Holt, M.J.J.:  
7th European Signal Processing 
Conference, Edinburgh, 
SIGNAL PROCESSING VII
THEORIES AND APPLICATIONS

Proceedings of EUSIPCO-94
Seventh European Signal Processing Conference
Edinburgh, Scotland, U.K.
13-16 September 1994

Edited by

M.J.J. Holt
University of Technology
Loughborough, U.K.

C.F.N. Cowan
University of Technology
Loughborough, U.K.

P.M. Grant
University of Edinburgh
Edinburgh, U.K.

W.A. Sandham
University of Strathclyde
Glasgow, U.K.

VOLUME II

European Association for Signal Processing
ESPL CH-1015, P.O. Box 134, CH 1000, Lausanne, Switzerland
EURASIP PRESIDENT’S MESSAGE

EURASIPCO conferences are organised in Europe every alternate year by EURASIP, the EUROpean Association for Signal Processing. The previous one was held in Brussels (1992) and the next one will take place in Trieste, Italy, in September 1996.

EURASIP membership is open to all individuals and institutions active or interested in signal processing. Basically, it is a scientific community, with members all over the world, which has the objective of disseminating research results and new technologies, and also of developing communication and friendly relations between its members as well as with other societies. This is achieved through meetings, conferences and workshops, but also perhaps more efficiently through the following set of regular publications:

The NEWSLETTER is the main communication link between members, providing information about the situation of the society and its activities.

SIGNAL PROCESSING is the journal dedicated to the general aspects of the field: theory, implementation and applications.

SPEECH COMMUNICATION is edited jointly with ESCA, the European Speech Communication Association, and covers the theoretical and experimental aspects of speech communication processes.

IMAGE COMMUNICATION is the journal dedicated to the theory and practice of image processing in the communication area.

As a well-established community, EURASIP has an impact on its environment which goes beyond its directly visible output.

The scientific life in Europe is dominated by two different sets of actions: the European research and education programmes and the standardisation activities. Although EURASIP, so far, has been unable to interact officially with either of these, it is, in fact, involved through its members, who belong to expert teams and who work in international consortia, and indeed it helps and facilitates participation.

On one specific issue EURASIP can have a significant contribution: developing relationships with eastern European countries for education and research.

The success of EURASIPCO-94 is just another illustration of the strength and vitality of the signal processing technical area. Considering the evolution of relations between countries within this continent and in the whole world, as well as the sustained pace of technical progress, it seems increasingly important to have strong associations in Europe like EURASIP, to accompany the technical and social evolution in the best possible harmony.

Therefore, I encourage all of you to join our society, not only to benefit from the material advantages it offers, such as reduced subscription rates for journals and reduced fees for conferences and workshops, but also to take an active part in improving our scientific environment and, at the same time, improving our human society as a whole.

M. Bellanger

President of EURASIP
CHAIRMAN’S MESSAGE

It is our pleasure to welcome you to EUSIPCO-94 in Edinburgh. In addition to being the capital city of Scotland, Edinburgh is a major centre for banking, investment and the legal profession, as well as accommodating a thriving electronics industry in its environs. GEC-Marconi Avionics, formerly Ferranti, is a major employer within Edinburgh itself, with Hewlett Packard who came to Scotland in the 1960s, and Digital Equipment Corporation (DEC) in the 1980s, located at South Queensferry. Many other microelectronics companies such as Motorola, NEC and SUN Microsystems, are located close-by in the central belt.

EUSIPCO-94 will be located in the Appleton Tower complex of the University of Edinburgh, close to the High Street and Edinburgh Old Town. There, we will see and hear the 490 paper presentations which have been selected from the 760 submitted summaries offered to us. Selecting this sub-set of papers for presentation was a difficult and time-consuming task, and we are very grateful to our panel of more than 120 international experts who willingly gave their time to advise us in this.

On behalf of the Organising Committee, we wish you a warm welcome to Edinburgh, a successful EUSIPCO-94, and an enjoyable and valuable time reading the many papers compiled in this three volume Proceedings.

C.F.N. Cowan                                      P.M. Grant
General Chairman                                 Co-Chairman (Edinburgh)
EUSIPCO-94 ORGANISING COMMITTEE

GENERAL CHAIRMAN
Colin Cowan
Dept. of Electronic & Electrical Engineering
University of Technology
Loughborough LE11 3TU
England, U.K.
Tel: +44 (0)9 223 468
Fax: +44 (0)9 222 830
E-Mail: c.f.m.cowan@lut.ac.uk

CO-CHAIRMAN
Peter Grant
Dept. of Electrical Engineering
University of Edinburgh
Edinburgh EH9 3JL
Scotland, U.K.
Tel: +44 (31) 650 5569
Fax: +44 (31) 650 6534
E-Mail: pmg@ee.ed.ac.uk

TECHNICAL PROGRAMME
Murray Holt
Dept. of Electronic & Electrical Engineering
University of Technology
Loughborough LE11 3TU
England, U.K.
Tel: +44 (0)9 222 847
Fax: +44 (0)9 222 854
E-Mail: m.holt@lut.ac.uk

FINANCE
Bernard Mulgrew
Dept. of Electrical Engineering
University of Edinburgh
Edinburgh EH9 3JL
Scotland, U.K.
Tel: +44 (31) 650 5580
Fax: +44 (31) 650 6554
E-Mail: bernie@ee.ed.ac.uk

EXHIBITS
Ed Chapell
Avonbank Cottage
Blackstone Road
Avonbridge
Falkirk FK1 2LB
Scotland, U.K.
Tel: +44 [32] 486 351

REGISTRATION
Roger Clarke
Dept. of Computing & Electrical Engineering
Heriot Watt University
Riccarton
Edinburgh EH14 4AS
Scotland, U.K.
Tel: +44 [31] 451 3323
Fax: +44 [31] 451 3327
E-Mail: rjc@cee.hw.ac.uk

PUBLICATIONS
Bill Sandham
Dept. of Electronic & Electrical Engineering
University of Strathclyde
Glasgow G1 1XW
Scotland, U.K.
Tel: +44 [41] 552 4400
Fax: +44 [41] 552 2487
E-Mail: wbs@spd.eee.strath.ac.uk

LOCAL ARRANGEMENTS
Stephen McLaughlin
Dept. of Electrical Engineering
University of Edinburgh
Edinburgh EH9 3JL
Scotland, U.K.
Tel: +44 [31] 650 5578
Fax: +44 [31] 650 6554
E-Mail: sml@ee.ed.ac.uk

PUBLICITY
Keith Brown
Dept. of Computing & Electrical Engineering
Heriot Watt University
Riccarton
Edinburgh EH16 4AS
Scotland, U.K.
Tel: +44 [31] 451 3351
Fax: +44 [31] 451 3327
E-Mail: keb@cee.hw.ac.uk

ADMINISTRATION
Anne Hall
St Leonard’s Hall
University of Edinburgh
Edinburgh EH16 5AY
Scotland, U.K.
Tel: +44 [31] 667 1971
Fax: +44 [31] 668 3217
<table>
<thead>
<tr>
<th>Name</th>
<th>Initial(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arce</td>
<td>G.</td>
</tr>
<tr>
<td>Arp</td>
<td>F.</td>
</tr>
<tr>
<td>Atal</td>
<td>B. S.</td>
</tr>
<tr>
<td>Basseville</td>
<td>M.</td>
</tr>
<tr>
<td>Bellanger</td>
<td>M.</td>
</tr>
<tr>
<td>Bentley</td>
<td>P.</td>
</tr>
<tr>
<td>Benveniste</td>
<td>A. I.</td>
</tr>
<tr>
<td>Bernstein</td>
<td>R.</td>
</tr>
<tr>
<td>Bershad</td>
<td>N. J.</td>
</tr>
<tr>
<td>Biemond</td>
<td>J.</td>
</tr>
<tr>
<td>Bisset</td>
<td>D.</td>
</tr>
<tr>
<td>Boehme</td>
<td>J. F.</td>
</tr>
<tr>
<td>Boite</td>
<td>R.</td>
</tr>
<tr>
<td>Boles</td>
<td>P.</td>
</tr>
<tr>
<td>Bonnet</td>
<td>M.</td>
</tr>
<tr>
<td>Boyce</td>
<td>J.</td>
</tr>
<tr>
<td>Caelen</td>
<td>J.</td>
</tr>
<tr>
<td>Cain</td>
<td>G. D.</td>
</tr>
<tr>
<td>Cappellini</td>
<td>V.</td>
</tr>
<tr>
<td>Chapman</td>
<td>R.</td>
</tr>
<tr>
<td>Chaudhuri</td>
<td>B. B.</td>
</tr>
<tr>
<td>Christie</td>
<td>P.</td>
</tr>
<tr>
<td>Clarke</td>
<td>R.</td>
</tr>
<tr>
<td>Coltuc</td>
<td>D.</td>
</tr>
<tr>
<td>Comon</td>
<td>P.</td>
</tr>
<tr>
<td>Constantinides</td>
<td>A. G.</td>
</tr>
<tr>
<td>Cowan</td>
<td>C. F.</td>
</tr>
<tr>
<td>Cuperman</td>
<td>V.</td>
</tr>
<tr>
<td>Datta</td>
<td>S.</td>
</tr>
<tr>
<td>Davies</td>
<td>E. R.</td>
</tr>
<tr>
<td>Descout</td>
<td>R.</td>
</tr>
<tr>
<td>Devinlin</td>
<td>M.</td>
</tr>
<tr>
<td>Dewilde</td>
<td>P. M.</td>
</tr>
<tr>
<td>Doncarli</td>
<td>C.</td>
</tr>
<tr>
<td>Downton</td>
<td>A. C.</td>
</tr>
<tr>
<td>Duhamel</td>
<td>P.</td>
</tr>
<tr>
<td>Durrani</td>
<td>T. S.</td>
</tr>
<tr>
<td>Elliott</td>
<td>S. J.</td>
</tr>
<tr>
<td>Feng</td>
<td>G.</td>
</tr>
<tr>
<td>Fettweis</td>
<td>A.</td>
</tr>
<tr>
<td>Fidler</td>
<td>J. K.</td>
</tr>
<tr>
<td>Figueiras-Vidal</td>
<td>A. R.</td>
</tr>
<tr>
<td>Fraser</td>
<td>A.</td>
</tr>
<tr>
<td>Frei</td>
<td>W.</td>
</tr>
<tr>
<td>Galand</td>
<td>C.</td>
</tr>
<tr>
<td>Gersho</td>
<td>A.</td>
</tr>
<tr>
<td>Gharavi</td>
<td>H.</td>
</tr>
<tr>
<td>Giannakis</td>
<td>G. B</td>
</tr>
<tr>
<td>Goulty</td>
<td>N. R.</td>
</tr>
<tr>
<td>Granlund</td>
<td>G. H.</td>
</tr>
<tr>
<td>Grant</td>
<td>P. M.</td>
</tr>
<tr>
<td>Griffiths</td>
<td>J. W.</td>
</tr>
<tr>
<td>Haralick</td>
<td>R. M.</td>
</tr>
<tr>
<td>Haskey</td>
<td>P.</td>
</tr>
<tr>
<td>Hatton</td>
<td>L.</td>
</tr>
<tr>
<td>Hess</td>
<td>W.</td>
</tr>
<tr>
<td>Heute</td>
<td>U.</td>
</tr>
<tr>
<td>Holt</td>
<td>M. J.</td>
</tr>
<tr>
<td>Huang</td>
<td>T. S.</td>
</tr>
<tr>
<td>Hudson</td>
<td>J.</td>
</tr>
<tr>
<td>Jones</td>
<td>S. R.</td>
</tr>
<tr>
<td>Jordan</td>
<td>J. R.</td>
</tr>
<tr>
<td>Kaiser</td>
<td>J. F.</td>
</tr>
<tr>
<td>Kittler</td>
<td>J.</td>
</tr>
<tr>
<td>Kotropoulos</td>
<td>C.</td>
</tr>
<tr>
<td>Kunt</td>
<td>M.</td>
</tr>
<tr>
<td>Lacoume</td>
<td>J. L.</td>
</tr>
<tr>
<td>Lagunas</td>
<td>M. A.</td>
</tr>
<tr>
<td>Laver</td>
<td>J.</td>
</tr>
<tr>
<td>Leggett</td>
<td>M.</td>
</tr>
<tr>
<td>Lind</td>
<td>L. F.</td>
</tr>
<tr>
<td>Ljung</td>
<td>L.</td>
</tr>
<tr>
<td>Lowe</td>
<td>D.</td>
</tr>
<tr>
<td>Macchi</td>
<td>O.</td>
</tr>
<tr>
<td>Manolakis</td>
<td>D.</td>
</tr>
<tr>
<td>Maragos</td>
<td>P.</td>
</tr>
<tr>
<td>Martinez</td>
<td>C.</td>
</tr>
<tr>
<td>McCanny</td>
<td>J. V.</td>
</tr>
<tr>
<td>McDonnell</td>
<td>J. T.</td>
</tr>
<tr>
<td>McLaughlin</td>
<td>S.</td>
</tr>
<tr>
<td>McQuilllin</td>
<td>R.</td>
</tr>
<tr>
<td>McWhirter</td>
<td>J. G.</td>
</tr>
<tr>
<td>Mersereau</td>
<td>R. M.</td>
</tr>
<tr>
<td>Moore</td>
<td>R.</td>
</tr>
<tr>
<td>Mulgrew</td>
<td>B.</td>
</tr>
<tr>
<td>Nickel</td>
<td>U.</td>
</tr>
<tr>
<td>Nikias</td>
<td>C. L.</td>
</tr>
<tr>
<td>Noll</td>
<td>P.</td>
</tr>
<tr>
<td>Parish</td>
<td>D. J.</td>
</tr>
<tr>
<td>Peardon</td>
<td>L.</td>
</tr>
<tr>
<td>Piper</td>
<td>J.</td>
</tr>
<tr>
<td>Pitas</td>
<td>I.</td>
</tr>
<tr>
<td>Proakis</td>
<td>J. G.</td>
</tr>
<tr>
<td>Proudler</td>
<td>I. K.</td>
</tr>
<tr>
<td>Ramponi</td>
<td>G.</td>
</tr>
<tr>
<td>Rix</td>
<td>H.</td>
</tr>
<tr>
<td>Roberts</td>
<td>S.</td>
</tr>
<tr>
<td>Rosenfeld</td>
<td>A.</td>
</tr>
<tr>
<td>Sandham</td>
<td>W. A.</td>
</tr>
<tr>
<td>Savva</td>
<td>H.</td>
</tr>
<tr>
<td>Schmale</td>
<td>B.</td>
</tr>
<tr>
<td>Schreiber</td>
<td>W. F.</td>
</tr>
<tr>
<td>Schuessler</td>
<td>H. W.</td>
</tr>
<tr>
<td>Serra</td>
<td>J.</td>
</tr>
<tr>
<td>Sicuranza</td>
<td>G. L.</td>
</tr>
<tr>
<td>Stock</td>
<td>D. T.</td>
</tr>
<tr>
<td>Smith</td>
<td>P. R.</td>
</tr>
<tr>
<td>South</td>
<td>C. R.</td>
</tr>
<tr>
<td>Stoica</td>
<td>P.</td>
</tr>
<tr>
<td>Sugrue</td>
<td>M. J.</td>
</tr>
<tr>
<td>Tattersall</td>
<td>G. D.</td>
</tr>
<tr>
<td>Theodoridis</td>
<td>S.</td>
</tr>
<tr>
<td>Tran</td>
<td>T.</td>
</tr>
<tr>
<td>Trancoso</td>
<td>I.</td>
</tr>
<tr>
<td>Viberg</td>
<td>M.</td>
</tr>
<tr>
<td>Ward</td>
<td>C. R.</td>
</tr>
<tr>
<td>Westall</td>
<td>F.</td>
</tr>
<tr>
<td>Wong</td>
<td>K.</td>
</tr>
<tr>
<td>Woods</td>
<td>J. W.</td>
</tr>
<tr>
<td>Xydeas</td>
<td>C. S.</td>
</tr>
<tr>
<td>Yegnanarayana</td>
<td>B.</td>
</tr>
<tr>
<td>Zamperoni</td>
<td>P.</td>
</tr>
<tr>
<td>Ziolkowski</td>
<td>A.</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

EUSIPCO-94 has been organised in co-operation with the following institutions:

- Institution of Electrical Engineers (IEE)
- IEEE Signal Processing Society
- Lothian & Edinburgh Enterprise Ltd (LEEL)
- Loughborough University of Technology
- University of Edinburgh
TECHNICAL PROGRAMME

NOMENCLATURE
All papers are of the following form:

\[ slp \]

where

\( s \) represents a Session Number (1-8)
or the Opening Session (OS)

\( l \) indicates a Lecture Theatre (A-E),
Poster Display Area (P)
or an Invited Paper (I)

\( p \) indicates a Paper Number (1-20)
OPENING SESSION

Time & Place: 9.00-10.30, McEwan Hall
Chairperson: C.F.N. Cowan, EUSIPCO-94 General Chairman

1OS.1 Opening Address
S. Sutherland, Principal and Vice Chancellor, University of Edinburgh

1OS.2 EURASIP Address
M.G. Bellanger, President of EURASIP

1OS.3 Digital video processing - the next mass market
J.R. Forrest, National Transcommunications Ltd, Winchester, U.K.

Session 1A: SPEECH SYNTHESIS

Time & Place: 11.00-12.30, Lecture Theatre A (Oral Session)
Chairperson: G. Carayannis, Institute for Language & Speech Processing, 22 Magari Street, Athens 11525, GREECE

1A.1 Word level stress as an intrinsic property of word level prosody and pitch modeling through a pencil of functions
M.A. Vlahakis, Institute for Language and Speech Processing, Athens, GREECE; S.F. Fotinea, National Technical University of Athens, Athens, GREECE; and G. Carayannis, Institute for Language and Speech Processing, Athens, GREECE

1A.2 Investigation on the audibility of glottal parameter variations in speech synthesis
J. Linden and J. Skoglund, Chalmers University of Technology, Goeteborg, SWEDEN

1A.3 On the ability of various speech models to smooth segment discontinuities in the context of text-to-speech synthesis by concatenation
T. Dutoit and H. Leich, Faculte Polytechnique de Mons, Mons, BELGIUM

1A.4 A novel algorithm based on sinusoidal modelling of speech for text-to-speech synthesiser
M. Larreategui and R.A. Carrasco, Staffordshire University, Stafford, U.K.

1A.5 A speech segmentation method based on the Itakura distance and its use for the construction of a segment inventory for TTS
M.A. Vlahakis, Institute for Language & Speech Processing, Athens, GREECE; S.F. Fotinea, National Technical University of Athens, Athens, GREECE; and G. Carayannis, Institute for Language & Speech Processing, Athens, GREECE
1B.1 Time recursive 3D segmentation for image sequences
M. Pardas and P. Salembier, ETSETB, Universitat Politecnica de Catalunya, Barcelona, SPAIN

1B.2 Multi-frame based segmentation of moving objects by combining luminance and motion
P. Schroeter and S. Ayer, Swiss Federal Institute of Technology, Lausanne, SWITZERLAND

1B.3 A low-level segmentation procedure for color images
R. Schettini and M. Suardi, IFCTR-ITIM, CNR, Milano, ITALY

1B.4 Image segmentation and texture classification using local thresholds and 2-D AR modelling
K. Chehdi and C. Cariou, ENSSAT, Lannion, FRANCE

1B.5 A new approach to dynamic thresholding
M.K. Yanni and E. Horne, University of Kent, Canterbury, U.K.

1C.1 Performance of TLS based noise reduction techniques on chaotic time series
J.K. Martin and A.K. Nandi, University of Strathclyde, Glasgow, U.K.

1C.2 State estimation of chaotic systems with application to information transmission and communication
G.C. Freeland and T.S. Durrani, University of Strathclyde, Glasgow, U.K.

1C.3 A method for determining the year dynamic model of ADCs
N.Y. Semyonova and V. Academy of Sciences, Riga, LATVIA

1C.4 A fast mutual information calculation algorithm
H-P. Bernhard and G. Kubin, Vienna University of Technology, Vienna, AUSTRIA

1C.5 Parametric estimation of probability density functions through a convolution of one-sided exponentials
J. Vidal, A. Bonafonte and J.A.R. Fonollosa, ETSETB, Universitat Politecnica de Catalunya, Barcelona, SPAIN
1D.1 Electrocardiogram enhancement by adaptive filtering  
A. Albiol and V. Almenar, ETSI Telecomunicacion, Universidad Politecnica de Valencia, Valencia, SPAIN

1D.2 Comparison of spectral analysis algorithms for use in spectral phonocardiography  
H.P. Sava and J.T.E. McDonnell, University of Edinburgh, Edinburgh, U.K.

1D.3 Adaptive noise removal from complex signals using the wavelet transform  

1D.4 Measuring shape variations of ECG waves through time-frequency representations  
B. Oficjalska, H. Rix, University of Nice, Sophia Antipolis, Valbonne, FRANCE; E. Chevalier, ESIM, Marseille, FRANCE; J. Fayn, INSERM U121, Lyon, FRANCE; A. Varenne,

1D.5 Combined representation of functional and morphological informations of human peripheral nerves  
C. Martinez, C. Doncarli, Ecole Centrale de Nantes, Nantes, FRANCE; and P. Guicheune, Centre Hospitalier Regional et Universitaire de Nantes, Nantes, FRANCE

1D.6 Wavelet analysis of cardiovascular signals  
P.M. Bentley and J.T.E. McDonnell, University of Edinburgh, Edinburgh, U.K.

Session 1E: SONAR

1E.1 Array shape calibration and source localization with an array of unknown geometry  
P. Forster, CNAM, Paris, FRANCE; and F. Martinerie, Thomson Sintra SM, Arcueil, FRANCE

1E.2 Multipath signal detection with different geometric properties for information about their properties  
V. Baronkin and A. Treniki, Institute for Nuclear Research, Moscow, RUSSIA

1E.3 Detection in non-homogeneous background environment  
L.A. Prastitis, Intercollege, Nicosia, CYPRUS; and S.D. Himonas, New York Institute of Technology, New York, U.S.A.

1E.4 Curve line extraction using statistical perceptual grouping  
J-C. Di Martino and F. Alexandre, CRIN-CNRS-INRIA, Vandoeuvre, FRANCE

1E.5 Continuous HMM tracking  
F. Martinerie, Thomson Sintra ASM, Arcueil, FRANCE
SESSION 2I: INVITED LECTURE

Time & Place: 1.30-2.20, Lecture Theatre A

2I.1 Intelligent signal processing: a way of the future
S. Haykin, McMaster University, Hamilton, Ontario, CANADA

SESSION 2A: SPEECH RECOGNITION

Time & Place: 2.20-5.50, Lecture Theatre A (Oral Session)
Chairperson: C.S. Xydeas, Dept. of Electronic & Electrical Engineering, University of Manchester, Oxford Road, Manchester M13 9PL, U.K.

2A.1 A study of robust isolated word speech recognition based on fuzzy methods
L. Cong, C.S. Xydeas and A.F. Erwood, University of Manchester, Manchester, U.K.

2A.2 Experiments with neural networks in isolated word recognition tasks
X. Menendez-Pidal, J. Macias-Guarasa, M. Leandro, J.A. Vallejo and J.M. Pardo, ETSI de Telecomunicacion (UPM), Madrid, SPAIN

2A.3 Subword segmentation alternatives for isolated and connected words recognition
A. Herrera, V.R. Algazi, K. Brown and D. Irvine, University of California, Davis, CA, U.S.A.

2A.4 Markov chain models in isolated word recognition
J. Dai, University of Nanjing, Nanjing, CHINA

2A.5 Influence of vector quantization on isolated word recognition
V. Fontaine, H. Leich, J. Hennebert and M. Hasler, Faculte Polytechnique de Mons, Mons, BELGIUM

2A.6 A monolithic speech recognizer based on fully recurrent neural networks
K. Kasper, H. Reiningter, D. Wolf and H. Wuest, Johann Wolfgang Goethe-Universitaet, Frankfurt am Main, GERMANY

2A.7 Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task

2A.8 Large-vocabulary continuous-speech recognition by partitioned graph search
C.C. Chiu, Feng Chia University, Taichung, TAIWAN, R.O.C.; J.R. Deller, Michigan State University, East Lansing, MICHIGAN, and C.G. Venkatesh, CTA Incorporated, Bedford, MA, U.S.A.

