambda}) = \sum_{Q_i} P(O,Q_i|\lambda) \log P(O,Q_i|\bar{\lambda}) \tag{12}$$

Then we have

$$\begin{aligned} Q(\lambda,\bar{\lambda}) - Q(\lambda,\lambda) &= \sum_{Q} P(O,Q_i|\log \frac{P(O,Q_i|\bar{\lambda})}{P(O,Q_i|\bar{\lambda})} \\ &\le \sum_{Q_i} P(O,Q_i|\lambda)\left(\frac{P(O,Q_i|\bar{\lambda})}{P(O,Q_i|\lambda)} - 1\right) \\ &= P(O|\bar{\lambda}) - P(O|\lambda) \end{aligned}$$

Hence if $Q(\lambda,\lambda) > Q(\lambda,\lambda)$, then $P(O|\lambda) > P(O|\lambda)$.

With eqs.(3) (4) and (1), we have

$$\begin{aligned} Q(\lambda,\bar{\lambda}) = \sum_{Q_i} P(O,Q_i|\lambda) \sum_{t=1}^{T} \{ & \log\overline{\pi_{s_t}} \\ & + \log\overline{b_{s_t}}(O_t) + \log\binom{N-1}{s_t-1} + (s_t-1)\log\overline{P_{t-1}} \\ & + (N-s_t)\log(1-\overline{P_{t-1}}) \} \end{aligned} \tag{13}$$

Here we only concern ourselves with the estimation of P_i. The optimum probability is thus the solution of the following equation

$$\begin{aligned} 0 &= \frac{\partial}{\partial \overline{P_i}} Q(\lambda,\bar{\lambda}) \\ &= \sum_{Q_i} P(O,Q_i|\lambda) \sum_{t\in T_i(s)} \left\{ \frac{s_t-1}{\overline{P_i}} - \frac{N-s_t}{1-\overline{P_i}} \right\} \end{aligned} \tag{14}$$

where $T_i(s)$ is the set $\{t:s_{t-1}=i\}$. Interchanging the order of summation, we get

$$\frac{1}{\overline{P_i}(1-\overline{P_i})} \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{s\in S_{ij}(t)} P(O,s|\lambda)(j-N\overline{P_i}) = 0 \tag{15}$$

where $S_{ij}(t)=\{s:s_{t-1}=i,s_t=j\}$. For $P_i\neq 0,1$, we get

$$\overline{P_i} = \frac{\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{s\in S_{ij}(t)} (j-1)P(O,s|\lambda)}{(N-1)\sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{s\in S_{ij}(t)} P(o,s|\lambda)} \tag{16}$$

Apparently $P_i<1$. From eqs.(8)(9), we have

$$\overline{P_i} = \frac{\sum_{t=1}^{T} \sum_{j=1}^{N} (j-1)\alpha_{t-1}(i)a_{ij}b_j(O_t)\beta_t(j)}{(N-1)\sum_{t=1}^{T} \alpha_{t-1}(i)\beta_{t-1}(i)} \tag{17}$$

The estimation of the output and initial distributions will be the same as in the case of discrete symbols or multivariate observations[4,5].

5. Conclusion

We have presented an extension of hidden Markov model theory. This extension makes the HMM fully parametrically representable and thus increases the versatility of the model. We expect that the new method can model the speech signal better. Application of the method to speech recognition is currently under investigation.

Reference

[1] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.ASSP-37, 11, 1641-1648, 1989

[2] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High performance connected digit recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.ASSP-37, 8, 1214-1225, 1989

[3] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markovian process to automatic speech recognition," B.S.T.J., vol.62, no.4, 1035-1074, Apr. 1983

[4] L. R. Liporace,"Maximum likelihood estimation for multivariate observation of Markov sources," IEEE Trans. Inform. Theory, vol.IT-28, 729-34, Sept. 1986

[5] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer,"Speech recognition with continuous parameter hidden Markov models," ICASSP, 40-43,1988