2A.9 A new reestimation formula for hidden Markov models to enforce state transition constraints
P. Schauble, Swiss Federal Institute of Technology, Zurich, SWITZERLAND

2A.10 Continuous speech phoneme recognition using two stage probabilistic modeling
R.N.V. Sitaram and T.V. Sreenivas, Indian Institute of Science, Bangalore, INDIA

Page No.
Session 2B: IMAGE CODING I

Time & Place: 2.20-5.50, Lecture Theatre B (Oral Session)

Chairperson: M. Ghanbari, Dept. of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester, Essex CL4 3SQ, U.K.

2B.1 A classifier constrained vector quantization algorithm for the compression of image sequence frames in DCT domain
M.A. Turker and M. Severcan, Scientific and Technical Research Council of Turkey, Ankara, TURKEY

2B.2 Multi-criterion segmentation for image sequence coding
C. Gu and M. Kunt, Swiss Federal Institute of Technology, Lausanne, SWITZERLAND

2B.3 Feature-based video coding using mathematical morphology
J.R. Casas and L. Torres, ETSETB, Universitat Politecnica de Catalunya, Barcelona, SPAIN

2B.4 Temporal linking of motion-based segmentation for object-oriented image sequence coding
V. Garcia-Gardinero, C. Labit and L. Bonnaud, IRISA, Rennes, FRANCE

2B.5 Very low bit-rate segmentation-based video coding
F. Eryurtlu, A.M. Kondoza and B.G. Evans, University of Surrey, Guildford, U.K.

2B.6 Object-based stereoscopic coding: vector field estimation and object segmentation
R.E.H. Franich, R.L. Lagendijk and J. Biemond, Delft University of Technology, Delft, NETHERLANDS

2B.7 Planetary image data compression via piecewise polynomial approximation
R. Sandau and O. Frauenberger, Institute for Space Sensor Technology, Berlin, GERMANY

2B.8 Adaptive transform image coding based on variable-shape-blocks and directional autocorrelation models
I. Matsuda, S. Itoh and T. Utsunomiya, Science University of Tokyo, Tokyo, JAPAN

2B.9 BSP tree coding of images using symmetry information
C. Saraceno and R. Leonardi, University of Brescia, Brescia, ITALY

2B.10 Polygonal segmentation for iterated transformations systems based image coding
E. Reusens, Swiss Federal Institute of Technology, Lausanne, SWITZERLAND

Session 2C: HIGHER ORDER STATISTICS I (SPECIAL SESSION)

Time & Place: 2.20-5.50, Lecture Theatre C (Oral Session)

Chairperson: A.K. Nandi, Dept. of Electronic & Electrical Engineering, University of Strathclyde, 204 George Street, Glasgow G1 1XW, U.K.

2C.1 A singular value decomposition for higher-order tensors and application to independent component analysis
L. De Lathauwer, B. De Moor and J. Vandewalle, K.U. Leuven, Heverlee, BELGIUM

2C.2 DOA estimation based on generalized Eigendecomposition of generalized cumulants
G. Scarano and G. Jacovitti, University of Rome "La Sapienza", Rome, ITALY
2C.3 Adaptive source separation without pre-whitening
B. Laheld and J-F. Cardoso, Telecom Paris, Paris, FRANCE

2C.4 Separation of a polynomial mixture of independent sources using higher-order
output moment matrices
M. Krob and M. Benidir, Ecole Superieure d'Electricite, Gif-sur-Yvette, FRANCE

2C.5 Localization of moving sources using sensor arrays and cyclic HOS
G.B. Giannakis and S. Shamsunder, University of Virginia, Charlottesville, VA, U.S.A.

2C.6 Blind identification of nonminimum phase filters by maximizing the kurtosis
M. Boumahdi and J-L. Lacoume, CEPHAG/ENSIEG, Saint-Martin d'Heres, FRANCE

2C.7 Identification of FIR sytems using third order cumulants
A.K. Nandi, University of Strathclyde, Glasgow, U.K.

2C.8 ARMA identification using high-order statistics based linear methods - an unified
presentation
G. Favier, D. Dembele and J-L. Peyre, CNRS/UNSA, Valbonne, FRANCE

2C.9 A robust approach to least squares solutions in HOC based system identification
A.G. Stogioglou and S. McLaughlin, University of Edinburgh, Edinburgh, U.K.

2C.10 Efficient algorithms for parametric Volterra system identification
N. Kalouptsidis, G-O. Glentis and P. Koukoulas, University of Athens, Athens, GREECE

Session 2D: SEISMIC SIGNAL PROCESSING (SPECIAL SESSION)

Time & Place: 2.20-5.50, Lecture Theatre D (Oral Session)
Chairperson: W.A. Sandham, SPD, Dept. of Electronic & Electrical Engg., University of
Strathclyde, 204 George Street, Glasgow G1 1XW, U.K.

2D.1 How well do signal processing algorithms written independently agree? (invited
paper)
L. Hatton, 11 Carlton Road, New Malden, U.K.

2D.2 First break detection in seismic reflection data with Fuzzy ARTMAP neural
networks

2D.3 Seismic event tracking with PDA in an interpretation environment
M. Leggett, C.A. Woodham, W.A. Sandham and T.S. Durrani, University of Strathclyde,
Glasgow, U.K.

2D.4 Fast radon transform for detection of seismic reflections
P.A. Toft, Technical University of Denmark, Lyngby, DENMARK; and K.V. Hansen,
Odegaard & Danneskiold-Samsoe Aps, DENMARK

2D.5 Arrival times estimation by means of the Wavelet transform
G. Olmo and L. Lo Presti, Politecnico di Torini, Torini, ITALY

2D.6 Estimation of formation shear velocity using the wavelet transform in acoustic
logging
N. Thirion, CEPHAG-ENSIEG, Saint-Martin d'Heress, FRANCE; J. Mars, 62 Aloha
Street, San Francisco, CA, U.S.A.; J-L. Mari, Institut Francais du Petrole, Rueill-
Malmaison, FRANCE; and P. Volant, CEPHAG-ENSIEG, Saint-Martin d'Heress,
FRANCE
2D.7 Wavefield decomposition using parametric inversion
P. Gavrin, Institut Francais du Petrole, Pau Cedex, FRANCE; J. Mars, 62 Aloha Street,
San Francisco, CA, U.S.A.; E. Bouisson, P. Delmas and A. Essebar, CEPHAG-
ENSIEG, Saint-Martin d’Heres, FRANCE

2D.8 Comparative study of different methods for separating waves: application to VSP
data
J-L. Mari, Institut Francais du Petrole, Rueil-Malmaison, FRANCE; and F. Glangeaud,
CEPHAG-ENSIEG, Saint-Martin d’Heres, FRANCE

2D.9 Iterative reconstruction of multidimensional objects buried in inhomogeneous
elastic media
E.J. Bellegarda, IBM Research, Yorktown Heights, NY, U.S.A.; and T.M. Habashy,
Schlumberger-Doll Research, Ridgefield, CT, U.S.A.

2D.10 Blind identification of nonminimum phase ARMA filters: application to field
seismic data
M. Boumahdi, F. Glangeaud and J-L. Lacoume, CEPHAG-ENSIEG, Saint-Martin
d’Heres, FRANCE

Session 2E: DETECTION

Time & Place: 2.20-5.50, Lecture Theatre E (Oral Session)
Chairperson: A.G. Constantinides, Department of Electrical Engineering, Imperial College
London, Exhibition Road, London SW7 2BT, U.K.

2E.1 Adaptive detectors for multi-channel signals in code division multiple access
systems
A. Haimovich, R. Om and Y. Bar-Ness, New Jersey Institute of Technology, Newark,
NJ, U.S.A.

2E.2 A suboptimal receiver for multiple-access channels with intersymbol and
interchannel interference using the M-algorithm
W. Sauer-Greff, Universitaet Kaiserslautern, Kaiserslauern, GERMANY; and R.A.
Kennedy, Australian National University, Canberra, ACT, AUSTRALIA

2E.3 Analysis and interpretation of the reduced-rank generalized likelihood-ratio test
I.P. Kirsteins, SACLANT Undersea Research Centre, La Spezia, ITALY

2E.4 Determination of the signal-subspace dimension: a nonparametric approach based
on the eigenvectors analysis
V.L. Koliadin, Kharkov Aviation Institute, Kharkov, UKRAINE

2E.5 Detection of transient signals using time-frequency distributions
O. Lemoine, S. Icart and C. Berenguer, CNRS-UNSA, Valbonne, FRANCE

2E.6 Optimal space dimension for change detection problems
D. Brie, M. Tomczak and D. Dal Ponte, Centre de Recherche en Automatique de Nancy,
Vandoeuvre, FRANCE

2E.7 Abrupt changes detection via singular system analysis
P. Poignet and M. Guglielmi, Ecole Centrale de Nantes, Nantes, FRANCE

2E.8 The application of circular statistics to specific radar pulse train detection
S.D. Elton, Defence Science and Technology Organisation, Salisbury SA, AUSTRALIA;
and D.A. Gray, University of Adelaide, Adelaide SA, AUSTRALIA
Session 2P: MULTIDIMENSIONAL DSP

Time & Place: 2.20-5.50, Poster Rooms (Poster Session)
Chairperson: K. Brown, Dept. of Computing and Electrical Engineering, Heriot Watt University, Riccarton, Edinburgh Scotland EH14 4AS, U.K.

2P.1 Optimal stack filter design with symmetry constraints
I. Tabus, D. Petrescu, Polytechnic Institute of Bucharest, Bucharest, ROMANIA; and M. Gabbouj, Tampere University of Technology, Tampere, FINLAND

2P.2 Least squares reconstruction of band-limited images from projections
J. Le Roux, P.E. Lise and E. Zerbib, University of Nice, Sophia Antipolis, Valbonne, FRANCE

2P.3 Computer modelling of point spread function for 3D image restoration
M. Razaz, R.A. Lee, University of East Anglia, Norwich, U.K.; and P.J. Shaw, John Innes Institute, Norwich, U.K.

2P.4 Regularized myopic image deconvolution. Application to aperture synthesis in radioastronomy
R. Prost, INSA-Lyon, Villeurbanne, FRANCE; J.L. Burdeau, INT-Evry, Evry, FRANCE; S. Guilloteau, IRAM, Saint-Martin d’Heres, FRANCE; and R. Goutte, INSA-Lyon, Villeurbanne, FRANCE

2P.5 Fast implementation of the parametric projection filter for image restoration
M. Ouldmanmar, University of Science and Technology of Oran, Oran, ALGERIA; H. Ogawa, Tokyo Institute of Technology, Tokyo, JAPAN; M. Kechch and A. Ouamri, University of Science and Technology of Oran, Oran, ALGERIA

2P.6 On implementation of two-dimensional filter banks
R. Stasinski and T.A. Ramstad, The Norwegian Institute of Technology, Trondheim, NORWAY

2P.7 2-D extensions of 1-D filter banks/biorthogonal wavelets with selectivity and regularity constraints
H. Le Bihan and P. Siohan, CCETT, Cesson-Sevigne, FRANCE

2P.8 Shuffling based design technique for nonseparable linear phase 2-D filter banks
E. Anarim, B. Sankur, Bogazici University, Istanbul, TURKEY; H. Caglar, Marmara Research Center, TUBITAK, Gebze-Kocael, TURKEY; and I. Celusun, Bogazici University, Istanbul, TURKEY

2P.9 A linear operator approach for designing multirate filter banks
S. Wada and H. Inaba, Tokyo Denki University, Satima, JAPAN

2P.10 Comparison between two recent stability tests for 2-D digital recursive filters
M. Barret, Ecole Superieure D’Electricite, Metz, FRANCE; and M. Benidir, Universite de Paris-Sud, Paris, FRANCE
2P.11 Fast spatial interference canceler based on the Householder transformation  
R. Onn, New Jersey Institute of Technology, Newark, NJ, U.S.A.; and Y. Bar-Ness, Delft University of Technology, Delft, NETHERLANDS  

2P.12 Robust multi-source beamforming via LMS-like target tracking  

2P.13 Construction of aperiodic decorrelation arrays for imaging systems with coded source  
H.D. Schotten, Aachen University of Technology, Aachen, GERMANY; and J. Vell, Research Center Juelich GmbH, Juelich, GERMANY  

2P.14 Depth of marine pipeline burial determination using an array  
T. Hanson, ESIGETEL, Avon-Fontainbleau, FRANCE  

2P.15 A sparse approach to partially adaptive airborne radar  
I. Scott and B. Mulgrew, University of Edinburgh, Edinburgh, U.K.  

2P.16 A segmental waveform matching algorithm in the wavelet transform domain and its application to image processing.  
T. Saito, S. Watanabe, Kanagawa University, Yokohama, JAPAN; and C.K. Cheong, University of Tokyo, Tokyo, JAPAN  

2P.17 Frequency effect on effective correlation for translational and rotational invariance processing  
WIThDRAWn  
D. Grenier and S. Pare, Laval, Quebec City, CANADA  

2P.18 Accuracy of wave parameter estimation using polarization sensitive arrays  
D. Sidorovitch, D. Maiwald and J.F. Boehme, Ruhr University Bochum, Bochum, GERMANY  

2P.19 Realisation of complex coefficient digital filters based on hypercomplex arithmetic  
K. Ueda, K-I. Mizukami and S-I. Takahashi, Keio University, Yokohama, JAPAN  

2P.20 Realization methods of pure imaginary transmission zeros for complex wave digital filters  
O. Yamazaki and S-I. Takahashi, Keio University, Yokohama, JAPAN  

Session 3A:  SPEECH CODING I  

Time & Place:  9.00-12.30, Lecture Theatre A (Oral Session)  
Chairperson:  F. Westall, British Telecom Laboratories, Martlesham Heath, Ipswich, Suffolk IP5 7RE, U.K.  

3A.1 A network-controlled variable rate multipulse speech coder  
V. Abreu, C. Diz and D. Docampo, Universidad de Vigo, Vigo, SPAIN  

3A.2 A variable bit-rate wideband speech coder based on sub-band coding  
G. Hirn, H. Dia and G. Feng, Universite Stendhal, Grenoble, FRANCE  

3A.3 A framework for the analysis of variable bit-rate speech waveform encoders  

3A.4 Multistage vector quantization with dynamic bit allocation  
T. Eriksson, Chalmers University of Technology, Goteborg, SWEDEN
3A.5 ADPCM with non linear predictors
E. Mumolo, A. Carini and D. Francescato, Universita’ di Trieste, Trieste, ITALY

3A.6 Speech coding using bi-harmonic spectral modeling
C. Garcia-Mateo, J.L. Alba-Castro and E.R. Banga, Universidad de Vigo, Vigo, SPAIN

3A.7 Sinewave amplitude coding using high-order allpole models
R.J. McAulay, T.F. Quatieri, Massachusetts Institute of Technology, Lexington, MA, U.S.A.; and T.G. Champion, R/L/EKT, Hanscom AFB, U.S.A.

3A.8 Optimal wavelets for high quality speech coding
R. Kastantin, D. Stefanoiu, G. Feng, Universite Stendhal, Grenoble, FRANCE; N. Martin, CEPHAG, ENSIEG, St-Martin-d’Heres, FRANCE; and M. Mrayati, Institut Superieur des Sciences Appliquees et de Technologie, Damascus, SYRIA

3A.9 Performance of immittance spectral pairs (ISP) in LPC scalar quantizer schemes
S. Pellet and Y. Bistritz, Tel Aviv University, Tel Aviv, ISRAEL

3A.10 A high quality speech coding algorithm suitable for future INMARSAT systems
S. Yeldener, A.M. Kondoz and B.G. Evans, University of Surrey, Guildford, U.K.

Session 3B: IMAGE RECOGNITION

Time & Place: 9.00-12.30, Lecture Theatre B (Oral Session)
Chairperson: M. Kunt, Laboratoire de Traitement des Signaux, Dept d’Electricite, Ecole Polytechnique Federale Lausanne, 1015 Lausanne, SWITZERLAND

3B.1 Robust multiresolution estimation of parametric motion models in complex image sequences
J.M. Odobez and P. Bouthemy, IRISA/INRIA, Rennes, FRANCE

3B.2 A top-down 3D image sequence segmentation technique controlled by morphological tools
F. Marques, V. Vera and A. Gasull, ETSETB, Universitat Politecnica de Catalunya, Barcelona, SPAIN

3B.3 Efficient implementation of large masks with applications for ultrasonic images
A.K. Nandi, University of Strathclyde, Glasgow, U.K.; and R. Stasinski, Norwegian Institute of Technology, NORWAY

3B.4 Estimation of the multiscaled image edge curves and profiles via the discrete Hermite-Binomial transform
Y.H. Gu, Institute for Perception Research, NETHERLANDS

3B.5 Airport recognition using contextual information
L. Benquet and T. Quiguier, ONERA - DESSTD, Chatillon, FRANCE

3B.6 Image interpretation using symbolic data

3B.7 A dual active contour incorporating parametric shape description
S.R. Gunn and M.S. Nixon, University of Southampton, Southampton, U.K.

3B.8 The application of morphological filters in the colour analysis of complex images
L. Shafarenko, J. Kittler and M. Petrou, University of Surrey, Guildford, U.K.
3B.9  Optimal discrete linear filters for step edge detection in images
J. De Vriendt, University of Gent, Gent, BELGIUM

3B.10  A method for automatic inspection in the textile industry
T. Thomas and M. Cattoen, ENSEEIHT GTTSI, Toulouse, FRANCE

Session 3C:  HIGHER ORDER STATISTICS II (SPECIAL SESSION)

Time & Place:  9.00-12.30, Lecture Theatre C (Oral Session)
Chairperson:  P. Comon, Thomson - Sintra, BP 157, 06903 Sophia Antipolis Cedex, FRANCE

3C.1  ROC curves of skewness and kurtosis statistical tests: application to textures
C. Coroyer, C. Jorand and P. Duvaut, ETS/ENSEA, Cergy, FRANCE

3C.2  Performance of the Skewness-of-Gaussian (SoG) edge extractor
G. Ramponi and S. Carrato, University of Trieste, Trieste, ITALY

3C.3  Speckle masking in the frequency domain: comparison of several algorithms on actual astrophysical data
D. Rossille and J. Le Roux, University of Nice Sophia Antipolis, Valbonne, FRANCE

3C.4  Bispectral processing of short-exposure astronomical images
J.C. Dainty and P. Negrete-Regagnon, Imperial College, London, U.K.

3C.5  New adaptive estimation of the fourth-order cumulant: application to transient detection, blind deconvolution and timing recovery in communication
P.O. Amblard, J.M. Brossier and N. Charkani, CEPHAG/ENSIEG, Saint-Martin d'Heres, FRANCE

3C.6  On robust estimation of cumulants
A.K. Nandi and D. Maempel, University of Strathclyde, Glasgow, U.K.

3C.7  Statistical hypothesis testing using higher-order cumulant spectra
J.W. Dalle Molle, Imperial College, London, U.K.; P. Duvaut,

3C.8  Some bounds on second and third order correlations for bandlimited random processes

3C.9  Estimation of time delays in the blind mixture problem
P. Comon, Thomson-Sintra, Sophia Antipolis, FRANCE

3C.10  Harmonic retrieval in non-Gaussian A/D prefiltering
X-D. Zhang, Y-C. Liang, University, Beijing, CHINA

WITHDRAWN
3D.1 Real-time estimation of the frequency of sinusoidal signals distorted by harmonics
T. Lobos, P. Ruczewski and J. Szymanda, Technical University of Wroclaw, Wroclaw, POLAND

3D.2 A comparative study between two instantaneous frequency estimators
I. Vincent, C. Doncarli, Ecole Centrale de Nantes, Nantes, FRANCE; and F. Auger, IUT de Saint-Nazaire, Saint Nazaire, FRANCE

3D.3 Effect of phase errors on resolution of MUSIC
D.H. Brandwood, Roke Manor Research, Romsey, U.K.

3D.4 A maximum likelihood estimation algorithm for autocovariance matrices of bandlimited processes
T. Chonovel, Franco-Polish School of New Information and Communication Technologies, Poznan, POLAND

3D.5 Linear frequency modulation signal parameters estimation using time-dependent modelling
S. Kovacs, CEPHAG-ENSIEG, Saint-Martin d’Heres, FRANCE; J.M. Piasco, Ecole Centrale de Nantes, Nantes, FRANCE; S. Jespers, CERDSM, Six-Four-Les-Plages, FRANCE; and M. Guglielmi, Ecole Centrale de Nantes, Nantes, FRANCE

3D.6 Cross-terms analysis in Wigner distribution for multicomponent linear FM signals
L. Qiu, Nanyang Technological University, SINGAPORE

3D.7 A method of the vibrating type system impulse response estimation
R. Makowski, Technical University of Wroclaw, Wroclaw, POLAND

3D.8 Parameter estimation of a Middleton class-A noise model by means of the Hough transform
M. Karrakchou, Swiss Federal Institute of Technology, Lausanne, SWITZERLAND

3D.9 Recursive least-squares estimation of continuous models via generalised Poisson moment functionals
H. Garnier, P. Sibille and A. Richard, CRAN - CNRS, Vandoeuvre, FRANCE

3D.10 A statistical inference approach to the parameter estimation of K-distributed noise
L. Vergara, J.A. Megias, Universidad Politecnica de Valencia, Valencia, SPAIN; and P.M. Shankar, Drexel University, Philadelphia, PA, U.S.A.
Session 3E: RADAR

Time & Place: 9.00-12.30, Lecture Theatre E (Oral Session)
Chairperson: J.F. Boehme, Ruhr-Universitat Bochum, Lehstu fiur Signaltheorie, 44780 Bochum, GERMANY

3E.1 Adaptive radar detection in compound-Gaussian clutter
E. Conte, M. Lops and G. Ricci, Universita di Napoli "Federico II", Napoli, ITALY 526

3E.2 Detection analysis of the clutter map CFAR detector with noncoherent integration
C-J. Kim, H-J. Lee, Electronics Telecommunication Research Institute, Taegon, KOREA; and H-S. Lee, Korea Advanced Institute of Science & Technology, KOREA 530

3E.3 Radar clutter rejection: implementation and comparison of adaptive noise cancellation schemes
C. Berenguer and G. Alengrin, Universite de Nice Sophia-Antipolis, Nice, FRANCE 534

3E.4 Clutter pre-whitening process in airborne MTI
X. Sun, Y-N. Peng, Q-G. Liu and D-J. Lu, Tsinghua University, Beijing, CHINA 538

3E.5 A new joint-domain adaptive processing scheme for radars based time-space architecture
G. Liao, Z. Bao and Y. Zhong, University of Xi'an, CHINA WITHDRAWN

3E.6 Antenna phase error of an aperture synthesis radiometer and consequences on the resulting image
P. Thibaut and F. Castanie, ENSEEIHT/GAPSE, Toulouse, FRANCE 541

3E.7 An adaptive nonlinear filtering technique for the initialisation of ISAR autofocusing algorithms
F. Berizzi, E. Dalle Mese, F. Gini and G. Pinelli, Universita di Pisa, Pisa, ITALY 545

3E.8 A robust detection algorithm based on the spectral density
C.R. Guarino, IBM, Gaithersburg, Maryland, U.S.A. 549

3E.9 Using wavelet transform in radar target classification
H-H. Fan, Z-M. Feng, Y-N. Peng and D-J. Lu, Tsinghua University, Beijing, CHINA 553

3E.10 A velocity processor for ultrawideband radars
S. Hussain and H. Al-Ahmad, University of Bradford, Bradford, U.K. 556

Session 3P: IMAGE CODING II

Time & Place: 9.00-12.30, Poster Rooms (Poster Session)

3P.1 Generalized run-length coding for SNR-scalable image compression
S. Simon, University of Ghent, Gent, BELGIUM 560

3P.2 A class of robust learning vector quantizers
R. Yang, M. Gabbouj, Tampere University of Technology, Tampere, FINLAND; I. Pitas and C. Kotropoulos, University of Thessaloniki, Thessaloniki, GREECE 564

3P.3 On deriving a perceptually uniform color space for image coding
F. Cucurullo, S. Curinga, A.A. Grattarola and L. Nobili, University of Genoa, Genova, ITALY 568
3P.4 Three JPEG image coding schemes and their robustness to transmission errors
M. Abdat and M.G. Bellanger, CNAM, Paris, FRANCE

3P.5 A highly parallel neural network architecture for lossy image compression
J. Jiang, Bolton Institute of Higher Education, Bolton, U.K.

3P.6 A comparison between raw and range focused SAR data compression
J-M. Moureaux, P. Gauthier, M. Barlaud, Universite de Nice Sophia Antipolis, Vaibonne, FRANCE; and P. Bellemain, Aerospatiale Etablissement de Cannes, Cannes la Bocca, FRANCE

3P.7 A new VLC decoding technique using code classification
Y.H. Kim, W.J. Kim and S-D. Kim, KAIST, Taejon, KOREA

3P.8 Solution of the inverse problem for IFS models of images by an optimization algorithm
A.M. Carminelli Gregor and A. Bonat, Universita’ di Trieste, Trieste, ITALY

3P.9 An enhancement of the JBIG algorithm for tone image compression
C.W. Jee, Han Yang University, Seoul, KOREA; and J.S. Lee, Korea Academy of Industrial Technology, Seoul, KOREA

3P.10 Image compression and Shannond W. M.G. Beatty and M.M. Dos, University of York, York, U.K.

3P.11 Adaptive image coding using warped polynomial transforms
W. Philips, University of Gent, Gent, BELGIUM

3P.12 Image compression based on reduced rank approximation and adaptive error modelling
S. Chretien and I. Dologlou, Ecole Superieure d'Electricite, Gif-sur-Yvette, FRANCE

3P.13 Pruning the two-dimensional fast cosine transform algorithm
C.A. Christopoulos, VUB - ETRO (IRIS), Brussels, BELGIUM; and A.N. Skodras, University of Patras, Patras, GREECE

3P.14 Modelling of fractal coding schemes
B. Huertgen, Aachen University of Technology, Aachen, GERMANY

3P.15 Coding and transmission of TV sequences at fixed and variable bit rates
F. Argenti, Universita di Firenze, Firenze, ITALY; G. Benelli, Universita di Pavia, Pavia, ITALY; A. Garzelli and M. Rovai, Universita di Firenze, Firenze, ITALY

3P.16 Oriented quincunx subband decomposition for image processing
S. Comes and B. Macq, Universite Catholique de Louvain, Louvain-la-Neuve, BELGIUM

3P.17 Neural modeling for the compression of images through analysis and reconstruction
S. Guillaumeux, M. Klefstad-Sillonville and P. Leray, CCETT, Cesson-Sevigné, FRANCE

3P.18 On the use of transform domain information for concealment of errors in JPEG images

3P.19 Lossless and semi-lossy compression of CT medical images by content-driven encoding a Laplacian hyper-pyramid
L. Alparone, Universita di Firenze, Firenze, ITALY; S. Baronti and F. Lotti, IROE-CNR, Firenze, ITALY
Statistical models for variable bit rate video
A. van der Kolk, University of Twente, Enschede, NETHERLANDS; J. Balsells, J. Mata, J. Zamora and S. Sallent, E.T.S.E. Telecomunicacion, Barcelona, SPAIN

Session 4I: INVITED LECTURE

Time & Place: 1.30-2.20, Lecture Theatre B
Chairperson: P.M. Grant, Dept. of Electrical Engineering, University of Edinburgh, Mayfield Road, Edinburgh, Scotland EH9 3JL, U.K.

4I.1 Future advances in visual telecommunications
H. Gharavi, Loughborough University of Technology, Loughborough, U.K.

Session 4A: SPECTRAL ESTIMATION

Time & Place: 2.20-5.50, Lecture Theatre A (Oral Session)
Chairperson: T.S. Durrani, Dept. Elec. Science & Telecommunication, University of Strathclyde, 204 George Street, Glasgow G1 1XW, U.K.

4A.1 [Withdrawn]

4A.2 High-resolution bearing estimation by Fourier methods
J. Joutsensalo, Helsinki University of Technology, Espoo, FINLAND

4A.3 A robust and efficient MUSIC-based algorithm for tracking weak and closely spaced sinusoids
F. Geremsky, B. Yang and J.F. Boehme, Ruhr-Universitaet Bochum, Bochum, GERMANY

4A.4 MUSIC estimation of real-valued sinewave frequencies
P. Stoica, A. Eriksson and T. Soderstrom, Uppsala University, Uppsala, SWEDEN

4A.5 Computation of affine Wigner-Ville distributions by means of chirp z-transform
T.P. Zielinski, University of Mining & Metallurgy, Krakow, POLAND

4A.6 Wavelet packets for the estimation of parameters of localized sinusoids
C. vanden Branden Lambrecht and M. Karrakchou, Swiss Federal Institute of Technology, Lausanne, SWITZERLAND

4A.7 Improvements of NMR data quantitation by using the Cadzow procedure prior to any SVBD-based linear prediction method
A. Diop, A. Brigue and D. Graveron-Demilly, Universite Lyon I, Villeurbanne, FRANCE

4A.8 A design method for the dyadic wavelet representation: multiresolution signal decomposition
G. Yang and H. Leich, Faculte Polytechnique de Mons, Mons, BELGIUM
Session 4B: MOTION ESTIMATION

Time & Place: 2.20-5.50, Lecture Theatre B (Oral Session)
Chairperson: J. Biemond, Delft University of Technology, Department of Electrical Engineering, Mekelweg 4, 2626 CJ Delft, NETHERLANDS

4B.1 Adaptive three-dimensional motion-compensated coding of image sequences for multiresolution application
J-P. Leduc, IRISA / INRIA, Rennes, FRANCE

4B.2 Selective coding scheme based on motion vector information
J-B. Lee, S-D. Kim and S-Z. Lee, KAIST, Taejon, KOREA

4B.3 Advanced spatiotemporal propagation strategy for multiresolution motion estimation
L. Boroczky, P. Csillag, KFKI Research Institute for Measurement and Computing Techniques, Budapest, HUNGARY; and K. Fazekas, Technical University of Budapest, Budapest, HUNGARY

4B.4 Estimation of dense 2-D motion based on the constancy of intensity gradient
J. Konrad and P. Treves, INRS Telecommunications, Verdun, Quebec, CANADA

4B.5 Multi-level motion estimation for image sequence coding
H. Nicolas and F. Moscheni, Swiss Federal Institute of Technology, Lausanne, SWITZERLAND

4B.6 A new motion estimation technique for video coding
F. Moscheni, F. Dufaux, I. Moccagatta and H. Nicolas, Swiss Federal Institute of Technology, Lausanne, SWITZERLAND

4B.7 A new block-matching algorithm for estimating multiple image motion vectors
T. Saito and T. Komatsu, Kanagawa University, Yokohama, JAPAN

4B.8 Multiple motion estimation by robust parameter estimation over multiple frames
S. Ayer and P. Schroeter, Swiss Federal Institute of Technology, Lausanne, SWITZERLAND

4B.9 Motion estimation with violation of the constancy brightness constraint
M. Mattavelli and A. Nicoulin, Swiss Federal Institute of Technology, Lausanne, SWITZERLAND

4B.10 Moving segment detection in monocular image sequences under egomotion
W. Kasprzak and H. Niemann, Bavarian Research Center for Knowledge-Based Systems, Erlangen, GERMANY
Session 4C: ADAPTIVE SYSTEMS I

Time & Place: 2.20-5.50, Lecture Theatre C (Oral Session)
Chairperson: O. Macchi, Ecole Superieure d'Electricite, Lab des Signaux et Systemes, CNRS, Plateau du Moulon, 91192 Gif-Sur-Yvette Cedex, FRANCE

4C.1 Subspace-based adaptive algorithms for the blind identification/equalization of multichannel FIR filters

4C.2 Blind channel equalization using 2nd-order cyclostationary statistics
S.V. Schell, Penn State University, Pennsylvania, PA, U.S.A.; W.A. Gardner, University of California, Davis, CA, U.S.A.; and D.L. Smith, Penn State University, Pennsylvania, PA, U.S.A.

4C.3 Adaptive blind equalization and demodulation without channel and signal parameter extraction
J. Sala, G. Vazquez and J. Goldberg, Universitat Politecnica de Catalunya, Barcelona, SPAIN

4C.4 Stability and convergence analysis of the Constant Modulus Algorithm. Comments on finite equalization schemes and the stability of the Normalized CMA

4C.5 A fast adaptive multichannel algorithm and application to decision feedback equalization
S. Theodoridis, G.V. Moustakides and K. Berberidis, University of Patras, Patras, GREECE

4C.6 Genetic algorithms in the continuous space for recursive adaptive filter design
J.-M. Vesin, Swiss Federal Institute of Technology, Lausanne, SWITZERLAND

4C.7 A new adaptive equalizer based on the canonical piecewise linear model
C.J. Pantaleon-Prieto, ETSI Telecommunication, Universidad de Cantabria, Santander, SPAIN; and A.R. Figueiras-Vidal, ETSI de Telecommunication, Universidad Politecnica de Madrid, Madrid, SPAIN

4C.8 The fast subsampled-updating fast transversal filter (FSU FTF) RLS algorithm for adapting long FIR filters
D.T.M. Stock and K. Maouche, Institut EURECOM, Sophia Antipolis, FRANCE

4C.9 A set of algorithms linking NLMS and RLS algorithms
M. Montazeri, CRPE/CNET/CNRS, Issy-les-Moulineaux, FRANCE; and P. Duhamel, ENST/SIG, Paris, FRANCE

4C.10 An a priori error bound for the Steiglitz-McBride method in reduced order cases
P.A. Regalia, Institut National des Telecommunications, Evry, FRANCE; and M. Mboup, Universite Rene Descartes, Paris, FRANCE
Session 4D: ARRAY PROCESSING I: DIRECTION FINDING

Time & Place: 2.20-5.50, Lecture Theatre D (Oral Session)
Chairperson: B. Ottersten, Department of Signals, Sensors and Systems, Royal Institute of Technology, 10044 Stockholm, SWEDEN

4D.1 A cyclic method for signal-selective DOA estimation
G. Gelli, L. Izzo and L. Paura, Universita di Napoli "Federico II", Napoli, ITALY

4D.2 The potential gain in using spectral information in passive localization of sources
H. Messer, Tel Aviv University, Tel Aviv, ISRAEL

4D.3 Analysis of LMS-like source tracking for robust adaptive beamforming

4D.4 A parametric approach for extended source localization
S. Valaee, P. Kabal, McGill University, Montreal, Quebec, CANADA; and B. Champagne, Universite du Quebec, Verdun, Quebec, CANADA

4D.5 Maximum likelihood source separation for discrete sources
A. Belouchrani and J-F. Cardoso, Telecom Paris, Paris, FRANCE

4D.6 Interpolation of distorted towed arrays for source bearing estimation
A. Marsal and S. Marcos, CNRS-ESR, Gif-sur-Yvette, FRANCE

4D.7 On the performance of source separation algorithms
J-F. Cardoso, Telecom Paris, Paris, FRANCE

4D.8 A state space method for direction finding of wide-band emitters
F. Vanpoucke and M. Moonen, Katholieke Universiteit Leuven, Heverlee, BELGIUM

4D.9 A new algorithm for direction of arrival tracking
M. Kosaroglu and Y. Tanik, Middle East Technical University, Ankara, TURKEY

4D.10 Optimally weighted ESPRIT for direction estimation
A. Eriksson, P. Stoica and T. Soderstrom, Uppsala University, Uppsala, SWEDEN

Session 4E: NEURAL NETWORKS I

Time & Place: 2.20-5.50, Lecture Theatre E (Oral Session)
Chairperson: G. Ramponi, Universita degli Studi di Trieste, Dip. Elettrotecnica Elettronica, Via A. Valerio 10, 34127 Trieste, ITALY

4E.1 An architecture of one-class-one-net perceptrons based human face recognition system
P-R. Hong, J-B. Ren, S-C. Tsay and B-C. Chieu, National Taiwan Institute of Technology, Taipei, TAIWAN, R.O.C.

4E.2 Artificial neural networks for attitude determination from visual image processing
J. Seijas, SENER Ingenieria y Sistemas S.A., Madrid, SPAIN; and J.L. Sanz-Gonzalez, ETSI de Telecomunicacion, Universidad Politecnica de Madrid, Madrid, SPAIN

4E.3 A neural approach to lips movements modeling
E. Casella, F. Lavagetto and R. Miani, University of Genoa, Genova, ITALY
A neural network approach for modeling the vocal tract
K. Deergha Rao, Osmania University, Hyderabad, INDIA; and G. Sridhar, Indian Institute of Technology, Bombay, INDIA 804

Using the gamma net in speaker recognition
A. Teixeira and F. Vaz, Universidade de Aveiro, Aveiro, PORTUGAL 808

Speaker identification using networks of radial basis functions

Dimensionality and dynamic temporal structure of a time-delay neural network for speech processing
V. Rodellar, P. Gomez, V. Nieto, Universidad Politecnica de Madrid, Madrid, SPAIN; R. Romera and J. Muruzabal, Universidad Carlos III, Getafe Madrid, SPAIN 816

A neural network based solution for multiple-target radar tracking
V. Schmidlin, G. Favier, University of Nice Sophia Antipolis, Valbonne, FRANCE; and B. Tomasini, Systella Technologies, Carqueiranne, FRANCE 820

Application studies of co-occurrence features in neural network based texture classification
C. Shang and K. Brown, Heriot-Watt University, Edinburgh, U.K. 824

A new structure for noise filtering
S. Munoz-Munoz, Instituto Madrileño de Tecnologia, Madrid, SPAIN; and J.L. Sanz-Gonzalez, ETSI Telecomunicacion, UPM, Madrid, SPAIN 828

Session 4Pa: NONLINEAR DSP (SPECIAL POSTER SESSION)

Time & Place: 2.20-5.50, Poster Room A (Poster Session)

Chairperson: M.D. Macleod, Engineering Department, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.

Optimum genetic algorithms for the design of stack-filters
N. Harvey and S. Marshall, University of Strathclyde, Glasgow, U.K. 832

Genetic optimization of piecewise self-affine fractal interpolation with application to speech modeling
P. Agati and E. Mumolo, Universita di Trieste, Trieste, ITALY 836

An optimization algorithm for fractal encoding of graytone 2D images
A.M. Carminelli Gregori, E. Mumolo and I. Bonat, Universita di Trieste, Trieste, ITALY 840

Trainable hybrid filter structures in particle detector readout systems
S.J. Inkinen, European Laboratory for Particle Physics, CERN, Geneva, SWITZERLAND; and Y. Neuvo, Nokia Group, Helsinki, FINLAND 844

Skeleton redundancy reduction based on a generalization of convexity
R. Kresch and D. Malah, Israel Institute of Technology, Haifa, ISRAEL 848

Improvements to the 'Top Hat' transform, used for analysing pigmented patches on flower petals
S.J. Impey and J.A. Bangham, University of East Anglia, Norwich, U.K. 852
Session 4Pb: BIOMEDICAL PROCESSING

Time & Place: 2.20-5.50, Poster Room B (Poster Session)
Chairperson: J.J. Soraghan, Dept. of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, Scotland G1 1XW,

4Pb.1 Knowledge-based segmentation of medical images
M.J. Carreira, D. Cabello, A. Mosquera, M.G. Penedo and J.M. Pardo, Universidade de Santiago, Santiago de Compostela, SPAIN

4Pb.2 Application of the optimized Canny-Deriche filter for edges detection on cardiac scintigraphic images
M. Girardier, P. Gouton, C. Milan, D. Micollet, Universite de Bourgogne, Dijon, FRANCE; and J.L. Pelletier, Centre de Medicine Nucleaire du Parc, Dijon, FRANCE

4Pb.3 Image sequence analysis for on-line observation of thrombus formation in blood vessels
R. Toenjes, Universitaet Hannover, Hannover, GERMANY

4Pb.4 Pattern recognition in normal and dyskaryotic cervical cell images using statistical texture analysis
F. Hallouche, O.R. Hinton, A.E. Adams, University of Newcastle upon Tyne, Newcastle upon Tyne, U.K.; D.P. Surtees, V. Wadehra, Newcastle General Hospital, Newcastle upon Tyne, U.K.; and G.V. Sherbet, Royal Victoria Hospital, Newcastle upon Tyne, U.K.

4Pb.5 Image analysis tool for embryo qualification based on morphological indices

4Pb.6 Texture analysis of ultrasonographic endoscopy images, using the Master Classifier Method
P. Wilinski, B. Solaiman, C. Roux, E.N.S.T. de Bretagne, Brest, FRANCE; and M. Robaszkiewicz, Universitaire de Brest, Brest, FRANCE

4Pb.7 Transient detection by a time-scale representation applied to biological signals
B. Sankur, E.C. Guler, Y.P. Kahya, Bogazici University, Istanbul, TURKEY; and T. Engin, Technical University of Istanbul, Istanbul, TURKEY
Stochastic model of awake and sleep EEG
A.C. Rosa, A. Fred and J.M. Leitao, Technical University of Lisbon, Lisbon, PORTUGAL

Automatic left ventricular boundary extraction in echocardiographic images using neural networks and snakes
I. Hunter J.J. Soraghan,

Conformation radiotherapy optimisation using distributed simulated annealing
S. Scarth, W.A. Sandham and R. Wilkinson, University of S’athclyde, Glasgow, U.K.

Session 5A: SPEECH CODING II

Time & Place: 9.00-12.30, Lecture Theatre A (Oral Session)
Chairperson: A.D. Fagan, Dept. of Electrical & Electronic Engineering, University College Dublin, Belfield, Dublin 4, EIRE

An enhanced adaptive codebook for a CELP coder
Y. Qian, G. Chahine and P. Kabal, McGill University, Montreal, PQ, CANADA

On improving wideband CELP speech coders
C. McElroy, B. Murray and A.D. Fagan, University College Dublin, Dublin, IRELAND

CELP coding with data rates below 4 kbit/s using phonetically oriented excitations
U. Balss, U. Kipper, H. Reininger and D. Wolf, Johann Wolfgang Goethe-Universitaet Frankfurt am Main, GERMANY

An efficient codebook structure for CELP
B. Murray and A.D. Fagan, University College Dublin, Dublin, IRELAND

Improving CELP quality voice by weighting MSE distance
M.A. Ferrer-Ballester, ETSI Telecomunicacion, Universidad Las Palmas, Las Palmas, SPAIN; and A.R. Figueiras-Vidal, ETSI Telecomunicacion, UPM, Madrid, SPAIN

Combined CELP speech coding and quadrature amplitude modulation
A. Fuldseth, Norwegian Institute of Technology, Trondheim, NORWAY

CELP coder: vector quantization of pitch predictor parameters determined in open-loop configuration
I. Kovalinka, B. Lukovic and M. Markovic, Institute of Applied Mathematics & Electronics, Belgrade, YUGOSLAVIA

Interpolation of autoregressive processes at discontinuities: application to LPC based speech coding
J.S. Erkelens and P.M.T. Broersen, Delft University of Technology, Delft, NETHERLANDS

Efficient quantization of LPC parameters using a mixed LSP and PARCOR representation
C.F. Chan and K.W. Law, City Polytechnic of Hong Kong, Kowloon, HONG KONG

Natural quality parametric speech coding at 2.4 Kbit/s
A. Das and A. Gersh, University of California, Santa Barbara, CA, U.S.A.
Session 5B: NONLINEAR DSP I (SPECIAL SESSION)

Time & Place: 9.00-12.30, Lecture Theatre B (Oral Session)
Chairperson: I. Pitas, Department of Electrical Engineering, Aristotle University of Thessaloniki, Thessaloniki 54006, GREECE

5B.1 Optimal recursive weighted order statistic filters for efficient high speed implementations
L.E. Lucke and R.A. Kroenke, University of Minnesota, Minneapolis, MN, U.S.A.

5B.2 Adaptive skeletonization using multistage Boolean and stack filtering
D. Petrescu, I. Tabus, Polytechnic Institute of Bucharest, Bucharest, ROMANIA; and M. Gabboj, Tampere University of Technology, Tampere, FINLAND

5B.3 Optimal L-filters for vector magnitude filtering
N. Nikolaidis and I. Pitas, University of Thessaloniki, Thessaloniki, GREECE

5B.4 Angular filtering in the context of frequency estimation
T. Loupas, Commonwealth Scientific Industrial and Research Organisation, Chatswood NSW, AUSTRALIA

5B.5 Introducing the fuzzy median filter
F. Russo and G. Ramponi, Universita degli Studi di Trieste, Trieste, ITALY

5B.6 Lack-of-fit detection using the run-distribution test
A.W. Fitzgibbon and R.B. Fisher, Edinburgh University, Edinburgh, U.K.

5B.7 A time and slope domain theory of morphological systems: slope transforms and max-min dynamics
P. Maragos, Georgia Institute of Technology, Atlanta, GA, U.S.A.

5B.8 Chamfer distances in anisotropic 3D images
J.F. Mangin, I. Bloch, J. Lopez-Krahe, Telecom Paris, Paris, FRANCE; and V. Frouin, CEA/DRIPP/SHFI, Orsay, FRANCE

5B.9 Skeleton extraction from noisy binary images using statistical morphological filters and simulated annealing
C.S. Regazzoni and A.N. Venetsanopoulos, University of Genoa, Genova, ITALY

5B.10 The Euclidean distance transform on curved space (EDTDCS) with application to image compression
P.J. Toivanen, Lappeenranta University of Technology, Lappeenranta, FINLAND

Session 5C: COMMUNICATIONS I

Time & Place: 9.00-12.30, Lecture Theatre C (Oral Session)
Chairperson: M.G. Bellanger, C.N.A.M., 292 Rue St. Martin, 75141 Paris Cedex 03, FRANCE

5C.1 A new approach for scrambling speech signals prior to analysis-by-synthesis LPC coding
D.J. Hiotakakos, C.S. Xydeas and C.A. Boyd, University of Manchester, Manchester, U.K.

5C.2 Speech transmission over a channel with a very low SNR
R. Shani, F. Piman and D. Wulich, Ben-Gurion University of the Negev, Beer-Sheva, ISRAEL
5C.3 The Danish candidate for the GSM half-rate speech channel
K.B. Mikkelsen, Tele Danmark Research, Horsholm, DENMARK; K.J. Larsen, Technical University of Denmark, DENMARK; H.B. Hansen, H. Nielsen and H. Jakobsen, Tele Danmark Research, Horsholm, DENMARK

5C.4 Low complexity video subband coding and transmission over ATM
P. Scotton, C. Galand, IBM France, Recherches, La Gaude, FRANCE; and J. Menez, Université de Nice, Parc Valrose, Antibes, Nice, FRANCE

5C.5 DIAMANT: All digital frequency division multiplexing for 10 Gbit/s fibre-optic CATV distribution system
H.G. Goecckler, ANT Nachrichtentechnik GmbH, Backnang, GERMANY

5C.6 Transmultiplexing by recursive (IIR) polyphase structures for onboard processing (OBP) satellites
G.R. Dumasfahani, T.G. Jeans and B.G. Evans, University of Surrey, Guildford, U.K.

5C.7 Real time implementation of the GMDF alpha algorithm on a multi DSP TMS320C40 board
G. Le Tourneur, J.P. Thomas and A. Gilleo, France Telecom, Lannion, FRANCE

5C.8 A maximum likelihood solution to blind identification of multichannel FIR filters
K. Abed-Meraim and E. Moulines, Telecom Paris, Paris, FRANCE

5C.9 A method for classification of analogue modulated radio signals
P.A.J. Nagy, National Defence Research Establishment, Linkoping, SWEDEN

5C.10 Coexistence of asymmetric digital subscriber lines using discrete multitone signalling with TI lines
T. Pollet and M. Moeneclaey, University of Ghent, Gent, BELGIUM

Session 5D: DIGITAL FILTERING I

Time & Place: 9.00-12.30, Lecture Theatre D (Oral Session)

Chairperson: U. Heute, Institute for Network & System Theory, Christian Albrechts University, Kaiserstr 2, 24143 Kiel, GERMANY

5D.1 Design of M-band wavelet filters with maximum regularity
H. Caglar, Marmara Research Center, TUBITAK, Gebze-Kocael, TURKEY; O. Alkin, Southern Illinois University, Edwardsville, IL, U.S.A.; B. Sankur and E. Anarim, Bogazici University, Istanbul, TURKEY

5D.2 Time-domain design of multirate filter banks and wavelets
A. Al-Adnani, R. Chapman and T.S. Durrani, University of Strathclyde, Glasgow, U.K.

5D.3 A new method for efficient convolution in frequency domain by non-uniform partitioning
G.P.M. Egelmeers and P.C.W. Sommen, Eindhoven University of Technology, Eindhoven, NETHERLANDS

5D.4 A spectral factorization approach to 3-channel analysis/synthesis systems
A. Al-Adnani and R. Chapman, University of Strathclyde, Glasgow, U.K.

5D.5 Optimal dyadic filter banks for subband coding and zonal sampling
A. Mertins, Hamburg University of Technology, Hamburg, GERMANY

5D.6 Novel complex FIR kth degree differentiator of variable delay
E. Hermanowicz, Technical University of Gdansk, Gdansk, POLAND
5D.7  A new method for noise cancellation
E. Bataillou and H. Rix, Universite de Nice Sophia Antipolis, Valbonne, FRANCE

5D.8  An example of blind source identification by linear prediction
E. Moisan, CEPHAG - ENSIEG, Saint-Martin d’Heres, FRANCE

5D.9  Discrete multichannel H-Infinity filtering
M.J. Grimble, University of Strathclyde, Glasgow, U.K.

5D.10 An identification method of FIR digital filters in frequency-domain
C. Serviere and V. Capdevielle, CEPHAG-ENSIEG, Saint-Martin d’Heres, FRANCE

Session 5E:  SIGNAL MODELLING

Time & Place:  9.00-12.30, Lecture Theatre E (Oral Session)

Chairperson:  S. Theodoridis, Dept. of Computer Engineering & Informatics, University of Patras, Patras 26500, GREECE

5E.1  Reconstruction from linear attenuating observation models using Markov chain priors
M. Nikolova and A. Mohammad-Djafari, Ecole Superieure d’Electricite, Gif-sur-Yvette, FRANCE

5E.2  Autoregressive with n exogenous inputs (ARX-sup-n) model for evoked potentials analysis: performance analysis
L. Capitiano, G.C. Filligoi, A. Cecchini and S. Cerutti, Universita degli Studi di Roma, Rome, ITALY

5E.3  Autorsimilar processes and Kalman filtering
O. Magre and M. Guglielmi, Ecole Centrale de Nantes, Nantes, FRANCE

5E.4  System theoretical approach to quantum mechanical phenomena
Y. Mondon, S. Uezono and M. Nagamatsu, Shimane University, Matsue, JAPAN

5E.5  Application of genetic algorithms for wavelet networks signal modelling
A. Prochazka and V. Sys, Prague Institute of Chemical Technology, Prague, CZECH REPUBLIC

5E.6  A new spherically invariant joint distribution model for image signals
F. Mueller, Aachen University of Technology, Aachen, GERMANY

5E.7  Optimality conditions for a finite series of Kautz functions derivable from Nth order allpass sections
A.C. den Brinker, Eindhoven University of Technology, Eindhoven, NETHERLANDS

5E.8  Nonlinear channel output behaviour in presence of error correcting code
A. Benhallam, Ecole Nationale de l’Aviation Civile, Toulouse, FRANCE

5E.9  Some weighted objective approaches for sparse deconvolution
I. Santamaria-Caballerio, ETSI de Telecomunicacion, Universidad de Cantabria, Santander, SPAIN; and A.R. Figueiras-Vidal, ETSI de Telecomunicacion, Universidad Politecnica de Madrid, Madrid, SPAIN

5E.10  Isomorphism between continuous-time system and discrete-time system based on differentiability of the input signal
T. Horiuchi and K. Toraichi, University of Tsukuba, Ibaraki, JAPAN
SP.1 M-Sequence detection by higher order signal correlation techniques
E.S. Warner, B. Mulgrew and P.M. Grant, University of Edinburgh, Edinburgh, U.K.

SP.2 Blind estimation of ARMA systems
D. Boss and K.D. Kammeyer, Technische Universitat Hamburg-Harburg, Hamburg, GERMANY

SP.3 Time delay / scale estimation using waveland theory - higher order wavelets
M. Devin, National Engineering Laboratory, East Kilbride, Glasgow, U.K.; and T.S. Durran, University of Strathclyde, Glasgow, U.K.

SP.4 Quadratic identification using bispectrum
O. Michel, ENS Lyon, Lyon, FRANCE; and P. Ruiz, ESM Marseille, Puy-Ste-Reparde, FRANCE

SP.5 A multi-time definition of the Wigner-Ville distribution; L-Wigner distribution
L. Stanovic and S. Stanovic, Elektroteknisk Fakultet, Montenegro, YUGOSLAVIA

SP.6 Cost functions for motion estimation based on HOS
E. Sayrol, A. Gasull and J.A.R. Fonollosa, ETSETB, Universitat Politecnica de Catalunya, Barcelona, SPAIN

SP.7 Bispectral analysis of periodic signals in noise: theory, interpretation and condition monitoring applications

SP.8 MIMO Volterra system input/output relations for cyclic higher-order statistics
L. Izzo, A. Napolitano and L. Paura, Universita di Napoli "Federico II", Napoli, ITALY

SP.9 Ventricle late potential detection from bispectral analysis of ST-segments
C.A. Speirs, J.J. Soraghan and R.W. Stewart, University of Strathclyde, Glasgow, U.K.

SP.10 Identification of nonlinearities in vowel generation

SP.11 A new adaptive algorithm for AR parameters estimation using cumulants
A. Meziane, A. Adib and D. Aboutajdine, Universite Mohamed V, Rabat, MOROCCO

SP.12 Time delay estimation using a single cumulant
C-C. Feng and C-Y. Chi, National Tsing Hua University, Hsinchu, TAIWAN, R.O.C.

SP.13 Horizontal-slice cumulant matrices for parameter estimation of exponentially damped sinusoids
D.P. Ruiz, M.C. Carrion, A. Gallego and A. Medouri, Universidad de Granada, Granada, SPAIN

SP.14 Performances analysis of non Gaussian linear AR parameters detection using the bispectrum
WITHDRAWN
A. Ferrari, G. Alengrin and C. Tenenbaum, Universite de Nice Sophia-Antipolis, Nice, FRANCE

SP.15 Performance analysis of non Gaussian linear AR parameters detection using the bispectrum
WITHDRAWN
A. Ferrari, G. Alengrin and C. Tenenbaum, Universite de Nice Sophia-Antipolis, Nice, FRANCE

xxxii
Identification of a quadratic system using only output cumulants
P. Bondon and M. Krob, Ecole Superieure d'Electricite, Gif-sur-Yvette, FRANCE

Blind separation of sources: a comparative study of HOS-based and decorrelation-based solutions

Complex self-adaptive algorithm for source separation based on high-order contrast
E. Moreau and O. Macchi, Ecole Superieure d'Electricite, Gif-sur-Yvette, FRANCE

A general adaptive algorithm for non Gaussian source separation without any constraint
F. Harroy, J-L. Lacoume, CEPHAG - ENSIEG, Saint-Martin d'Heres, FRANCE; and M.A. Lagunas, ETSETB, Universitat Politecnica de Catalunya, Barcelona, SPAIN

Non-linearity on the control error for adaptive recursive filters
S. Cherif, Ecole Superieure des Postes et Telecommunications, Tunis, TUNISIA; C. Vignat, S. Marcos, Ecole Superieure d'Electricite, Gif-sur-Yvette, FRANCE; and M. Jaidane, Ecole Nationale d'Ingenieurs de Tunis, Tunis, TUNISIA

Session 6I: INVITED LECTURE

Time & Place: 1.30-2.20, Lecture Theatre A
Chairperson: B. Mulgrew, Dept. of Electrical Engineering, University of Edinburgh, Mayfield Road, Edinburgh, Scotland EH9 3JL, U.K.

Variable-rate speech coding
A. Gersho, University of California, Santa Barbara, CA, U.S.A.

Session 6A: SPEECH ENHANCEMENT

Time & Place: 2.20-5.50, Lecture Theatre A (Oral Session)
Chairperson: P. Noll, Technische Universitaet Berlin, Institut fur Fernmeldetechnik, Einsteinufer 25, 10587 Berlin, GERMANY

A signal subspace approach for noise reduction of speech signals
S.H. Jensen, Katholieke Universiteit Leuven, Heverlee, BELGIUM; P.C. Hansen, S.D. Hansen and J.A. Sorensen, Technical University of Denmark, Lyngby, DENMARK

Bandwidth enhancement of narrow-band speech signals
H. Carl, Grundig Electronics, Fuerth, GERMANY; and U. Heute, Christian Albrechts University, Kiel, GERMANY

Speech enhancement based on minimum statistics
R. Martin, Aachen University of Technology, Aachen, GERMANY

Speech enhancement algorithms using MMSE estimation with interframe constraint
Z. Cao, B. Tian and X. Wang, Tsinghua University, Beijing, CHINA
6A.5 Speech enhancement based on an auditory model
F. Mekuria and T. Fjallbrant, University of Linkoping, Linkoping, SWEDEN

6A.6 Some robust speech enhancement techniques using higher order AR estimation
J.M. Salavedra, ETSETB, Universitat Politiecnica de Catalunya, Barcelona, SPAIN; E. Masgrau, University of Zaragoza, Zaragoza, SPAIN; and A. Moreno, ETSETB, Universitat Politiecnica de Catalunya, Barcelona, SPAIN

6A.7 Channel noise detection and suppression in G721 decoded speech for CAI CT2 applications
D.J. Hiotakakos, C.S. Xycides, M.A. Ireton, University of Manchester, Manchester, U.K.; and S.M. Asghar, Advanced Micro Devices, Austin, Texas, U.S.A.

6A.8 An environment-adaptive noise reduction neural network for reliable speech recognition
M. Trompf and H. Eckhardt, Alcatel SEL Research Center, Stuttgart, GERMANY

6A.9 Speech enhancement using sub-band decomposition and comparison with full-band techniques
G. Faucon, R. Le Bouquin and A. Akbari Azirani, Universite de Rennes I, Rennes, FRANCE

6A.10 Adaptive speech enhancement with diverse sub-band processing
D.R. Campbell, University of Paisley, Paisley, Scotland, U.K.

Session 6B: SUBBAND IMAGE CODING

Time & Place: 2.20-5.50, Lecture Theatre B (Oral Session)

Chairperson: H. Gharavi, Dept. of Electronic & Electrical Engineering, Loughborough University of Technology, Ashby Road, Loughborough, Leics. LE11 3TU, U.K.

6B.1 Very low bit-rate colour video coding scheme using sub-band motion estimation/compensation
K.H. Goh, J.J. Soraghan and T.S. Durrani, University of Strathclyde, Glasgow, U.K.

6B.2 Subband coding of images using hierarchical quantization
M. Domanski and R. Swierczynski, Politechnika Poznanska, Poznan, POLAND

6B.3 Linear phase wavelet transforms for low bitrate image coding
E.A.B. da Silva and M. Ghanbari, University of Essex, Colchester, U.K.

6B.4 Adaptive subband VQ for image coding
S.P. Voukelatos, J.J. Soraghan and T.S. Durrani, University of Strathclyde, Glasgow, U.K.

6B.5 Image subband coding with adaptive IIR filter banks - Automatic filter selection
R. Sorhus and J.H. Husoy, Rogaland University Center, Stavanger, NORWAY

6B.6 A region-based discrete wavelet transform
H.J. Barnard, J.H. Weber and J. Biemond, Delft University of Technology, Delft, NETHERLANDS

6B.7 A study on FIR filters for subband coding of images
H. Benoit-Cattin, A. Baskurt, F. Peyrin and R. Goutte, INSA-Lyon, Villeurbanne, FRANCE
6B.8 Design of recursive filter banks for subband coding of images
M. Domanski, Politechnika Poznanska, Poznan, POLAND

6B.9 On the choice of multirate filter banks for image coding
P. Onno and C. Guillemot, CCETT, Cesson-Sevigne, FRANCE

6B.10 Image coding by block prediction of wavelet coefficients
R. Rinaldo and G. Calvagno, Universita degli Studi di Padova, Padova, ITALY

Session 6C: ADAPTIVE SYSTEMS II

Time & Place: 2.20-5.50, Lecture Theatre C (Oral Session)


6C.1 A fast time-series adaptive filtering algorithm based on the QRD inverse-updates method

6C.2 Fast anti-lattice algorithm

6C.3 The performance of eight recursive least squares adaptive filtering algorithms in a limited precision environment

6C.4 Fast SCAR algorithm for least squares adaptive filtering
T. Schneider, Technische Hochschule Darmstadt, Darmstadt, GERMANY

6C.5 Stability guaranteed adaptive algorithm and its application to active noise control
Y. Shinohara, M. Kajiki, H. Ohmori and A. Sano, Keio University, Yokohama, JAPAN

6C.6 Inverse factorization type algorithms for H-infinity adaptive filtering
H. Sakai, Kyoto University, Kyoto, JAPAN

6C.7 The application of servo theory to the LMS algorithm
T.J. Moir, University of Paisley, Paisley, Scotland, U.K.

6C.8 An adaptive subspace filter for noise reduction
G. Doblinger, Vienna University of Technology, Vienna, AUSTRIA

6C.9 Block adaptive identification with predictive structure: an application to acoustic echo cancellation

6C.10 Recursive algorithm for elimination of measurement noise and impulsive disturbances from ARMA signals
M. Niedzwiecki, Technical University of Gdansk, Gdansk, POLAND
Session 6D: ARRAY PROCESSING II

Time & Place: 2.20-5.50, Lecture Theatre D (Oral Session)
Chairperson: J.P. Lecadre, IRISA, Campus de Beaulieu, 35042 Rennes Cedex, FRANCE

6D.1 Robust self-calibration of the maximum likelihood method in array processing
A. Flieeller, P. Larzabal, L.E.Si.R., E.N.S., Cachan, FRANCE; and H. Clergeot, L.E.Si.R., Cayenne, FRANCE

6D.2 Interference rejection range improvement for linear arrays with a randomized sensor positioning scheme
J.M. Paez-Borrallo, J. Orea and M. Bermejo, ETSI Telecomunicacion, UPM, Madrid, SPAIN

6D.3 A quiescent pattern control strategy for adaptive arrays
P.G. Richardson, Defence Research Agency, Malvern, U.K.

6D.4 Novel architecture for studying complex behaviour in non-linear processing systems
C.T. Pointon, R.A. Carrasco, Staffordshire University, Stafford, U.K.; and M. Gell, B.T. Laboratories, Ipswich, U.K.

6D.5 Adaptive instrumental-variable method for robust direction-of-arrival estimation
P. Stoica, M. Cedervall and T. Soderstrom, Uppsala University, Uppsala, SWEDEN

6D.6 An adaptive subspace algorithm based on QR decomposition and on perturbation analysis
J.P. Delmas, Institut National des Telecommunications, Evry, FRANCE

6D.7 An extension of the constant-modulus array for periodic power signals
J.R. Cerqueires and J.A. Fernandez-Rubio, ETSETB, Universitat Politecnica de Catalunya, Barcelona, SPAIN

6D.8 Robust direction finding in the presence of spatially correlated noise
B. Goransson, Royal Institute of Technology, Stockholm, SWEDEN

6D.9 Simultaneous multiuser demodulation based on digital array processing
A. Perez-Neira and M.A. Lagunas, ETSETB, Universitat Politecnica de Catalunya, Barcelona, SPAIN

6D.10 The singular value decomposition and its application in the presence of correlated arrivals
M. Elmarazay, 16 Dalton Street, Bondi, N.S.W., AUSTRALIA WITHDRAWN

Session 6E: DSP HARDWARE

Time & Place: 2.20-5.50, Lecture Theatre E (Oral Session)

6E.1 Implementation of complex DSP systems using high-level design tools
M. Freericks, A. Fauth, TU Berlin, Berlin, GERMANY; and A. Knoll, Universitaet Bielefeld, Bielefeld, GERMANY

6E.2 Synthesis of a dedicated signal processor via algorithmic/hardware trade-offs
M. Auguin, P. Balestra, F. Boeri, C. Carriere and A. Giuliani, Universite de Nice Sophia-Antipolis, Nice, FRANCE

xxxvi
6E.3 Fast adaptive algorithms implementation: algorithm/DSP-architecture interaction
M. Ouaou, Universite Mohamed V, Rabat, MOROCCO; P. Baylou, M. Najim, ENSERB, Talence, FRANCE; and M. Zyoute, Universite Mohamed V, Rabat, MOROCCO

6E.4 A VLSI mixed signal processing system
A. Alvarez and A.B. Premkumar, Nanyang Technological University, SINGAPORE

6E.5 Implementation of digital filters on reconfigurable field-programmable gate arrays
H. Leich and J. Hancq, Faculte Polytechnique de Mons, Mons, BELGIUM

6E.6 Hardware-efficient implementation of oversampled time-reversed linear phase IIR filters
Q. Huang, Swiss Federal Institute of Technology, Zurich, SWITZERLAND; and P.T. Maguire, University of East Anglia, Norwich, U.K.

6E.7 A high speed bit stream generator for HDTV
Y-G. Park, C. Sohn and B-U. Lee, DAEWOO Electronics, Seoul, KOREA

6E.8 Very high speed edge detector chip using the optimised Canny-Deriche filter
E. Bourennane, C. Milan, M. Paindavoine, Universite de Bourgogne, Dijon, FRANCE; and M. Robert, Laboratoire LIRMM, Montpellier, FRANCE

6E.9 Systolic arrays for modified covariance spectral estimation used with ultrasonic doppler blood flow detectors
S.J. Bellis, University College North Wales, Bangor, Wales, U.K.; W.P. Marnane, University College Cork, Cork, IRELAND; D. Wilde, IRISA, Rennes, FRANCE; and P.J. Fish, University College North Wales, Bangor, Wales, U.K.

6E.10 High level synthesis of a non-linear distortion corrector for PWM DACs
M. Sandler, A. Paul and H. Malik, King’s College London, London, U.K.

Session 6P: PATTERN RECOGNITION

Time & Place: 2.20-5.50, Poster Rooms (Poster Session)

6P.1 A visual surveillance system for autonomous vehicle risk avoidance
G.L. Foresti, P. Matteucci, C.S. Regazzoni and S. Spaggiari, University of Genoa, Genova, ITALY

6P.2 Pose estimation based on symmetry of 3-D object
W. Wen and B-Z. Yuan, Northern Jiaotong University, Beijing, CHINA

6P.3 Noise robust and rotation invariant texture classification
T.N. Tan, University of Reading, Reading, U.K.

6P.4 Wavelet-based edge detection and classification
J.R. Beltran and J. Navarro, Universidad de Zaragoza, Zaragoza, SPAIN

6P.5 Recognizing partially occluded colored objects
R. Schettini, IFCTR-ITIM, CNR, Milano, ITALY

6P.6 Fractional Brownian motion: a model for image texture
R. Jennane and R. Harba, Universite d’Orleans, Orleans, FRANCE

xxxvii
Finding particles in a polymeric matrix by means of Hough transform
V. Ballarin and E. Moler, Facultad de Ingenieria, Mar del Plata, ARGENTINA

Measurement of crowd density using image processing
J.H. Yin, S.A. Velastin and A.C. Davies, King's College London, London, U.K.

Range imaging via spatially encoded coloured spots
C.J. Davies and M.S. Nixon, University of Southampton, Southampton, U.K.

Image sequence processing to supervise fish passes
N. Castignolles, M. Cattoen, INP-ENSEEIHT, Toulouse, FRANCE; and M. Larinier, CSP-CEMAGREF, FRANCE

Time variant system identification for car engine signal analysis
M. Wagner, Ruhr-Universitaet Bochum, GERMANY; E. Karlsson, Uppsala University, Uppsala, SWEDEN; D. Koenig and C. Toerk, Ruhr-Universitaet Bochum, Bochum, GERMANY

A model based detection scheme for non synchronous multidimensional transient
P. Beauséjour and P. Gaillard, Universite de Technologie de Compiègne Troyes, Troyes, FRANCE

Optimal estimation of fractal models of digital sequences by means of genetic algorithms
P. Agati and E. Mumolo, Universita' di Trieste, Trieste, ITALY

On two-point resolution in partially coherent light: a parameter estimation approach
A.J. den Dekker, Delft University of Technology, Delft, NETHERLANDS

Pattern recognition technics to improve knock detection in spark ignition engines
F. Molinaro, Universite de la Reunion, Saint-Denis, FRANCE; and F. Castanie, ENSEEIHT-GAPSE, Toulouse, FRANCE

Classification of non-stationary sinusoidal signals
E. Le Carpentier, I. Vincent, Ecole Centrale de Nantes, Nantes, FRANCE; F. Auger, IUT de Saint-Nazaire, Saint-Nazaire, FRANCE; and C. Doncarli, Ecole Centrale de Nantes, Nantes, FRANCE

On-line handwriting recognition using supervised hidden Markov models
J.R. Bellegarda, D. Nahamoo, K.S. Nathan and E.J. Bellegarda, IBM Research, Yorktown Heights, NY, U.S.A.

Unconstrained letter-segmentation of off-line cursive script using contour information

A function to express hand-drawn letters
M. Kamada, Ibaraki University, Ibaraki, JAPAN; K. Toraiichi, University of Tsukuba, Ibaraki, JAPAN; C.Y. Huang and W.E. Pan, Ibaraki University, Ibaraki, JAPAN

A technique CLEAN for detecting multiple straight lines in a binary image using discrete Hough transform
R.C. Agrawal and R.K. Shevgaonkar, Indian Institute of Technology, Bombay, INDIA
Session 7A: DIGITAL AUDIO

Time & Place: 9.00-12.30, Lecture Theatre A (Oral Session)

Chairperson: B. Mulgrew, Dept. of Electrical Engineering, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JL, U.K.

7A.1 Subband ADPCM coding for wideband audio signals using analysis-by-synthesis quantization scheme
M.T. Chu and C.F. Chan, City Polytechnic of Hong Kong, Kowloon, HONG KONG
1449

7A.2 Dynamic bit allocation in subband coding of wideband audio with multipulse LPC
P. Menardi, G.A. Mian and G. Riccardi, University of Padova, Padova, ITALY
1453

7A.3 A global theoretical auditory model for application on audio coders design and objective perceptual assessment
A.S. Pena, Universidad de Vigo, Vigo, SPAIN
1457

7A.4 Hi-fi audio CODEC employing variable frame bit allocation
1461

7A.5 A method of noise reduction with FIR dynamic filtering
1465

7A.6 Modeling and compensation of nonlinear distortion in horn loudspeakers
H. Schurer, A.P. Berkhoff, C.H. Slump and O.E. Herrmann, University of Twente, Enschede, NETHERLANDS
1468

7A.7 Consumer audio DSP applications using Motorola's DSP56004
P. Atherton, Motorola Semiconductors, East Kilbride, Glasgow, U.K.
1472

7A.8 A digital signal processing audiological workstation
D. Sweeney, R.W. Stewart and E. Pirie, University of Strathclyde, Glasgow, U.K.
1478

Session 7B: NONLINEAR DSP II (SPECIAL SESSION)

Time & Place: 9.00-12.30, Lecture Theatre B (Oral Session)

Chairperson: G.L. Sicuranza, Dept. Elettrotecnica Elettronica Infor., University degli Studi di Trieste, Via A. Valerio 10, 34127 Trieste, ITALY

7B.1 A stability condition for time-varying bilinear systems and its applications in adaptive filtering
J. Lee, Industrial Technology Research Institute, Hsinchu, TAIWAN, R.O.C.; and V.J. Mathews, University of Utah, Salt Lake City, Utah, U.S.A.
1482

7B.2 Unsharp masking with nonlinear filters
T-H. Yu, Chinese University of Hong Kong, Shatin - NT, HONG KONG; and S.K. Mitra, University of California, Santa Barbara, CA, U.S.A.
1485

7B.3 Optimal order for nonlinear prediction
A. Poncet and G.S. Moschytz, Swiss Federal Institute of Technology, Zurich, SWITZERLAND
WITHDRAWN

xxxix
7B.4 An adaptive invariant transform using neural network techniques
S. Kroener, Technische Universität Hamburg-Harburg, Hamburg, GERMANY; R.
Moratz, Universität Bielefeld, Bielefeld, GERMANY; and H. Burkhardt, Technische
Universität Hamburg-Harburg, Hamburg, GERMANY

7B.5 Order estimation and nonlinear prediction with radial basis functions
C. Ris, Faculte Polytechnique de Mons, Mons, BELGIUM; H. Dedieu and M. Hasler,
Swiss Federal Institute of Technology, Lausanne, SWITZERLAND

7B.6 Marginal median learning vector quantizer
C. Kotropoulos, I. Pitas, University of Thessaloniki, Thessaloniki, GREECE; and M.
Gabbouj, Tampere University of Technology, Tampere, FINLAND

7B.7 Nonlinear filters for noise reduction
W. Knecht, M. Schenkel and G.S. Moschytz, Swiss Federal Institute of Technology,
Zurich, SWITZERLAND

7B.8 Smoothing 2-D or 3-D images using local classification
K. Haris, G. Tziritas and S. Orphanoudakis, University of Crete, Heraklion, Crete,
GREECE

7B.9 A robust L-estimator for filtering quantum-limited image sequences
R.P. Kleihorst, R.L. Lagendijk and J. Biemond, Delft University of Technology, Delft,
NETHERLANDS

7B.10 Nonlinear image restoration in coarse
J.B.T. Roerdink, University of Groningen, NETHERLANDS

Session 7C: COMMUNICATIONS II

Time & Place: 9.00-12.30, Lecture Theatre C (Oral Session)
Chairperson: F. Westall, BT Laboratories, Martlesham Heath, Ipswich, Suffolk IPS 7RE, U.K.

7C.1 State space modeling of time-varying multipath channels
M. Haardt and R. Pauli, Technical University of Munich, Munich, GERMANY

7C.2 MLSE antenna diversity equalization of a jammed frequency-selective fading
channel
P. Vila, F. Pipo, D. Peres, Thomson-CSF/RGS, Gennevilliers, FRANCE; and L. Fety,
CNAM, Paris, FRANCE

7C.3 Analysis of a globally convex algorithm for blind equalization of non minimum
phase channels
S. Zazo, J.M. Paez-Borrillo and I.A. Perez-Alvarez, ETSI de Telecomunicacion (UPM),
Madrid, SPAIN

7C.4 A comparison of six different non-linear equalisation techniques for digital
communications systems
C.P. Callender and C.F.N. Cowan, Loughborough University of Technology,
Loughborough, U.K.

7C.5 On the equalisation of the baseband LNL channel
N. Beamish and A.D. Fagan, University College Dublin, Dublin, IRELAND

Page No.
1489
1492
1496
1500
1504
1508
WITHDRAWN
1528
7C.6 Low complexity cancellation of non-linear impairments to data communications channels
C.P. Callender, C.F.N. Cowan, Loughborough University of Technology, Loughborough, U.K.; and S. Theodoridis, University of Patras, Patras, GREECE

7C.7 Adaptive arrays for non-selective and selective frequency channels with co-channel interference
O. Munoz and J. Fernandez, ETSE Telecommunicacion (UPC), Barcelona, SPAIN

7C.8 Performance of a stochastic gradient adaptive beamformer for sub-sea acoustic communications
O.R. Hinton, G.S. Howe and A.E. Adams, University of Newcastle upon Tyne, Newcastle upon Tyne, U.K.

7C.9 A diversity demodulation technique for mobile communications
A. Abrado, G. Benelli, A. Bini and A. Garzelli, Universita di Firenze, Firenze, ITALY

7C.10 The effect of a nonlinear amplifier on the bit error rate performance of OFDM and single carrier signals
T. Pollet, M. Van Bladel and M. Moeneclaey, University of Ghent, Gent, BELGIUM

Session 7D: DIGITAL FILTERING II

Time & Place: 9.00-12.30, Lecture Theatre D (Oral Session)

Chairperson: A. Fettweis, Ruhr Universitaet Bochum, Lehrst. f. Nachrichtentechnik, Universitaetsstr 150, 44780 Bochum, GERMANY

7D.1 The convergence rate of symmetric weighted median filters
H. Chen, Jilin University of Technology, Changchun, CHINA; R. Yang and M. Gabbouj, Tampere University of Technology, Tampere, FINLAND

7D.2 Fast approximation of the euclidean norm: application to vector median filtering
M. Barni, F. Bartolini and V. Cappellini, Universita di Firenze, Firenze, ITALY

7D.3 Fast algorithms for analyzing and designing weighted median filters
R. Yang, M. Gabbouj and Y. Neuvo, Tampere University of Technology, Tampere, FINLAND

7D.4 The KFilter: a new model of nonlinear systems with memory
A. Pages-Zamora and M.A. Lagunas, ETSETB, Universitat Politeneica de Catalunya, Barcelona, SPAIN

7D.5 Non-linear recursive smoothing filters in 1-D and 2-D
M. Macleod, Cambridge University, Cambridge, U.K.

7D.6 Time-varying filter banks for the analysis of pc processes
J-C. Pesquet, Ecole Superieure d'Electricite, Gif-sur-Yvette, FRANCE; and H. Krim, LIDS, M.I.T., Boston, MA, U.S.A.

7D.7 Convergence analysis of processing cost reduction method of NLMS algorithm partly skipping weight vector components adaptation
K. Takahashi and S. Mori, Keio University, Yokohama, JAPAN

7D.8 A fast Fermat number transform for long sequences
L-I. Alfredsson, Linkoping University, Linkoping, SWEDEN
7D.9  High order transformations for flexible IIR filter design

7D.10 Empiric and theoretical roundoff noise formulas for digital filter realizations
J.L. Sanz-Gonzalez and D. Andina, ETSI Telecomunicacion-UPM, Madrid, SPAIN

Session 7E: DSP SOFTWARE

Time & Place: 9.00-12.30, Lecture Theatre E (Oral Session)

Chairperson: P. Duhamel, Departement Signal, E.N.S.T., 46 Rue Barrault, 75634 Paris Cedex 13, FRANCE

7E.1 Towards a European reference package for practical signal processing
C. Capdessus, O. Labordette, J. Thiel and E. Jolivet, Universite d'Orleans, Orleans, FRANCE

7E.2 An applicative DSP development environment
D. Freeman, British Telecom Laboratories, Ipswich, U.K.

7E.3 Optimizing DSP programs using the multirate retiming transformation
V. Zivojnovic, S. Ritz and H. Meyr, Aachen University of Technology, Aachen, GERMANY

7E.4 Direct software bridge Matlab-transputer boards
J. Kadlec, Institute of Information Theory and Automation, Prague, CZECH REPUBLIC

7E.5 A multiprocessor algorithm scheduler for signal and image processing
M. Razaz and K.A. Marlow, University of East Anglia, Norwich, U.K.

7E.6 A transputer implementation of pattern matching by invariant moments
G. Hall and F. Dechamp, University of Central Lancashire, Preston, U.K.

7E.7 A software structure for real-time parallel image processing
E. Rendon, L. Salgado, J.M. Menendez and N. Garcia, E.T.S.I. de Telecomunicacion (UPM), Madrid, SPAIN

7E.8 A real-time implementation of the CCITT G.728 16kb/s LD-CELP fixed-point algorithm on the Motorola DSP56156
M. Murphy, Motorola Ltd, East Kilbride, Glasgow, U.K.; and C. Cox, Signals and Software Ltd., Harrow, U.K.

7E.9 A fast 1D sieve transform for multiscale signal decomposition
J.A. Bangham, S.J. Imprey and F.W.D. Woodhams, University of East Anglia, Norwich, U.K.

7E.10 Efficient implementation of FFT-like algorithms on MIMD systems
N. Jungclaus and M. Noelle, Technische Universitat Hamburg-Harburg, Hamburg, GERMANY
Session 7P: SPEECH PROCESSING

Time & Place: 9.00-12.30, Poster Rooms (Poster Session)

Chairperson: D. Van Compernolle, ESAT Laboratory, K.U. Leuven, K. Mercierlaan 94, 3001 Heverlee, BELGIUM

7P.1 Using MLPS as probability generators vs. as labelers: a comparative study
P. Le Cerf and D. Van Compernolle, K.U. Leuven, Heverlee, BELGIUM

7P.2 On the separability of speech signals
J. Ming, A.D. Irvine and F.J. Smith, Queen’s University of Belfast, Belfast, Northern Ireland, U.K.

7P.3 Fuzzy smoothing of HMM parameters using Parzen’s window with application to speech recognition
J. Dai, University of Nanjing, Nanjing, CHINA

7P.4 Large vocabulary Mandarin tone recognition by TDNN technique
G-S. Poo, National University of Singapore, Kent Ridge, SINGAPORE

7P.5 A study of the effect of pitch on LPC spectral matching metrics
J. Crestel and M. Guitton, ENSAT/LASTI, Lannion, FRANCE

7P.6 The noise robustness of auditory front-ends in HMM based speech recognisers
I.R. Gransden and S.W. Beet, University of Sheffield, Sheffield, U.K.

7P.7 Non-stationary prediction of frame-based speech data
S.W. Beet, L. Baghai-Ravary and M.O. Tokhi, University of Sheffield, Sheffield, U.K.

7P.8 Application of singularity detection with wavelets for pitch estimation of speech signals
N. Gonzalez and D. Docampo, Universidad de Vigo, Vigo, SPAIN

7P.9 Study of a VLSI implementation of a noise reduction algorithm for digital hearing aids
S. Grassi, A. Heubi, M. Ansorge and F. Pellandini, University of Neuchatel, Neuchatel-Serrières, SWITZERLAND

7P.10 A general waveform interpolation structure for speech coding
W.B. Kleijn and J. Haagen, A T & T Bell Laboratories, Murray Hill, NJ, U.S.A.

7P.11 Applications of speech processing using an AM-FM modulation model and energy operators
A. Potamianos and P. Maragos, Georgia Institute of Technology, Atlanta, Georgia, U.S.A.

7P.12 Double frequency and time-frequency analyses of cyclostationary speech sounds
B. Ravera and C. d’Alessandro, LIMSI-CNRS, Orsay, FRANCE

7P.13 A novel DyWTVT approach for continuous speech pitch estimation
F.J. Ainsic, B.L. Burrows and R.A. Carrasco, Staffordshire University, Stafford, U.K.

7P.14 Two pass robust pitch extraction algorithm using the Dyadic wavelet transform
E. Lukasik and S. Grochowelewski, Technical University of Poznan, Poznan, POLAND

7P.15 An algorithm for the estimation of glottal closure instants using the sequential detection of abrupt changes in speech signals
C. Murgia, I. Mann and G. Feng, Universite Stendhal, Grenoble, FRANCE

7P.16 The variation of the lip radiation impedance in a reverberant enclosure
C.J. Bleakley and R. Scaife, Dublin City University, Dublin, IRELAND

Page No. 1629

Page No. 1633

Page No. 1637

Page No. 1641

Page No. 1645

Page No. 1649

Page No. 1653

Page No. 1657

Page No. 1661

Page No. 1665

Page No. 1669

Page No. 1673

Page No. 1677

Page No. 1681

Page No. 1685

Page No. 1689
7P.17 New speech separation algorithms
E. Mumolo and M. Romanin, Universita' di Trieste, Trieste, ITALY

7P.18 New speech processing algorithms based on the sinusoidal model
E. Mumolo and L. Salvi, Universita' di Trieste, Trieste, ITALY

7P.19 A neural network based adaptive noise reduction filter for speech recognition systems
K. Kasper, H. Reininger and D. Wolf, Johann Wolfgang Goethe-Universitaet, Frankfurt am Main, GERMANY

7P.20 Noise classification using vector quantization
N. Nicol, M. Falkhausen, H. Reininger, D. Wolf, Johann Wolfgang Goethe-Universitaet, Frankfurt am Main, GERMANY; S. Euler and J. Zinke, Telenorma Bosch-Telekom, Frankfurt, GERMANY

Session 8Ia: INVITED LECTURE

Time & Place: 1.30-2.20, Lecture Theatre A
Chairperson: E. Chapel, Avonbank Cottage, Blackstone Road, Avonbridge, Falkirk, Scotland FK1 2LB, U.K.

8Ia.1 Engineering algorithms for parallel VLSI implementation
E.F. Deprettere, Delft University of Technology, Delft, NETHERLANDS

Session 8Ib: INVITED LECTURE

Time & Place: 1.30-2.20, Lecture Theatre B
Chairperson: R.J. Clarke, Dept. of Computing and Electrical Engineering, Heriot Watt University, Riccarion, Edinburgh, Scotland EH14 4AS, U.K.

8Ib.1 ITU-T and ISO video compression standards: where we are now and how we got there
G. Morrison, British Telecom Laboratories, Ipswich, U.K.
Session 8A: ADAPTIVE SIGNAL PROCESSING

Time & Place: 2.20-5.50, Lecture Theatre A (Oral Session)
Chairperson: C.R. South, British Telecom Laboratories, Martlesham Heath, Ipswich, Suffolk IP5 7RE, U.K.

8A.1 Different approaches for a high-resolution narrow-band spectrum
A.N. Hossen, Ruhr-University, Bochum, GERMANY; and U. Heute, Christian Albrechts University, Kiel, GERMANY

8A.2 Performance analysis of two algorithms for tracking of multiple noisy cisoids
P. Tchavsky, Institute of Information Theory & Automation, Prague, CZECH REPUBLIC; and P. Handel, Ericsson Radio System AB, Stockholm, SWEDEN

8A.3 Behaviour of cascade resonator-in-a-loop adaptive filters for tracking multiple sinusoids

8A.4 Autoregressive estimation on signals presenting abrupt changes
T. Robert and C. Mailhes, ENSEEIHT/GAPSE, Toulouse, FRANCE

8A.5 Spectral estimation and performance optimisation for the cyclostationary PPM process in the presence of jitter
J.M.H. Elmirghani and R.A. Cryan, Manchester Metropolitan University, Manchester, U.K.

8A.6 High accuracy spectral estimation using the Legendre nonuniform discrete Fourier transform
V.E. Neagoe, Polytechnical University of Bucharest, Bucharest, ROMANIA

8A.7 Estimation of the optimal convergence factor for acoustic echo cancellation in a noisy environment
I. Marx, Technische Hoschule Darmstadt, Darmstadt, GERMANY

8A.8 Predictors based on interpolated FIR filters
O. Vainio, Tampere University of Technology, Tampere, FINLAND

8A.9 Acoustic echo control combined with two orthogonalizing techniques
C. Antweiler and A. Schmitz, Aachen University of Technology, Aachen, GERMANY

8A.10 An optimal microphone array for speech reception in a car
J. Rex and S.J. Elliott, University of Southampton, Southampton, U.K.

Session 8B: IMAGE RESTORATION AND RECONSTRUCTION

Time & Place: 2.20-5.50, Lecture Theatre B (Oral Session)
Chairperson: M. Gabouj, Signal Processing Laboratory, Tampere University of Technology, PO Box 553, 33101 Tampere, FINLAND

8B.1 A deinterlacing and 4:2:2 to HDTV upconversion algorithm using motion compensation
C. Reillo, Telefonica I+D, Madrid, SPAIN

8B.2 Removal of replacement noise in motion picture sequences using 3D autoregressive modelling
A. Kokarum, Cambridge University, Cambridge, U.K.
8B.3 Motion compensated conversion from interlaced to progressive formats
L. Vandendorpe, L. Cuvelier, B. Maisin, P. Queluz and P. Delogne, Universite Catholique de Louvain, Louvain-la-Neuve, BELGIUM

8B.4 Adaptive weighted median filtering based on local statistics
A. Taguchi, Musashi Institute of Technology, Tokyo, JAPAN; T. Sun and M. Gabbouj, Tampere University of Technology, Tampere, FINLAND

8B.5 Two stage neural networks filters for adaptive restoration of images degraded by both blur and noise
H. Younal, M. Janati-I, L. Elismaili and A. Kada, Universite Mohamed V de Rabat, Rabat, MOROCCO

8B.6 Adaptive separable weighted median filters for image processing
T. Sun, M. Gabbouj and Y. Hwang, University of Technology, Tampere, FINLAND

8B.7 Adaptive filters for edge-preserving smoothing of airborne SAR image speckle noise
C.R. Moloney and S. Ward, Memorial University of Newfoundland, St. John’s, Newfoundland, CANADA

8B.8 3D Wiener filtering for noise suppression in motion picture sequences using overlapped processing
A. Kokkarum, Cambridge University, Cambridge, U.K.

8B.9 An extended Kalman filtering strategy for image feature extraction
C. Xu and S.A. Velastin, King’s College London, London, U.K.

8B.10 Blind and locally adaptive image restoration in the framework of a multiscale Gabor representation
G. Cristobal and R. Navarro, Consejo Superior de Investigaciones Científica, Madrid, SPAIN

Session 8C: NEURAL NETWORKS II

Time & Place: 2.20-5.50, Lecture Theatre C (Oral Session)

Chairperson: G.D. Cain, School of Electronic & Manufacturing Systems, University of Westminster, New Cavendish Street, London W1M 8JS, U.K.

8C.1 Consideration of generalized optimization methods for labeling problems
T. Iwama, T. Horiuchi and K. Toraiuchi, University of Tsukuba, Ibaraki, JAPAN

8C.2 Weight initialisation and node selection in the construction of Casasent network classifiers

8C.3 Regular and fast chaotic neural network learning of translation invariant pattern recognition
V.E. Bondarenko, Moscow Research Institute, Moscow, RUSSIA

8C.4 A fast Kalman filter based new algorithm for training feedforward neural networks
F. Fnaiech, ENSET, Tunis, TUNISIA; D. Bastard, V. Buzenac, R. Settineri and M. Najim, Universite de Bordeaux I, Talence, FRANCE
8C.5 Learning algorithms for neural networks with finite temporal dynamic
N. Benvenuto, Universita di Padova, Padova, ITALY; F. Piazza and A. Uncini, Universita di Ancona, Ancona, ITALY

8C.6 Clustering in non-stationary pattern recognition systems
M. Markovic, M. Milosavljevic, Institute of Applied Mathematics & Electronics, Belgrade, YUGOSLAVIA; and B. Kovacevic, Faculty of Electrical Engineering, Belgrade, YUGOSLAVIA

8C.7 Adiabatic layering: beyond hierarchical multi-scale optimization
B. Truyen and J. Cornelis, Free University of Brussels, Brussels, BELGIUM

8C.8 Reconstruction of PAM signals using a multilayer perceptron with a multilevel sigmoidal function
K. Hacioglu and M. Abdelhafez, Eastern Mediterranean University, Mersin, TURKEY

8C.9 Optimal functional-link-net-based linear feedforward and decision feedback equalizers
A. Hussain, J.J. Soraghan and T.S. Durrani, University of Strathclyde, Glasgow, U.K.

8C.10 Improving the radial basis function networks for homogeneous nonstationary time series prediction

Session 8D: ESTIMATION II

Time & Place: 2.20-5.50, Lecture Theatre D (Oral Session)

Chairperson: D.T.M. Slock, Institut EURECOM, 2229 Route des Cretes, BP 193, 06904 Sophia Antipolis Cedex, FRANCE

8D.1 On-line constrained deconvolution
N. Souilah and G. Thomas, Ecole Centrale de Lyon, Ecully, FRANCE

8D.2 Towards globally convergent blind equalization of constant modulus signals: a bilinear approach
C.B. Papadis and D.T.M. Slock, Institut EURECOM, Sophia Antipolis, FRANCE

8D.3 Efficient algorithms for instrumental variable system identification
G-O. Glentis and N. Kalouptsidis, University of Athens, Athens, GREECE

8D.4 Optimal decimation-interpolation based parameter estimation
J. Xin and A. Sano, Keio University, Yokohama, JAPAN

8D.5 Prediction of the asymptotic and threshold behaviour of Bayesian signal parameter estimators
A. Quinn, Trinity College, University of Dublin, Dublin, IRELAND

8D.6 A method solving normal equation by ARMA lattice filter realization algorithm
M. Haseyama, N. Nagai and N. Miki, Hokkaido University, Sapporo, JAPAN

8D.7 Estimating the Kullback-Leibler information for autoregressive model order selection in finite samples
H.E. Wensink and P.M.T. Broersen, Delft University of Technology, Delft, NETHERLANDS
8D.8 On the performance of AR order selection methods
J.R. Dickie and A.K. Nandi, University of Strathclyde, Glasgow, U.K.

8D.9 Information criteria and abrupt changes in probability laws
O. Colot, C. Olivier, P. Courtellemont, Universite de Rouen, Mont-Saint-Aignan, FRANCE; and A. El Matouat, Ecole Normale Superieure, Fez Principale, MOROCCO

8D.10 Estimation of mixing distributions. Application to a classification problem in signal processing
F. Brouaye, SUPELEC, Gif-sur-Yvette, FRANCE

Session 8E: PARALLEL AND NOVEL ARCHITECTURES

Time & Place: 2.20-5.50, Lecture Theatre E (Oral Session)

Chairperson: J.V. McCanny, Department of Electrical Engineering, Queens University of Belfast, Ashby Building, Stranmillis Road, Belfast BT9 5AH, N. IRELAND

8E.1 Mapping directionally weighted recursive least squares computations on an SVD updating array
M. Moonen, Katholieke Universiteit Leuven, Heverlee, BELGIUM

8E.2 CORDIC-based approximate rotations for SVD and QRD
J. Goetze, Technical University of Munich, Munich, GERMANY

8E.3 Implementation of adaptive signal processing architectures based on dynamically reconfigurable FPGAs
P. Lysaght and H.P. Dick, University of Strathclyde, Glasgow, U.K.

8E.4 A low power VLSI architecture with an application to adaptive algorithms for digital hearing aids
A. Heubi, S. Grassi, M. Ansorge and F. Pellandini, University of Neuchatel, Neuchatel-Serrières, SWITZERLAND

8E.5 Solving simultaneous linear equations over GF(p)
S. Fenn, M. Benaissa and D. Taylor, University of Huddersfield, Huddersfield, U.K.

8E.6 Digital filters for cochlea implants using mixed mode gate arrays
T. Olbrich and M. Shaw, University of Central Lancashire, Preston, U.K.

8E.7 Computational structures for recursive digital filters using the delta-operator
D.I. Patel and R.M. Goodall, Loughborough University of Technology, Loughborough, U.K.

8E.8 Single-clock delay-commutator for SR-FFT pipeline implementation
J. Garcia, J.A. Michell and A.M. Buron, Universidad de Cantabria, Santander, SPAIN

8E.9 On the implementation of the generalized stack algorithm and the comparison of sorting strategies
K. Achtmann and W. Rupprecht, Universitaet Kaiserslautern, Kaiserslautern, GERMANY

8E.10 A parallel approach to fractal-based image coder
F.J. Gonzalez-Serrano, O.W. Marquez and J.L. Alba-Castro, Universidad de Vigo, Vigo, SPAIN
AUTHOR INDEX

NOMENCLATURE

All papers are of the following form:

\[ s_l p \]

where

- \( s \) represents a Session Number (1-8) or the Opening Session (OS)
- \( l \) indicates a Lecture Theatre (A-E), Poster Display Area (P) or an Invited Paper (I)
- \( p \) indicates a Paper Number (1-20)

Chairs are indicated by:

\[ s_l \text{ Chair} \]
<table>
<thead>
<tr>
<th>Name</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdat, M.</td>
<td>3P.4</td>
</tr>
<tr>
<td>Abdelhafiz, M.</td>
<td>8C.8</td>
</tr>
<tr>
<td>Abed Meraim, K.</td>
<td>5C.8</td>
</tr>
<tr>
<td>Aboutajbine, D.</td>
<td>5F.11</td>
</tr>
<tr>
<td>Abramovich, Y.I.</td>
<td>4A.1</td>
</tr>
<tr>
<td>Abrardo, A.</td>
<td>7C.9</td>
</tr>
<tr>
<td>Abreu, V.</td>
<td>3A.1</td>
</tr>
<tr>
<td>Achtmann, K.</td>
<td>8E.9</td>
</tr>
<tr>
<td>Adams, A.E.</td>
<td>4Pb.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>7C.8</td>
</tr>
<tr>
<td>Adib, A.</td>
<td>5P.11</td>
</tr>
<tr>
<td>Affes, S.</td>
<td>2P.12</td>
</tr>
<tr>
<td>&quot;</td>
<td>4D.3</td>
</tr>
<tr>
<td>Agati, P.</td>
<td>4Pa.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>6P.13</td>
</tr>
<tr>
<td>Agrawal, R.C.</td>
<td>6P.20</td>
</tr>
<tr>
<td>Akbari Azizani, A.</td>
<td>6A.9</td>
</tr>
<tr>
<td>Al-Adnani, A.</td>
<td>5D.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>5D.4</td>
</tr>
<tr>
<td>Al-Ahmad, H.</td>
<td>3E.10</td>
</tr>
<tr>
<td>Alba-Castro, J.L.</td>
<td>3A.6</td>
</tr>
<tr>
<td>&quot;</td>
<td>8E.10</td>
</tr>
<tr>
<td>Albiol, A.</td>
<td>1D.1</td>
</tr>
<tr>
<td>Alengrin, G.</td>
<td>3E.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>5P.14</td>
</tr>
<tr>
<td>Alexandre, F.</td>
<td>1E.4</td>
</tr>
<tr>
<td>Alfredsson, L-I.</td>
<td>7D.8</td>
</tr>
<tr>
<td>Algazi, V.R.</td>
<td>2A.3</td>
</tr>
<tr>
<td>Alkin, O.</td>
<td>5D.1</td>
</tr>
<tr>
<td>Almenar, V.</td>
<td>1D.1</td>
</tr>
<tr>
<td>Alparone, L.</td>
<td>3P.19</td>
</tr>
<tr>
<td>Alshatti, W.</td>
<td>4Pa.8</td>
</tr>
<tr>
<td>Alvarex, A.</td>
<td>6E.4</td>
</tr>
<tr>
<td>Amblard, P.O.</td>
<td>3C.5</td>
</tr>
<tr>
<td>Anzlou, M.</td>
<td>2A.7</td>
</tr>
<tr>
<td>Anarim, E.</td>
<td>2P.8</td>
</tr>
<tr>
<td>&quot;</td>
<td>5D.1</td>
</tr>
<tr>
<td>Ancin, F.J.</td>
<td>7P.13</td>
</tr>
<tr>
<td>Andina, D.</td>
<td>7D.10</td>
</tr>
<tr>
<td>Ansoe, M.</td>
<td>7E.9</td>
</tr>
<tr>
<td>&quot;</td>
<td>8E.4</td>
</tr>
<tr>
<td>Antweiller, C.</td>
<td>8A.9</td>
</tr>
<tr>
<td>Argenti, F.</td>
<td>3P.15</td>
</tr>
<tr>
<td>Asghar, S.M.</td>
<td>6A.7</td>
</tr>
<tr>
<td>Atherton, P.</td>
<td>7A.7</td>
</tr>
<tr>
<td>Auger, F.</td>
<td>3D.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>6P.16</td>
</tr>
<tr>
<td>Auguin, M.</td>
<td>6E.2</td>
</tr>
<tr>
<td>Ayer, S.</td>
<td>1B.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>4B.8</td>
</tr>
<tr>
<td>Baghban-Ravari, L.</td>
<td>7P.7</td>
</tr>
<tr>
<td>Bahl, L.R.</td>
<td>2A.7</td>
</tr>
<tr>
<td>Balakrishnan-Aiyer, S.V.</td>
<td>2A.7</td>
</tr>
<tr>
<td>Balestra, P.</td>
<td>6E.2</td>
</tr>
<tr>
<td>Ballarin, V.</td>
<td>6P.7</td>
</tr>
<tr>
<td>Balseaks, J.</td>
<td>3P.20</td>
</tr>
<tr>
<td>Balss, J.</td>
<td>5A.3</td>
</tr>
<tr>
<td>Banga, E.R.</td>
<td>3A.6</td>
</tr>
<tr>
<td>Bangham, J.A.</td>
<td>4P.6</td>
</tr>
<tr>
<td>Bangham, J.A.</td>
<td>4P.7</td>
</tr>
<tr>
<td>&quot;</td>
<td>7E.9</td>
</tr>
<tr>
<td>Bao, Z.</td>
<td>3E.5</td>
</tr>
<tr>
<td>Bar-Ness, Y.</td>
<td>2E.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>2P.11</td>
</tr>
<tr>
<td>Barlaud, M.</td>
<td>3P.6</td>
</tr>
<tr>
<td>Barnard, H.J.</td>
<td>6B.6</td>
</tr>
<tr>
<td>Barni, M.</td>
<td>7D.2</td>
</tr>
<tr>
<td>Baronkin, V.</td>
<td>1E.2</td>
</tr>
<tr>
<td>Baronti, S.</td>
<td>3P.19</td>
</tr>
<tr>
<td>Barret, M.</td>
<td>2P.10</td>
</tr>
<tr>
<td>Bartolini, F.</td>
<td>7D.2</td>
</tr>
<tr>
<td>Baskurt, A.</td>
<td>6B.7</td>
</tr>
<tr>
<td>Bastard, D.</td>
<td>8C.4</td>
</tr>
<tr>
<td>Bataillou, E.</td>
<td>5D.7</td>
</tr>
<tr>
<td>Baylou, P.</td>
<td>6E.3</td>
</tr>
<tr>
<td>Beamish, N.</td>
<td>7C.5</td>
</tr>
<tr>
<td>Beaty, M.G.</td>
<td>3P.10</td>
</tr>
<tr>
<td>Beausseury, P.</td>
<td>6P.12</td>
</tr>
<tr>
<td>Beet, S.W.</td>
<td>7P.6</td>
</tr>
<tr>
<td>&quot;</td>
<td>7P.7</td>
</tr>
<tr>
<td>Bellanger, M.G.</td>
<td>3P.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>5C Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>6C.2</td>
</tr>
<tr>
<td>Bellegarda, E.J.</td>
<td>2D.9</td>
</tr>
<tr>
<td>&quot;</td>
<td>6P.17</td>
</tr>
<tr>
<td>Bellegarda, J.R.</td>
<td>2A.7</td>
</tr>
<tr>
<td>&quot;</td>
<td>6P.17</td>
</tr>
<tr>
<td>Bellemain, P.</td>
<td>3P.6</td>
</tr>
<tr>
<td>Bellis, S.J.</td>
<td>6E.9</td>
</tr>
<tr>
<td>Belouchrani, A.</td>
<td>4D.5</td>
</tr>
<tr>
<td>Beltran, J.R.</td>
<td>6P.4</td>
</tr>
<tr>
<td>Benissa, M.</td>
<td>8E.5</td>
</tr>
<tr>
<td>Benelli, G.</td>
<td>3P.15</td>
</tr>
<tr>
<td>&quot;</td>
<td>7C.9</td>
</tr>
<tr>
<td>Benhallam, A.</td>
<td>5E.8</td>
</tr>
<tr>
<td>Benidir, M.</td>
<td>2C.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>2P.10</td>
</tr>
<tr>
<td>Benoît-Cattin, H.</td>
<td>6B.7</td>
</tr>
<tr>
<td>Benquel, L.</td>
<td>3B.5</td>
</tr>
<tr>
<td>Bentley, P.M.</td>
<td>1D.6</td>
</tr>
<tr>
<td>Benvenuto, N.</td>
<td>8C.5</td>
</tr>
<tr>
<td>Berberidès, K.</td>
<td>4C.5</td>
</tr>
<tr>
<td>Berenguier, C.</td>
<td>2E.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>3E.3</td>
</tr>
<tr>
<td>Berizzi, F.</td>
<td>3E.7</td>
</tr>
<tr>
<td>Berkhoff, A.P.</td>
<td>7A.6</td>
</tr>
<tr>
<td>Bermejo, M.</td>
<td>6D.2</td>
</tr>
<tr>
<td>Bernhard, H-P.</td>
<td>1C.4</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Biemond, J.</td>
<td>2B.6</td>
</tr>
<tr>
<td></td>
<td>4B Chair</td>
</tr>
<tr>
<td></td>
<td>6B.6</td>
</tr>
<tr>
<td></td>
<td>7B.9</td>
</tr>
<tr>
<td>Bini, A.</td>
<td>7C.9</td>
</tr>
<tr>
<td>Bistritz, Y.</td>
<td>3A.9</td>
</tr>
<tr>
<td>Bleakley, C.J.</td>
<td>7P.16</td>
</tr>
<tr>
<td>Bloch, I.</td>
<td>3B.6</td>
</tr>
<tr>
<td></td>
<td>5B.8</td>
</tr>
<tr>
<td>Boehme, J.F.</td>
<td>2P.18</td>
</tr>
<tr>
<td></td>
<td>3E Chair</td>
</tr>
<tr>
<td></td>
<td>4A.1</td>
</tr>
<tr>
<td></td>
<td>4A.3</td>
</tr>
<tr>
<td>Boeri, F.</td>
<td>6E.2</td>
</tr>
<tr>
<td>Bolon, P.</td>
<td>4Pa.8</td>
</tr>
<tr>
<td>Bonafonte, A.</td>
<td>1C.5</td>
</tr>
<tr>
<td>Bonat, I.</td>
<td>3P.8</td>
</tr>
<tr>
<td></td>
<td>4Pa.3</td>
</tr>
<tr>
<td>Bondarenko, V.E.</td>
<td>8C.3</td>
</tr>
<tr>
<td>Bondon, P.</td>
<td>5P.15</td>
</tr>
<tr>
<td>Bonnauad, L.</td>
<td>2B.4</td>
</tr>
<tr>
<td>Boroczky, L.</td>
<td>4B.3</td>
</tr>
<tr>
<td></td>
<td>4Pb.5</td>
</tr>
<tr>
<td>Boss, D.</td>
<td>5P.2</td>
</tr>
<tr>
<td>Bouisson, E.</td>
<td>2D.7</td>
</tr>
<tr>
<td>Bournahdi, M.</td>
<td>2C.6</td>
</tr>
<tr>
<td></td>
<td>2D.10</td>
</tr>
<tr>
<td>Bourennane, E.</td>
<td>6E.8</td>
</tr>
<tr>
<td>Bouthersy, P.</td>
<td>3B.1</td>
</tr>
<tr>
<td>Boyce, J.F.</td>
<td>2D.2</td>
</tr>
<tr>
<td>Boyd, C.A.</td>
<td>5C.1</td>
</tr>
<tr>
<td>Brandwood, D.H.</td>
<td>3D.3</td>
</tr>
<tr>
<td>Brie, D.</td>
<td>2E.6</td>
</tr>
<tr>
<td></td>
<td>4A.9</td>
</tr>
<tr>
<td>Briguet, A.</td>
<td>4A.7</td>
</tr>
<tr>
<td>Broersen, P.M.T.</td>
<td>5A.8</td>
</tr>
<tr>
<td></td>
<td>8D.7</td>
</tr>
<tr>
<td>Brossier, J.M.</td>
<td>3C.5</td>
</tr>
<tr>
<td>Brouayre, F.</td>
<td>8D.10</td>
</tr>
<tr>
<td>Brown, K.</td>
<td>2A.3</td>
</tr>
<tr>
<td></td>
<td>2P Chair</td>
</tr>
<tr>
<td></td>
<td>4E.9</td>
</tr>
<tr>
<td>Burdeau, J.L.</td>
<td>2P.4</td>
</tr>
<tr>
<td>Burkhardt, H.</td>
<td>7B.4</td>
</tr>
<tr>
<td>Buren, A.M.</td>
<td>8E.8</td>
</tr>
<tr>
<td>Burrows, B.L.</td>
<td>7P.13</td>
</tr>
<tr>
<td>Buzenac, V.</td>
<td>8C.4</td>
</tr>
<tr>
<td>Cabello, D.</td>
<td>4Pa.1</td>
</tr>
<tr>
<td>Caglar, H.</td>
<td>2P.8</td>
</tr>
<tr>
<td></td>
<td>5D.1</td>
</tr>
<tr>
<td>Cain, G.D.</td>
<td>7D.9</td>
</tr>
<tr>
<td></td>
<td>8A.3</td>
</tr>
<tr>
<td></td>
<td>8C Chair</td>
</tr>
<tr>
<td>Callender, C.P.</td>
<td>7C.4</td>
</tr>
<tr>
<td>Callender, C.P.</td>
<td>7C.6</td>
</tr>
<tr>
<td>Calvagno, G.</td>
<td>6B.10</td>
</tr>
<tr>
<td>Campbell, D.R.</td>
<td>6A.10</td>
</tr>
<tr>
<td>Cao, Z.</td>
<td>6A.4</td>
</tr>
<tr>
<td>Capdessus, C.</td>
<td>7E.1</td>
</tr>
<tr>
<td>Capdevielle, V.</td>
<td>5D.10</td>
</tr>
<tr>
<td>Capitanio, L.</td>
<td>5E.2</td>
</tr>
<tr>
<td>Cappellini, V.</td>
<td>7D.2</td>
</tr>
<tr>
<td>Carayannis, G.</td>
<td>1A Chair</td>
</tr>
<tr>
<td></td>
<td>1A.1</td>
</tr>
<tr>
<td></td>
<td>1A.5</td>
</tr>
<tr>
<td>Cardoso, J-F.</td>
<td>2C.3</td>
</tr>
<tr>
<td></td>
<td>4D.5</td>
</tr>
<tr>
<td></td>
<td>4D.7</td>
</tr>
<tr>
<td>Carini, A.</td>
<td>3A.5</td>
</tr>
<tr>
<td>Cariou, C.</td>
<td>1B.4</td>
</tr>
<tr>
<td>Carl, H.</td>
<td>6A.2</td>
</tr>
<tr>
<td>Carminelli Gregori, A.M.</td>
<td>3P.8</td>
</tr>
<tr>
<td></td>
<td>4Pa.3</td>
</tr>
<tr>
<td>Carrasco, R.A.</td>
<td>1A.4</td>
</tr>
<tr>
<td></td>
<td>6D.4</td>
</tr>
<tr>
<td></td>
<td>7P.13</td>
</tr>
<tr>
<td>Carrato, S.</td>
<td>3C.2</td>
</tr>
<tr>
<td>Carreira, M.J.</td>
<td>4Pb.1</td>
</tr>
<tr>
<td>Carriere, C.</td>
<td>6E.2</td>
</tr>
<tr>
<td>Carrion, M.C.</td>
<td>5P.13</td>
</tr>
<tr>
<td>Casas, J.R.</td>
<td>2B.3</td>
</tr>
<tr>
<td>Casella, E.</td>
<td>4E.3</td>
</tr>
<tr>
<td>Castanie, F.</td>
<td>3E.6</td>
</tr>
<tr>
<td></td>
<td>6E.15</td>
</tr>
<tr>
<td>Castignolles, N</td>
<td>6F.10</td>
</tr>
<tr>
<td>Cattoen, M.</td>
<td>3B.10</td>
</tr>
<tr>
<td></td>
<td>6P.10</td>
</tr>
<tr>
<td>Cecchini, A.</td>
<td>5E.2</td>
</tr>
<tr>
<td>Cedervall, M.</td>
<td>6D.5</td>
</tr>
<tr>
<td>Celasun, I.</td>
<td>2P.8</td>
</tr>
<tr>
<td>Cerquides, J.R.</td>
<td>6D.7</td>
</tr>
<tr>
<td>Cerutti, S.</td>
<td>5E.2</td>
</tr>
<tr>
<td>Chahine, G.</td>
<td>5A.1</td>
</tr>
<tr>
<td>Champagne, B.</td>
<td>4D.4</td>
</tr>
<tr>
<td>Champion, T.G</td>
<td>3A.7</td>
</tr>
<tr>
<td>Chan, C.F.</td>
<td>5A.9</td>
</tr>
<tr>
<td></td>
<td>7A.1</td>
</tr>
<tr>
<td>Chang, K-H.</td>
<td>7A.4</td>
</tr>
<tr>
<td>Chapell, E.</td>
<td>8la Chair</td>
</tr>
<tr>
<td>Chapman, R.</td>
<td>5D.2</td>
</tr>
<tr>
<td></td>
<td>5D.4</td>
</tr>
<tr>
<td>Charkani, N.</td>
<td>3C.5</td>
</tr>
<tr>
<td>Chaumeron, T.</td>
<td>1D.3</td>
</tr>
<tr>
<td>Chehdi, K.</td>
<td>1B.4</td>
</tr>
<tr>
<td>Chen, H.</td>
<td>7D.1</td>
</tr>
<tr>
<td>Chen, S.</td>
<td>8C.10</td>
</tr>
<tr>
<td>Cheong, C.K.</td>
<td>2P.16</td>
</tr>
<tr>
<td>Cherif, S.</td>
<td>5P.19</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Chevalier, E.</td>
<td>1D.4</td>
</tr>
<tr>
<td>Chi, C-Y.</td>
<td>5P.12</td>
</tr>
<tr>
<td>Chia, L.T.</td>
<td>3P.18</td>
</tr>
<tr>
<td>Chieu, B-C.</td>
<td>4E.1</td>
</tr>
<tr>
<td>Chiu, C.C.</td>
<td>2A.8</td>
</tr>
<tr>
<td>Chng, E.S.</td>
<td>8C.10</td>
</tr>
<tr>
<td>Chonovel, T.</td>
<td>3D.4</td>
</tr>
<tr>
<td>Choukri, K.</td>
<td>2E.9</td>
</tr>
<tr>
<td>Chretien, S.</td>
<td>3P.12</td>
</tr>
<tr>
<td>Christopoulos,</td>
<td>C.A.</td>
</tr>
<tr>
<td></td>
<td>3P.13</td>
</tr>
<tr>
<td>Chu, M.T.</td>
<td>7A.1</td>
</tr>
<tr>
<td>Clarke, R.J.</td>
<td>1B Chair</td>
</tr>
<tr>
<td></td>
<td>8Bb Chair</td>
</tr>
<tr>
<td>Clergeot, H.</td>
<td>6D.1</td>
</tr>
<tr>
<td>Cocurullo, F.</td>
<td>3P.3</td>
</tr>
<tr>
<td>Colot, O.</td>
<td>8D.9</td>
</tr>
<tr>
<td>Comes, S.</td>
<td>3P.16</td>
</tr>
<tr>
<td>Comon, P.</td>
<td>3C Chair</td>
</tr>
<tr>
<td></td>
<td>3C.9</td>
</tr>
<tr>
<td>Cong, L.</td>
<td>2A.1</td>
</tr>
<tr>
<td>Constantinides, A.G.</td>
<td>2E Chair</td>
</tr>
<tr>
<td></td>
<td>3C.8</td>
</tr>
<tr>
<td>Conte, E.</td>
<td>3E.1</td>
</tr>
<tr>
<td>Cornelis, J.</td>
<td>8C.7</td>
</tr>
<tr>
<td>Coroyer, C.</td>
<td>3C.1</td>
</tr>
<tr>
<td>Courtellemont, P.</td>
<td>8D.9</td>
</tr>
<tr>
<td>Cowan, C.F.N.</td>
<td>21 Chair</td>
</tr>
<tr>
<td></td>
<td>6C.3</td>
</tr>
<tr>
<td></td>
<td>7C.4</td>
</tr>
<tr>
<td></td>
<td>7C.6</td>
</tr>
<tr>
<td></td>
<td>7E Chair</td>
</tr>
<tr>
<td>Cox, C.</td>
<td>7E.8</td>
</tr>
<tr>
<td>Crestet, J.</td>
<td>7P.5</td>
</tr>
<tr>
<td>Cristobal, G.</td>
<td>8B.10</td>
</tr>
<tr>
<td>Cryan, R.A.</td>
<td>8A.5</td>
</tr>
<tr>
<td>Csillag, P.</td>
<td>4B.3</td>
</tr>
<tr>
<td>Curinga, S.</td>
<td>3P.3</td>
</tr>
<tr>
<td>Cuvelier, L.</td>
<td>8B.3</td>
</tr>
<tr>
<td>Dai, J.</td>
<td>2A.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>7P.3</td>
</tr>
<tr>
<td>Dainty, J.C.</td>
<td>3C.4</td>
</tr>
<tr>
<td>Dal Ponte, D.</td>
<td>2B.6</td>
</tr>
<tr>
<td>&quot;</td>
<td>4A.9</td>
</tr>
<tr>
<td>Dalle Mese, E.</td>
<td>3E.7</td>
</tr>
<tr>
<td>Dalle Molle, J.W.</td>
<td>3C.7</td>
</tr>
<tr>
<td>Danesfahani, G.R.</td>
<td>5C.6</td>
</tr>
<tr>
<td>Das, A.</td>
<td>5A.10</td>
</tr>
<tr>
<td>Davies, A.C.</td>
<td>6P.8</td>
</tr>
<tr>
<td>Davies, C.J.</td>
<td>6P.9</td>
</tr>
<tr>
<td>De Lathauer, L.</td>
<td>2C.1</td>
</tr>
<tr>
<td>De Moor, B.</td>
<td>2C.1</td>
</tr>
<tr>
<td>De Viendt, J.</td>
<td>3B.9</td>
</tr>
<tr>
<td>Deschamps, F.</td>
<td>7E.6</td>
</tr>
<tr>
<td>Dedieu, H.</td>
<td>7B.5</td>
</tr>
<tr>
<td>Deergha Rao, K.</td>
<td>4B.4</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Eriksson, A.</td>
<td>4D.10</td>
</tr>
<tr>
<td>Eriksson, T.</td>
<td>3A.4</td>
</tr>
<tr>
<td>Erkens, J.S.</td>
<td>5A.8</td>
</tr>
<tr>
<td>Erwood, A.F.</td>
<td>2A.1</td>
</tr>
<tr>
<td>Eryurtu, F.</td>
<td>2B.5</td>
</tr>
<tr>
<td>Esebar, A.</td>
<td>2D.7</td>
</tr>
<tr>
<td>Estola, K.P.</td>
<td>4Pa.10</td>
</tr>
<tr>
<td>Euler, S.</td>
<td>7P.20</td>
</tr>
<tr>
<td>Evans, B.G.</td>
<td>2B.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>3A.0</td>
</tr>
<tr>
<td>&quot;</td>
<td>3A.10</td>
</tr>
<tr>
<td>&quot;</td>
<td>5C.6</td>
</tr>
<tr>
<td>Fackrell, J.W.A.</td>
<td>5P.7</td>
</tr>
<tr>
<td>Fagan, A.D.</td>
<td>5A Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>5A.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>5A.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>7C.5</td>
</tr>
<tr>
<td>Falkhausen, M.</td>
<td>7P.20</td>
</tr>
<tr>
<td>Fan, H-H.</td>
<td>3E.9</td>
</tr>
<tr>
<td>Fauson, G.</td>
<td>6A.9</td>
</tr>
<tr>
<td>Fauth, A.</td>
<td>6E.1</td>
</tr>
<tr>
<td>Favier, G.</td>
<td>2C.8</td>
</tr>
<tr>
<td>&quot;</td>
<td>4E.8</td>
</tr>
<tr>
<td>Fayn, J.</td>
<td>1D.4</td>
</tr>
<tr>
<td>Fazekas, K.</td>
<td>4B.3</td>
</tr>
<tr>
<td>Fazekas, Z.</td>
<td>4Pb.5</td>
</tr>
<tr>
<td>Feng, C-C.</td>
<td>5P.12</td>
</tr>
<tr>
<td>Feng, G.</td>
<td>3A.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>3A.8</td>
</tr>
<tr>
<td>&quot;</td>
<td>7P.15</td>
</tr>
<tr>
<td>Feng, Z-M.</td>
<td>3E.9</td>
</tr>
<tr>
<td>Fenn, S.</td>
<td>8E.5</td>
</tr>
<tr>
<td>Fernandez, J.</td>
<td>7C.7</td>
</tr>
<tr>
<td>Fernandez-Rubio, J.A.</td>
<td>6D.7</td>
</tr>
<tr>
<td>Ferrari, A.</td>
<td>5P.14</td>
</tr>
<tr>
<td>Ferrer-Ballester, M.A.</td>
<td>5A.5</td>
</tr>
<tr>
<td>Fettweis, A.</td>
<td>7D Chair</td>
</tr>
<tr>
<td>Fety, L.</td>
<td>6C.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>7C.2</td>
</tr>
<tr>
<td>Figueiras-Vidal, A.R.</td>
<td>4C.7</td>
</tr>
<tr>
<td>&quot;</td>
<td>5A.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>5E.9</td>
</tr>
<tr>
<td>Filligoi, G.C.</td>
<td>5E.2</td>
</tr>
<tr>
<td>Fish, P.J.</td>
<td>6E.9</td>
</tr>
<tr>
<td>Fisher, R.B.</td>
<td>5B.6</td>
</tr>
<tr>
<td>Fitzgibbon, A.W.</td>
<td>5B.6</td>
</tr>
<tr>
<td>Fjallbrant, T.</td>
<td>6A.5</td>
</tr>
<tr>
<td>Flieller, A.</td>
<td>6D.1</td>
</tr>
<tr>
<td>Fniech, F.</td>
<td>8C.4</td>
</tr>
<tr>
<td>Fonollosa, J.A.R.</td>
<td>1C.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>5P.6</td>
</tr>
<tr>
<td>Fontaine, V.</td>
<td>2A.5</td>
</tr>
<tr>
<td>Foresti, G.L.</td>
<td>6P.1</td>
</tr>
<tr>
<td>Forster, P.</td>
<td>1E.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fotinea, S.F.</td>
<td>1A.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>1A.5</td>
</tr>
<tr>
<td>Francescato, D.</td>
<td>3A.5</td>
</tr>
<tr>
<td>Franich, R.E.H.</td>
<td>2B.6</td>
</tr>
<tr>
<td>Frauenberger, O.</td>
<td>2B.7</td>
</tr>
<tr>
<td>Fred, A.</td>
<td>4Pb.8</td>
</tr>
<tr>
<td>Fredrickson, S.E.</td>
<td>4E.6</td>
</tr>
<tr>
<td>Freeland, G.C.</td>
<td>1C.2</td>
</tr>
<tr>
<td>Freeman, D.</td>
<td>7E.2</td>
</tr>
<tr>
<td>Freericks, M.</td>
<td>6E.1</td>
</tr>
<tr>
<td>Frouin, V.</td>
<td>5B.8</td>
</tr>
<tr>
<td>Fulseth, A.</td>
<td>5A.6</td>
</tr>
<tr>
<td>Gabbouj, M.</td>
<td>2P.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>3P.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>5B.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>7B.6</td>
</tr>
<tr>
<td>&quot;</td>
<td>7D.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>7D.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>8B Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>8B.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>8B.6</td>
</tr>
<tr>
<td>Gaillard, P.</td>
<td>6P.12</td>
</tr>
<tr>
<td>Galand, C.</td>
<td>5C.4</td>
</tr>
<tr>
<td>Gallego, A.</td>
<td>5P.13</td>
</tr>
<tr>
<td>Garcia, J.</td>
<td>8E.8</td>
</tr>
<tr>
<td>Garcia, N.</td>
<td>7E.7</td>
</tr>
<tr>
<td>Garcia-Garduno, V.</td>
<td>2B.4</td>
</tr>
<tr>
<td>Garcia-Mateo, C.</td>
<td>3A.6</td>
</tr>
<tr>
<td>Gardner, W.A.</td>
<td>4C.2</td>
</tr>
<tr>
<td>Garnier, H.</td>
<td>3D.9</td>
</tr>
<tr>
<td>Garzelli, A.</td>
<td>3P.15</td>
</tr>
<tr>
<td>&quot;</td>
<td>7C.9</td>
</tr>
<tr>
<td>Gasull, A.</td>
<td>3B.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>5P.6</td>
</tr>
<tr>
<td>Gauthier, P.</td>
<td>3P.6</td>
</tr>
<tr>
<td>Gavin, P.</td>
<td>2D.7</td>
</tr>
<tr>
<td>Gazor, S.</td>
<td>2P.12</td>
</tr>
<tr>
<td>&quot;</td>
<td>4D.3</td>
</tr>
<tr>
<td>Gell, M.</td>
<td>6D.4</td>
</tr>
<tr>
<td>Gelli, G.</td>
<td>4D.1</td>
</tr>
<tr>
<td>Gersemé, F.</td>
<td>4A.3</td>
</tr>
<tr>
<td>Gersho, A.</td>
<td>5A.10</td>
</tr>
<tr>
<td>Gesbert, D.</td>
<td>4C.1</td>
</tr>
<tr>
<td>Ghanbari, M.</td>
<td>2B Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>6B.3</td>
</tr>
<tr>
<td>Gharavi, H.</td>
<td>41.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>6B Chair</td>
</tr>
<tr>
<td>Giannakis, G.B.</td>
<td>2C.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>5P Chair</td>
</tr>
<tr>
<td>Gilloire, A.</td>
<td>5C.7</td>
</tr>
<tr>
<td>Gini, F.</td>
<td>3E.7</td>
</tr>
<tr>
<td>Girardier, M.</td>
<td>4Pb.2</td>
</tr>
<tr>
<td>Giulieri, A.</td>
<td>6E.2</td>
</tr>
<tr>
<td>Glangeaud, F.</td>
<td>2D.8</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Glancesaud, F.</td>
<td>2D.10</td>
</tr>
<tr>
<td>Glentis, G-O.</td>
<td>2C.10</td>
</tr>
<tr>
<td>&quot;</td>
<td>8D.3</td>
</tr>
<tr>
<td>Goeckler, H.G.</td>
<td>5C.5</td>
</tr>
<tr>
<td>Goetzte, J.</td>
<td>8E.2</td>
</tr>
<tr>
<td>Goh, K.H.</td>
<td>6B.1</td>
</tr>
<tr>
<td>Goldberg, J.</td>
<td>4C.3</td>
</tr>
<tr>
<td>Gomez, P.</td>
<td>4E.7</td>
</tr>
<tr>
<td>Gonzalez, N.</td>
<td>7P.8</td>
</tr>
<tr>
<td>Gonzalez-Serrano, F.J.</td>
<td>8E.10</td>
</tr>
<tr>
<td>Goodall, R.M.</td>
<td>8E.7</td>
</tr>
<tr>
<td>Gopalakrishnan, P.S.</td>
<td>2A.7</td>
</tr>
<tr>
<td>Goransson, B.</td>
<td>6D.8</td>
</tr>
<tr>
<td>Gorokhov, A.Y.</td>
<td>4A.1</td>
</tr>
<tr>
<td>Gouton, P.</td>
<td>4Pb.2</td>
</tr>
<tr>
<td>Goutte, R.</td>
<td>2P.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>6B.7</td>
</tr>
<tr>
<td>Gransden, I.R.</td>
<td>7P.6</td>
</tr>
<tr>
<td>Grant, P.M.</td>
<td>4I Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>5P.1</td>
</tr>
<tr>
<td>Grassi, S.</td>
<td>7P.9</td>
</tr>
<tr>
<td>&quot;</td>
<td>8E.4</td>
</tr>
<tr>
<td>Grattarola, A.A.</td>
<td>3P.3</td>
</tr>
<tr>
<td>Graveron-Demilly, D.</td>
<td>4A.7</td>
</tr>
<tr>
<td>Gray, D.A.</td>
<td>2E.8</td>
</tr>
<tr>
<td>Grenier, D.</td>
<td>2P.17</td>
</tr>
<tr>
<td>Grenier, Y.</td>
<td>2P.12</td>
</tr>
<tr>
<td>&quot;</td>
<td>4D.3</td>
</tr>
<tr>
<td>Griffiths, J.W.R.</td>
<td>3P.18</td>
</tr>
<tr>
<td>Grimble, M.J.</td>
<td>5D.9</td>
</tr>
<tr>
<td>Grochowalski, S.</td>
<td>7P.14</td>
</tr>
<tr>
<td>Gu, C.</td>
<td>2B.2</td>
</tr>
<tr>
<td>Gu, Y.H.</td>
<td>3B.4</td>
</tr>
<tr>
<td>Guarino, C.R.</td>
<td>3E.8</td>
</tr>
<tr>
<td>Guerin, P.</td>
<td>3B.6</td>
</tr>
<tr>
<td>Guglielmi, M.</td>
<td>2E.7</td>
</tr>
<tr>
<td>&quot;</td>
<td>3D.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>5E.3</td>
</tr>
<tr>
<td>Guineunece, P.</td>
<td>1D.5</td>
</tr>
<tr>
<td>Guillaudeau, S.</td>
<td>3A.17</td>
</tr>
<tr>
<td>Guillemon, C.</td>
<td>6B.9</td>
</tr>
<tr>
<td>Guilloteau, S.</td>
<td>2B.4</td>
</tr>
<tr>
<td>Guitton, M.</td>
<td>7P.5</td>
</tr>
<tr>
<td>Guler, E.C.</td>
<td>4Pb.7</td>
</tr>
<tr>
<td>Gunn, S.R.</td>
<td>3B.7</td>
</tr>
<tr>
<td>Haagen, J.</td>
<td>7P.10</td>
</tr>
<tr>
<td>Haardt, M.</td>
<td>7C.1</td>
</tr>
<tr>
<td>Habashy, T.M.</td>
<td>2D.9</td>
</tr>
<tr>
<td>Hacioglu, K.</td>
<td>8C.8</td>
</tr>
<tr>
<td>Haimovitch, A.</td>
<td>2E.1</td>
</tr>
<tr>
<td>Hail, G.</td>
<td>7E.6</td>
</tr>
<tr>
<td>Hallouche, F.</td>
<td>4Pb.4</td>
</tr>
<tr>
<td>Hammond, J.K.</td>
<td>5P.7</td>
</tr>
<tr>
<td>Haneq, J.</td>
<td>6E.5</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Ireton, M.A.</td>
<td>6A.7</td>
</tr>
<tr>
<td>Irvine, A.D.</td>
<td>7P.2</td>
</tr>
<tr>
<td>Irvine, D.</td>
<td>2A.3</td>
</tr>
<tr>
<td>Itoh, S.</td>
<td>2B.8</td>
</tr>
<tr>
<td>Iwama, T.</td>
<td>8C.1</td>
</tr>
<tr>
<td>Izzo, L.</td>
<td>4D.1</td>
</tr>
<tr>
<td>*</td>
<td>5P.8</td>
</tr>
<tr>
<td>Jacovitti, G.</td>
<td>2C.2</td>
</tr>
<tr>
<td>Jaidane-Saidane, M.</td>
<td>5P.19</td>
</tr>
<tr>
<td>Jakobsen, H.</td>
<td>5C.3</td>
</tr>
<tr>
<td>Janiti-I, M.</td>
<td>8B.5</td>
</tr>
<tr>
<td>Jeans, T.G.</td>
<td>5C.6</td>
</tr>
<tr>
<td>Jee, C.W.</td>
<td>3P.9</td>
</tr>
<tr>
<td>Jennane, R.</td>
<td>6P.6</td>
</tr>
<tr>
<td>Jensen, S.H.</td>
<td>6A.1</td>
</tr>
<tr>
<td>Jespers, S.</td>
<td>3D.5</td>
</tr>
<tr>
<td>Jiang, J.</td>
<td>3P.5</td>
</tr>
<tr>
<td>Jolivet, E.</td>
<td>7E.1</td>
</tr>
<tr>
<td>Jones, G.T.</td>
<td>3P.18</td>
</tr>
<tr>
<td>Jorand, C.</td>
<td>3C.1</td>
</tr>
<tr>
<td>Joutsensalo, J.</td>
<td>4A.2</td>
</tr>
<tr>
<td>Junglolaus, N.</td>
<td>7E.10</td>
</tr>
<tr>
<td>Jutten, C.</td>
<td>5P.16</td>
</tr>
<tr>
<td>Kabal, P.</td>
<td>4D.4</td>
</tr>
<tr>
<td>*</td>
<td>5A.1</td>
</tr>
<tr>
<td>Kada, A.</td>
<td>8B.5</td>
</tr>
<tr>
<td>Kadlec, J.</td>
<td>7E.4</td>
</tr>
<tr>
<td>Kahya, Y.P.</td>
<td>4Pb.7</td>
</tr>
<tr>
<td>Kajiki, M.</td>
<td>6C.5</td>
</tr>
<tr>
<td>Kale, I.</td>
<td>7D.9</td>
</tr>
<tr>
<td>Kalouptsidis, N.</td>
<td>2C.10</td>
</tr>
<tr>
<td>*</td>
<td>3D Chair</td>
</tr>
<tr>
<td>*</td>
<td>8D.3</td>
</tr>
<tr>
<td>Kamada, M.</td>
<td>6P.19</td>
</tr>
<tr>
<td>Kammeeyer, K.D.</td>
<td>5P.2</td>
</tr>
<tr>
<td>Karlsson, E.</td>
<td>6P.11</td>
</tr>
<tr>
<td>Karrakchou, M.</td>
<td>3D.8</td>
</tr>
<tr>
<td>*</td>
<td>4A.6</td>
</tr>
<tr>
<td>Kasper, K.</td>
<td>2A.6</td>
</tr>
<tr>
<td>*</td>
<td>7P.19</td>
</tr>
<tr>
<td>Kasprzak, W.</td>
<td>4B.10</td>
</tr>
<tr>
<td>Kastantin, R.</td>
<td>3A.8</td>
</tr>
<tr>
<td>Kecohe, M.</td>
<td>2P.5</td>
</tr>
<tr>
<td>Kennedy, R.A.</td>
<td>2E.2</td>
</tr>
<tr>
<td>Kim, C-J.</td>
<td>3E.2</td>
</tr>
<tr>
<td>Kim, J-I.</td>
<td>7A.4</td>
</tr>
<tr>
<td>Kim, S-D.</td>
<td>3P.7</td>
</tr>
<tr>
<td>*</td>
<td>4B.2</td>
</tr>
<tr>
<td>Kim, W.J.</td>
<td>3P.7</td>
</tr>
<tr>
<td>Kim, Y.H.</td>
<td>3P.7</td>
</tr>
<tr>
<td>Kipper, U.</td>
<td>5A.3</td>
</tr>
<tr>
<td>Kirsteins, I.P.</td>
<td>2E.3</td>
</tr>
<tr>
<td>Kittler, J.</td>
<td>3B.8</td>
</tr>
<tr>
<td>Klefstad-Sillonville, M.</td>
<td>3P.17</td>
</tr>
<tr>
<td>Kleihorst, R.P.</td>
<td>7B.9</td>
</tr>
<tr>
<td>Kleija, W.B.</td>
<td>7F.10</td>
</tr>
<tr>
<td>Knecht, W.</td>
<td>7B.7</td>
</tr>
<tr>
<td>Knoll, A.</td>
<td>6B.1</td>
</tr>
<tr>
<td>Koenig, D.</td>
<td>6P.11</td>
</tr>
<tr>
<td>Kokarum, A.</td>
<td>8B.2</td>
</tr>
<tr>
<td>*</td>
<td>8B.8</td>
</tr>
<tr>
<td>Koliadin, V.L.</td>
<td>2E.4</td>
</tr>
<tr>
<td>Komatsu, T.</td>
<td>4B.7</td>
</tr>
<tr>
<td>Kondo, A.M.</td>
<td>2B.5</td>
</tr>
<tr>
<td>*</td>
<td>3A.10</td>
</tr>
<tr>
<td>Konrad, J.</td>
<td>4B.4</td>
</tr>
<tr>
<td>Konvalinka, I.</td>
<td>5A.7</td>
</tr>
<tr>
<td>Kosaroglu, M.</td>
<td>4D.9</td>
</tr>
<tr>
<td>Kotropoulos, C.</td>
<td>3P.2</td>
</tr>
<tr>
<td>*</td>
<td>7B.6</td>
</tr>
<tr>
<td>Koukoulas, P.</td>
<td>2C.10</td>
</tr>
<tr>
<td>Kovacic, B.</td>
<td>8C.6</td>
</tr>
<tr>
<td>Kovaes, S.</td>
<td>3D.5</td>
</tr>
<tr>
<td>Kresch, R.</td>
<td>4Pa.5</td>
</tr>
<tr>
<td>Krim, H.</td>
<td>7D.6</td>
</tr>
<tr>
<td>Krob, M.</td>
<td>2C.4</td>
</tr>
<tr>
<td>*</td>
<td>5P.15</td>
</tr>
<tr>
<td>Kroener, S.</td>
<td>7B.4</td>
</tr>
<tr>
<td>Kroenke, R.A.</td>
<td>5B.1</td>
</tr>
<tr>
<td>Krukowski, A.</td>
<td>7D.9</td>
</tr>
<tr>
<td>Kubin, G.</td>
<td>1C.4</td>
</tr>
<tr>
<td>Kunt, M.</td>
<td>2B.2</td>
</tr>
<tr>
<td>*</td>
<td>3B Chair</td>
</tr>
<tr>
<td>Labit, C.</td>
<td>2B.4</td>
</tr>
<tr>
<td>Labourdette, O.</td>
<td>7E.1</td>
</tr>
<tr>
<td>Lacoome, J-L.</td>
<td>2C.6</td>
</tr>
<tr>
<td>*</td>
<td>2D.10</td>
</tr>
<tr>
<td>*</td>
<td>5P.18</td>
</tr>
<tr>
<td>*</td>
<td>6D.0</td>
</tr>
<tr>
<td>Lagendijk, R.L.</td>
<td>2B.6</td>
</tr>
<tr>
<td>*</td>
<td>7B.9</td>
</tr>
<tr>
<td>Lagunas, M.A.</td>
<td>1B Chair</td>
</tr>
<tr>
<td>*</td>
<td>5P.18</td>
</tr>
<tr>
<td>*</td>
<td>6D.9</td>
</tr>
<tr>
<td>*</td>
<td>7D.4</td>
</tr>
<tr>
<td>Laheld, B.</td>
<td>2C.3</td>
</tr>
<tr>
<td>Lambert, P.</td>
<td>4Pa.8</td>
</tr>
<tr>
<td>Lara-Barron, M.</td>
<td>3A.3</td>
</tr>
<tr>
<td>Larinier, M.</td>
<td>6P.10</td>
</tr>
<tr>
<td>Larrategui, M.</td>
<td>1A.4</td>
</tr>
<tr>
<td>Larsen, K.J.</td>
<td>5C.3</td>
</tr>
<tr>
<td>Larzabal, P.</td>
<td>6D.1</td>
</tr>
<tr>
<td>Lavagetto, F.</td>
<td>4E.3</td>
</tr>
<tr>
<td>Law, K.W.</td>
<td>5A.9</td>
</tr>
<tr>
<td>Le Bihan, H.</td>
<td>2P.7</td>
</tr>
<tr>
<td>Le Bouquin, R.</td>
<td>6A.9</td>
</tr>
<tr>
<td>Le Carpentier, E.</td>
<td>6P.16</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Le Cerf, P.</td>
<td>7P.1</td>
</tr>
<tr>
<td>Le Roux, J.</td>
<td>2P.2</td>
</tr>
<tr>
<td>*</td>
<td>3C.3</td>
</tr>
<tr>
<td>Le Tourneur, G.</td>
<td>5C.7</td>
</tr>
<tr>
<td>Leandro, M.</td>
<td>2A.2</td>
</tr>
<tr>
<td>Leduc, J-P.</td>
<td>4B.1</td>
</tr>
<tr>
<td>Lee, B-U.</td>
<td>6E.7</td>
</tr>
<tr>
<td>*</td>
<td>7A.4</td>
</tr>
<tr>
<td>Lee, H-J.</td>
<td>3E.2</td>
</tr>
<tr>
<td>Lee, H-S.</td>
<td>3E.2</td>
</tr>
<tr>
<td>Lee, J-B.</td>
<td>4B.2</td>
</tr>
<tr>
<td>Lee, J.</td>
<td>7B.1</td>
</tr>
<tr>
<td>Lee, J-S.</td>
<td>3P.9</td>
</tr>
<tr>
<td>Lee, R.A.</td>
<td>2P.3</td>
</tr>
<tr>
<td>Lee, S-Z.</td>
<td>4B.2</td>
</tr>
<tr>
<td>Leggett, M.</td>
<td>2D.3</td>
</tr>
<tr>
<td>Leigh, H.</td>
<td>1A.3</td>
</tr>
<tr>
<td>*</td>
<td>2A.5</td>
</tr>
<tr>
<td>*</td>
<td>4A.8</td>
</tr>
<tr>
<td>*</td>
<td>6E.5</td>
</tr>
<tr>
<td>Leitao, J.M.</td>
<td>4P.b.8</td>
</tr>
<tr>
<td>Lemoine, O.</td>
<td>2E.5</td>
</tr>
<tr>
<td>Leonardi, R.</td>
<td>2B.9</td>
</tr>
<tr>
<td>Leray, P.</td>
<td>3P.17</td>
</tr>
<tr>
<td>Levin, M.D.</td>
<td>6C.3</td>
</tr>
<tr>
<td>Li, Y-D.</td>
<td>3C.10</td>
</tr>
<tr>
<td>Liang, Y-C.</td>
<td>3C.10</td>
</tr>
<tr>
<td>Liao, G.</td>
<td>3E.5</td>
</tr>
<tr>
<td>Linden, J.</td>
<td>1A.2</td>
</tr>
<tr>
<td>Lise, P.E.</td>
<td>2P.2</td>
</tr>
<tr>
<td>Liu, Q-G.</td>
<td>3E.4</td>
</tr>
<tr>
<td>Lo Presti, L.</td>
<td>2D.5</td>
</tr>
<tr>
<td>Lobos, T.</td>
<td>3D.1</td>
</tr>
<tr>
<td>Lockhart, G.B.</td>
<td>3A.3</td>
</tr>
<tr>
<td>Lopez-Krahe, J.</td>
<td>5B.8</td>
</tr>
<tr>
<td>Lops, M.</td>
<td>3E.1</td>
</tr>
<tr>
<td>Lotti, F.</td>
<td>3P.19</td>
</tr>
<tr>
<td>Loupas, T.</td>
<td>5B.4</td>
</tr>
<tr>
<td>Lowry, A.</td>
<td>5P.10</td>
</tr>
<tr>
<td>Lu, D-J.</td>
<td>3E.4</td>
</tr>
<tr>
<td>*</td>
<td>3E.9</td>
</tr>
<tr>
<td>Lucke, L.E.</td>
<td>5B.1</td>
</tr>
<tr>
<td>Lukasik, E.</td>
<td>7P.14</td>
</tr>
<tr>
<td>Lukovic, B.</td>
<td>5A.7</td>
</tr>
<tr>
<td>Lysaght, P.</td>
<td>8E.3</td>
</tr>
<tr>
<td>Macchi, O.</td>
<td>4C Chair</td>
</tr>
<tr>
<td>*</td>
<td>5P.17</td>
</tr>
<tr>
<td>Macias-Guarasa, J.</td>
<td>2A.2</td>
</tr>
<tr>
<td>Macleod, M.</td>
<td>7D.5</td>
</tr>
<tr>
<td>*</td>
<td>7Pa Chair</td>
</tr>
<tr>
<td>Macq, B.</td>
<td>3P.16</td>
</tr>
<tr>
<td>Maempel, D.</td>
<td>3C.6</td>
</tr>
<tr>
<td>Magre, O.</td>
<td>5E.3</td>
</tr>
<tr>
<td>Maguire, P.T.</td>
<td>6E.6</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>Mailhes, C.</td>
<td>8A.4</td>
</tr>
<tr>
<td>Maison, B.</td>
<td>8B.3</td>
</tr>
<tr>
<td>Maitre, H.</td>
<td>3B.6</td>
</tr>
<tr>
<td>Maimwald, D.</td>
<td>3P.18</td>
</tr>
<tr>
<td>Makowski, R.</td>
<td>3D.7</td>
</tr>
<tr>
<td>Malah, D.</td>
<td>4P.a.5</td>
</tr>
<tr>
<td>Malik, H.</td>
<td>6E.10</td>
</tr>
<tr>
<td>Mangin, J.F.</td>
<td>5B.8</td>
</tr>
<tr>
<td>Mann, I.</td>
<td>7P.15</td>
</tr>
<tr>
<td>Maouche, K.</td>
<td>4C.8</td>
</tr>
<tr>
<td>Maragos, P.</td>
<td>5B.7</td>
</tr>
<tr>
<td>*</td>
<td>7P.11</td>
</tr>
<tr>
<td>Marcos, S.</td>
<td>4D.6</td>
</tr>
<tr>
<td>*</td>
<td>5P.19</td>
</tr>
<tr>
<td>Mari, J-L.</td>
<td>2D.6</td>
</tr>
<tr>
<td>*</td>
<td>2D.8</td>
</tr>
<tr>
<td>Markovic, M.</td>
<td>5A.7</td>
</tr>
<tr>
<td>Markovic, M.</td>
<td>8C.6</td>
</tr>
<tr>
<td>Marlow, K.A.</td>
<td>7E.5</td>
</tr>
<tr>
<td>Marmane, W.P.</td>
<td>6E.9</td>
</tr>
<tr>
<td>Marques, F.</td>
<td>3B.2</td>
</tr>
<tr>
<td>Marquez, O.W.</td>
<td>8E.10</td>
</tr>
<tr>
<td>Mars, J.</td>
<td>2D.6</td>
</tr>
<tr>
<td>*</td>
<td>2D.7</td>
</tr>
<tr>
<td>Marsal, A.</td>
<td>4D.6</td>
</tr>
<tr>
<td>Marshall, S.</td>
<td>4P.a.1</td>
</tr>
<tr>
<td>Martin, J.K.</td>
<td>1C.1</td>
</tr>
<tr>
<td>Martin, N.</td>
<td>3A.8</td>
</tr>
<tr>
<td>Martin, P.</td>
<td>7A.5</td>
</tr>
<tr>
<td>Martin, R.</td>
<td>6A.3</td>
</tr>
<tr>
<td>Martinerie, F.</td>
<td>1E.1</td>
</tr>
<tr>
<td>*</td>
<td>1E.5</td>
</tr>
<tr>
<td>Martinez, C.</td>
<td>1D.5</td>
</tr>
<tr>
<td>Marx, J.</td>
<td>8A.7</td>
</tr>
<tr>
<td>Masgrau, E.</td>
<td>6A.6</td>
</tr>
<tr>
<td>Mata, J.</td>
<td>3P.20</td>
</tr>
<tr>
<td>Mathews, V.J.</td>
<td>7B.1</td>
</tr>
<tr>
<td>Matsuda, I.</td>
<td>2B.8</td>
</tr>
<tr>
<td>Mattavelli, M.</td>
<td>4B.9</td>
</tr>
<tr>
<td>Matteucci, P.</td>
<td>6P.1</td>
</tr>
<tr>
<td>Mayrargue, S.</td>
<td>4C.1</td>
</tr>
<tr>
<td>Mboup, M.</td>
<td>4C.10</td>
</tr>
<tr>
<td>McAulay, R..</td>
<td>3A.7</td>
</tr>
<tr>
<td>McCanny, J.v.</td>
<td>8E Chair</td>
</tr>
<tr>
<td>McDonnell, J.T.E.</td>
<td>1D Chair</td>
</tr>
<tr>
<td>*</td>
<td>1D.2</td>
</tr>
<tr>
<td>*</td>
<td>1D.6</td>
</tr>
<tr>
<td>McElroy, C.</td>
<td>5A.2</td>
</tr>
<tr>
<td>McLaughlin, S.</td>
<td>1C Chair</td>
</tr>
<tr>
<td>*</td>
<td>2C.9</td>
</tr>
<tr>
<td>*</td>
<td>5P.10</td>
</tr>
<tr>
<td>McWhirter, J.G.</td>
<td>6C.1</td>
</tr>
<tr>
<td>Medouri, A.</td>
<td>5P.13</td>
</tr>
<tr>
<td>Megias, J.A.</td>
<td>3D.10</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Mekuria, F.</td>
<td>6A.5</td>
</tr>
<tr>
<td>Menardi, P.</td>
<td>7A.2</td>
</tr>
<tr>
<td>Menendez, J.M.</td>
<td>7E.7</td>
</tr>
<tr>
<td>Menendez-Pidal, X.</td>
<td>2A.2</td>
</tr>
<tr>
<td>Menez, J.</td>
<td>5C.4</td>
</tr>
<tr>
<td>Mertins, A.</td>
<td>5D.5</td>
</tr>
<tr>
<td>Messer, H.</td>
<td>4D.2</td>
</tr>
<tr>
<td>Meyr, H.</td>
<td>7E.3</td>
</tr>
<tr>
<td>Meziane, A.</td>
<td>5P.11</td>
</tr>
<tr>
<td>Mian, G.A.</td>
<td>7A.2</td>
</tr>
<tr>
<td>Miani, R.</td>
<td>4E.3</td>
</tr>
<tr>
<td>Michel, O.</td>
<td>5P.4</td>
</tr>
<tr>
<td>Michell, J.A.</td>
<td>8E.8</td>
</tr>
<tr>
<td>Micollet, D.</td>
<td>4Pb.2</td>
</tr>
<tr>
<td>Miki, N.</td>
<td>8D.6</td>
</tr>
<tr>
<td>Mikkelsen, K.B.</td>
<td>5C.3</td>
</tr>
<tr>
<td>Milan, C.</td>
<td>4Pb.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>6E.8</td>
</tr>
<tr>
<td>Milosavljevic, M.</td>
<td>8C.6</td>
</tr>
<tr>
<td>Ming, J.</td>
<td>7P.2</td>
</tr>
<tr>
<td>Mironov, S.A.</td>
<td>7A.5</td>
</tr>
<tr>
<td>Mitra, S.K.</td>
<td>7B.2</td>
</tr>
<tr>
<td>Mizukami, K-I.</td>
<td>2P.19</td>
</tr>
<tr>
<td>Moccona, I.</td>
<td>4B.6</td>
</tr>
<tr>
<td>Moenclaey, M.</td>
<td>5C.10</td>
</tr>
<tr>
<td>&quot;</td>
<td>7C.10</td>
</tr>
<tr>
<td>Mohammad Beglou, M.</td>
<td>6P.18</td>
</tr>
<tr>
<td>Mohammad-Djafari, A.</td>
<td>5E.1</td>
</tr>
<tr>
<td>Moir, T.J.</td>
<td>6C.7</td>
</tr>
<tr>
<td>Moisan, E.</td>
<td>5D.8</td>
</tr>
<tr>
<td>Moissan, H.</td>
<td>3B.6</td>
</tr>
<tr>
<td>Moler, E.</td>
<td>6P.7</td>
</tr>
<tr>
<td>Molinaro, F.</td>
<td>6P.15</td>
</tr>
<tr>
<td>Moloney, C.R.</td>
<td>8B.7</td>
</tr>
<tr>
<td>Monden, Y.</td>
<td>5E.4</td>
</tr>
<tr>
<td>Montazeri, M.</td>
<td>4C.9</td>
</tr>
<tr>
<td>Moonen, M.</td>
<td>4D.8</td>
</tr>
<tr>
<td>&quot;</td>
<td>8E.1</td>
</tr>
<tr>
<td>Moratz, R.</td>
<td>7B.4</td>
</tr>
<tr>
<td>Moreau, E.</td>
<td>5P.17</td>
</tr>
<tr>
<td>Moreno, A.</td>
<td>6A.6</td>
</tr>
<tr>
<td>Mori, S.</td>
<td>7D.7</td>
</tr>
<tr>
<td>Moscheni, F.</td>
<td>4B.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>4B.6</td>
</tr>
<tr>
<td>Moschytz, G.S.</td>
<td>7B.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>7B.7</td>
</tr>
<tr>
<td>Mosquera, A.</td>
<td>4Pb.1</td>
</tr>
<tr>
<td>Moulines, E.</td>
<td>2E.9</td>
</tr>
<tr>
<td>&quot;</td>
<td>5C.8</td>
</tr>
<tr>
<td>&quot;</td>
<td>6C.9</td>
</tr>
<tr>
<td>Mureaux, J-M.</td>
<td>3P.6</td>
</tr>
<tr>
<td>Moustakides, G.V.</td>
<td>4C.5</td>
</tr>
<tr>
<td>Mayrati, M.</td>
<td>3A.8</td>
</tr>
<tr>
<td>Mueller, F.</td>
<td>5E.6</td>
</tr>
<tr>
<td>Mulgrew, B.</td>
<td>2P.15</td>
</tr>
<tr>
<td>&quot;</td>
<td>5P.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>61 Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>7A Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>8C.10</td>
</tr>
<tr>
<td>Mumolo, E.</td>
<td>3A.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>3P.8</td>
</tr>
<tr>
<td>&quot;</td>
<td>4P.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>4Pa.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>6P.13</td>
</tr>
<tr>
<td>&quot;</td>
<td>7P.17</td>
</tr>
<tr>
<td>&quot;</td>
<td>7P.18</td>
</tr>
<tr>
<td>Munoz, O.</td>
<td>7C.7</td>
</tr>
<tr>
<td>Munoz-Munoz, S.</td>
<td>4E.10</td>
</tr>
<tr>
<td>Murgia, C.</td>
<td>7P.15</td>
</tr>
<tr>
<td>Murphy, M.</td>
<td>7E.8</td>
</tr>
<tr>
<td>Murray, B.</td>
<td>5A.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>5A.4</td>
</tr>
<tr>
<td>Muruzabal, J.</td>
<td>4E.7</td>
</tr>
<tr>
<td>Nagai, N.</td>
<td>8D.6</td>
</tr>
<tr>
<td>Nagamatsu, M.</td>
<td>5E.4</td>
</tr>
<tr>
<td>Ngy, P.A.</td>
<td>5C.9</td>
</tr>
<tr>
<td>Nahamoo, D.</td>
<td>2A.7</td>
</tr>
<tr>
<td>&quot;</td>
<td>6P.17</td>
</tr>
<tr>
<td>Najim, M.</td>
<td>6E.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>8C.4</td>
</tr>
<tr>
<td>Nandi, A.K.</td>
<td>1C.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>2C Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>2C.7</td>
</tr>
<tr>
<td>&quot;</td>
<td>3B.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>3C.6</td>
</tr>
<tr>
<td>&quot;</td>
<td>8D.8</td>
</tr>
<tr>
<td>Napolitano, A.</td>
<td>5P.8</td>
</tr>
<tr>
<td>Nathan, K.S.</td>
<td>6P.17</td>
</tr>
<tr>
<td>Navarro, J.</td>
<td>6P.4</td>
</tr>
<tr>
<td>Navarro, R.</td>
<td>8B.10</td>
</tr>
<tr>
<td>Neagoe, V.E.</td>
<td>8A.6</td>
</tr>
<tr>
<td>Negrete-Regagnon, P.</td>
<td>3C.4</td>
</tr>
<tr>
<td>Neuvo, Y.</td>
<td>4P.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>7D.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>8B.6</td>
</tr>
<tr>
<td>Nguyen-Thi, H.L.</td>
<td>5P.16</td>
</tr>
<tr>
<td>Nicol, N.</td>
<td>7P.20</td>
</tr>
<tr>
<td>Nicolas, H.</td>
<td>4B.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>4B.6</td>
</tr>
<tr>
<td>&quot;</td>
<td>4B.9</td>
</tr>
<tr>
<td>Niedzwiecki, M.</td>
<td>6C.10</td>
</tr>
<tr>
<td>Nielsen, H.</td>
<td>5C.3</td>
</tr>
<tr>
<td>Niemann, H.</td>
<td>4B.10</td>
</tr>
<tr>
<td>Nieto, V.</td>
<td>4E.7</td>
</tr>
<tr>
<td>Nikolaidis, N.</td>
<td>5B.3</td>
</tr>
<tr>
<td>Nikolova, M.</td>
<td>5E.1</td>
</tr>
<tr>
<td>Nixon, M.S.</td>
<td>3B.7</td>
</tr>
<tr>
<td>&quot;</td>
<td>6P.9</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Nobili, L.</td>
<td>3P.3</td>
</tr>
<tr>
<td>Neelle, M.</td>
<td>7E.10</td>
</tr>
<tr>
<td>Noll, P.</td>
<td>6A Chair</td>
</tr>
<tr>
<td>Odobez, J.M.</td>
<td>3B.1</td>
</tr>
<tr>
<td>Oficjalska, B.</td>
<td>1D.4</td>
</tr>
<tr>
<td>Ogawa, H.</td>
<td>2P.5</td>
</tr>
<tr>
<td>Ohmori, H.</td>
<td>6C.5</td>
</tr>
<tr>
<td>Olbrich, T.</td>
<td>8E.6</td>
</tr>
<tr>
<td>Olivier, C.</td>
<td>8D.9</td>
</tr>
<tr>
<td>Olmo, G.</td>
<td>2D.5</td>
</tr>
<tr>
<td>Orn, R.</td>
<td>2E.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>2P.11</td>
</tr>
<tr>
<td>Orno, P.</td>
<td>6B.9</td>
</tr>
<tr>
<td>Orea, J.</td>
<td>6D.2</td>
</tr>
<tr>
<td>Orphanoudakis, S.</td>
<td>7B.8</td>
</tr>
<tr>
<td>Ottersten, B.</td>
<td>4D Chair</td>
</tr>
<tr>
<td>Ouadou, M.</td>
<td>6E.3</td>
</tr>
<tr>
<td>Ouamri, A.</td>
<td>2P.5</td>
</tr>
<tr>
<td>Ouldmanmar, M.</td>
<td>2P.5</td>
</tr>
<tr>
<td>Paez-Borraillo, J.M.</td>
<td>6D.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>7C.3</td>
</tr>
<tr>
<td>Pages-Zamora, A.</td>
<td>7D.4</td>
</tr>
<tr>
<td>Paindavoine, M.</td>
<td>6E.8</td>
</tr>
<tr>
<td>Pan, W.E.</td>
<td>6P.19</td>
</tr>
<tr>
<td>Pantaleon-Prieto, C.J.</td>
<td>4C.7</td>
</tr>
<tr>
<td>Papadias, C.B.</td>
<td>8D.2</td>
</tr>
<tr>
<td>Pardas, M.</td>
<td>1B.1</td>
</tr>
<tr>
<td>Pardo, J.M.</td>
<td>2A.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>4Pb.1</td>
</tr>
<tr>
<td>Parish, D.J.</td>
<td>3P Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>3P.18</td>
</tr>
<tr>
<td>Park, Y-G.</td>
<td>6E.7</td>
</tr>
<tr>
<td>Parsons, A.T.</td>
<td>5P.7</td>
</tr>
<tr>
<td>Patel, D.I.</td>
<td>8E.7</td>
</tr>
<tr>
<td>Paul, A.</td>
<td>6E.10</td>
</tr>
<tr>
<td>Pauli, R.</td>
<td>7C.1</td>
</tr>
<tr>
<td>Puira, L.</td>
<td>4D.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>5P.8</td>
</tr>
<tr>
<td>Pellandini, F.</td>
<td>7P.9</td>
</tr>
<tr>
<td>&quot;</td>
<td>8E.4</td>
</tr>
<tr>
<td>Peller, S.</td>
<td>3A.9</td>
</tr>
<tr>
<td>Pelletier, J.L.</td>
<td>4Pb.2</td>
</tr>
<tr>
<td>Pena, A.S.</td>
<td>7A.3</td>
</tr>
<tr>
<td>Penedo, M.G.</td>
<td>4Pb.1</td>
</tr>
<tr>
<td>Peng, Y-N.</td>
<td>3E.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>3E.9</td>
</tr>
<tr>
<td>Perez-Alvarez, I.A.</td>
<td>7C.3</td>
</tr>
<tr>
<td>Perez-Neira, A.</td>
<td>6D.9</td>
</tr>
<tr>
<td>Perrier, A.</td>
<td>6C.9</td>
</tr>
<tr>
<td>Pesquet, J-C.</td>
<td>7D.6</td>
</tr>
<tr>
<td>Petresou, D.</td>
<td>2P.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>5B.2</td>
</tr>
<tr>
<td>Petrou, M.</td>
<td>3B.8</td>
</tr>
<tr>
<td>Peyre, J-L.</td>
<td>2C.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peyrin, F.</td>
<td>6B.7</td>
</tr>
<tr>
<td>Philips, W.</td>
<td>3P.11</td>
</tr>
<tr>
<td>Phillips, I.W.</td>
<td>3P.18</td>
</tr>
<tr>
<td>Piasco, J.M.</td>
<td>3D.5</td>
</tr>
<tr>
<td>Piazza, F.</td>
<td>8C.5</td>
</tr>
<tr>
<td>Pinelli, G.</td>
<td>3E.7</td>
</tr>
<tr>
<td>Pinnington, R.J.</td>
<td>5P.7</td>
</tr>
<tr>
<td>Pipon, F.</td>
<td>7C.2</td>
</tr>
<tr>
<td>Pirez, D.</td>
<td>7C.2</td>
</tr>
<tr>
<td>Pirie, E.</td>
<td>7A.8</td>
</tr>
<tr>
<td>Pitas, I.</td>
<td>3P.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>5B Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>5B.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>7B.6</td>
</tr>
<tr>
<td>Pliman, F.</td>
<td>5C.2</td>
</tr>
<tr>
<td>Poignet, P.</td>
<td>2E.7</td>
</tr>
<tr>
<td>Pointon, C.T.</td>
<td>6D.4</td>
</tr>
<tr>
<td>Pollet, T.</td>
<td>5C.10</td>
</tr>
<tr>
<td>&quot;</td>
<td>7C.10</td>
</tr>
<tr>
<td>Ponceot, A.</td>
<td>7B.3</td>
</tr>
<tr>
<td>Poo, G-S.</td>
<td>7P.4</td>
</tr>
<tr>
<td>Potamianos, A.</td>
<td>7P.11</td>
</tr>
<tr>
<td>Prado, J.</td>
<td>6C.9</td>
</tr>
<tr>
<td>Prastitis, L.A.</td>
<td>1E.3</td>
</tr>
<tr>
<td>Premkumar, A.B.</td>
<td>6E.4</td>
</tr>
<tr>
<td>Prochazka, A.</td>
<td>5E.5</td>
</tr>
<tr>
<td>Prost, R.</td>
<td>2P.4</td>
</tr>
<tr>
<td>Prouder, I.K.</td>
<td>6C Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>6C.1</td>
</tr>
<tr>
<td>Pye, C.J.</td>
<td>4P.7</td>
</tr>
<tr>
<td>Qian, Y.</td>
<td>5A.1</td>
</tr>
<tr>
<td>Qiu, L.</td>
<td>3D.6</td>
</tr>
<tr>
<td>Quarmby, D.</td>
<td>6E Chair</td>
</tr>
<tr>
<td>Quatieri, T.F.</td>
<td>3A.7</td>
</tr>
<tr>
<td>Queluz, P.</td>
<td>8B.3</td>
</tr>
<tr>
<td>Quiguer, T.</td>
<td>3B.5</td>
</tr>
<tr>
<td>Quinn, A.</td>
<td>8D.5</td>
</tr>
<tr>
<td>Ramponi, G.</td>
<td>3C.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>4B Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>5B.5</td>
</tr>
<tr>
<td>Ramstad, T.A.</td>
<td>2P.6</td>
</tr>
<tr>
<td>Ravera, B.</td>
<td>7P.12</td>
</tr>
<tr>
<td>Razaz, M.</td>
<td>2P.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>7E.5</td>
</tr>
<tr>
<td>Regalia, P.A.</td>
<td>4C.10</td>
</tr>
<tr>
<td>Regazzoni, C.S.</td>
<td>5B.9</td>
</tr>
<tr>
<td>&quot;</td>
<td>6P.1</td>
</tr>
<tr>
<td>Reillo, C.</td>
<td>8B.1</td>
</tr>
<tr>
<td>Reininger, H</td>
<td>2A.6</td>
</tr>
<tr>
<td>&quot;</td>
<td>5A.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>7P.19</td>
</tr>
<tr>
<td>&quot;</td>
<td>7P.20</td>
</tr>
<tr>
<td>Ren, J-B.</td>
<td>4E.1</td>
</tr>
<tr>
<td>Rendon, E.</td>
<td>7E.7</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Reusens, E.</td>
<td>2B.10</td>
</tr>
<tr>
<td>Rex, J.</td>
<td>8A.10</td>
</tr>
<tr>
<td>Riccardi, G.</td>
<td>7A.2</td>
</tr>
<tr>
<td>Ricci, G.</td>
<td>3E.1</td>
</tr>
<tr>
<td>Richard, A.</td>
<td>3D.9</td>
</tr>
<tr>
<td>Richardson, P.G.</td>
<td>6D.3</td>
</tr>
<tr>
<td>Rinaldo, R.</td>
<td>6B.10</td>
</tr>
<tr>
<td>Ris, C.</td>
<td>7B.5</td>
</tr>
<tr>
<td>Ritz, S.</td>
<td>7E.3</td>
</tr>
<tr>
<td>Rix, H.</td>
<td>1D.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>5D.7</td>
</tr>
<tr>
<td>Robaszkiewicz, M.</td>
<td>4Pb.6</td>
</tr>
<tr>
<td>Robert, M.</td>
<td>6E.8</td>
</tr>
<tr>
<td>Robert, T.</td>
<td>8A.4</td>
</tr>
<tr>
<td>Roberts, S.</td>
<td>1D.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>4E.6</td>
</tr>
<tr>
<td>Rodellar, V.</td>
<td>4E.7</td>
</tr>
<tr>
<td>Roerdink, J.B.T.</td>
<td>7B.10</td>
</tr>
<tr>
<td>Romanin, M.</td>
<td>7P.17</td>
</tr>
<tr>
<td>Romera, R.</td>
<td>4E.7</td>
</tr>
<tr>
<td>Rosa, A.C.</td>
<td>4Pb.8</td>
</tr>
<tr>
<td>Rossilie, D.</td>
<td>3C.3</td>
</tr>
<tr>
<td>Roux, C.</td>
<td>4Pb.6</td>
</tr>
<tr>
<td>Ruczewski, P.</td>
<td>3D.1</td>
</tr>
<tr>
<td>Ruiz, D.P.</td>
<td>5P.13</td>
</tr>
<tr>
<td>Ruiz, P.</td>
<td>5P.4</td>
</tr>
<tr>
<td>Rupprechert, W.</td>
<td>8E.9</td>
</tr>
<tr>
<td>Russo, F.</td>
<td>5B.5</td>
</tr>
<tr>
<td>Saito, T.</td>
<td>2P.16</td>
</tr>
<tr>
<td>&quot;</td>
<td>4B.7</td>
</tr>
<tr>
<td>Sakai, H.</td>
<td>6C.6</td>
</tr>
<tr>
<td>Sala, J.</td>
<td>4C.3</td>
</tr>
<tr>
<td>Salavedra, J.M.</td>
<td>6A.6</td>
</tr>
<tr>
<td>Salembier, P.</td>
<td>1B.1</td>
</tr>
<tr>
<td>Salgado, L.</td>
<td>7E.7</td>
</tr>
<tr>
<td>Sallent, S.</td>
<td>3P.20</td>
</tr>
<tr>
<td>Salvi, L.</td>
<td>7P.18</td>
</tr>
<tr>
<td>Sandau, R.</td>
<td>2B.7</td>
</tr>
<tr>
<td>Sandham, W.A.</td>
<td>2D Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>2D.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>4Pb.10</td>
</tr>
<tr>
<td>Sandler, M.</td>
<td>6E.10</td>
</tr>
<tr>
<td>Sankur, B.</td>
<td>2P.8</td>
</tr>
<tr>
<td>&quot;</td>
<td>4Pb.7</td>
</tr>
<tr>
<td>&quot;</td>
<td>5D.1</td>
</tr>
<tr>
<td>Sano, A.</td>
<td>6C.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>8D.4</td>
</tr>
<tr>
<td>Santamaria-Caballero, I.</td>
<td>5E.9</td>
</tr>
<tr>
<td>Sanz-Gonzalez, J.L.</td>
<td>4E.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>4E.10</td>
</tr>
<tr>
<td>&quot;</td>
<td>7D.10</td>
</tr>
<tr>
<td>Saraceno, C.</td>
<td>2B.9</td>
</tr>
<tr>
<td>Sauer-Greff, W.</td>
<td>2E.2</td>
</tr>
<tr>
<td>Sava, H.P.</td>
<td>1D.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sayrol, E.</td>
<td>5P.6</td>
</tr>
<tr>
<td>Seaife, R.</td>
<td>7P.16</td>
</tr>
<tr>
<td>Searano, G.</td>
<td>2C.2</td>
</tr>
<tr>
<td>Searth, S.</td>
<td>4Pb.10</td>
</tr>
<tr>
<td>Schauble, P.</td>
<td>2A.9</td>
</tr>
<tr>
<td>Schell, S.V.</td>
<td>4C.2</td>
</tr>
<tr>
<td>Schenkél, M.</td>
<td>7B.7</td>
</tr>
<tr>
<td>Schettini, R.</td>
<td>1B.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>6P.5</td>
</tr>
<tr>
<td>Schmidlin, V.</td>
<td>4E.8</td>
</tr>
<tr>
<td>Schmitz, A.</td>
<td>8A.9</td>
</tr>
<tr>
<td>Schneider, T.</td>
<td>6C.4</td>
</tr>
<tr>
<td>Schotten, H.D.</td>
<td>2P.13</td>
</tr>
<tr>
<td>Schroeter, P.</td>
<td>1B.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>4B.8</td>
</tr>
<tr>
<td>Schurer, H.</td>
<td>7A.6</td>
</tr>
<tr>
<td>Scott, I.</td>
<td>2P.15</td>
</tr>
<tr>
<td>Scotton, P.</td>
<td>5C.4</td>
</tr>
<tr>
<td>Seijas, J.</td>
<td>4E.2</td>
</tr>
<tr>
<td>Semyonova, N.Y.</td>
<td>1C.3</td>
</tr>
<tr>
<td>Serviere, C.</td>
<td>5D.10</td>
</tr>
<tr>
<td>Settineri, R.</td>
<td>8C.4</td>
</tr>
<tr>
<td>Severcan, M.</td>
<td>2B.1</td>
</tr>
<tr>
<td>Shafarenko, L.</td>
<td>3B.8</td>
</tr>
<tr>
<td>Shamshunder, S.</td>
<td>2C.5</td>
</tr>
<tr>
<td>Shang, C.</td>
<td>4E.9</td>
</tr>
<tr>
<td>Shani, R.</td>
<td>5C.2</td>
</tr>
<tr>
<td>Shankar, P.M.</td>
<td>3D.10</td>
</tr>
<tr>
<td>Shaw, M.</td>
<td>8E.6</td>
</tr>
<tr>
<td>Shaw, P.J.</td>
<td>2P.3</td>
</tr>
<tr>
<td>Sherbet, G.V.</td>
<td>4Pb.4</td>
</tr>
<tr>
<td>Shevgaonkar, R.K.</td>
<td>6P.20</td>
</tr>
<tr>
<td>Shinohara, Y.</td>
<td>6C.5</td>
</tr>
<tr>
<td>Sibille, P.</td>
<td>3D.9</td>
</tr>
<tr>
<td>Sicuranzo, G.L.</td>
<td>7B Chair</td>
</tr>
<tr>
<td>Sidorovitch, D.</td>
<td>2P.18</td>
</tr>
<tr>
<td>Simon, S.</td>
<td>3P.1</td>
</tr>
<tr>
<td>Siohan, P.</td>
<td>2P.7</td>
</tr>
<tr>
<td>Sitaram, R.N.V.</td>
<td>2A.10</td>
</tr>
<tr>
<td>Skodras, A.N.</td>
<td>3P.13</td>
</tr>
<tr>
<td>Skoglund, J.</td>
<td>1A.2</td>
</tr>
<tr>
<td>Stuck, D.T.M.</td>
<td>4C.8</td>
</tr>
<tr>
<td>&quot;</td>
<td>8D Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>8D.2</td>
</tr>
<tr>
<td>Shump, C.H.</td>
<td>7A.6</td>
</tr>
<tr>
<td>Smith, D.L.</td>
<td>4C.2</td>
</tr>
<tr>
<td>Smith, F.J.</td>
<td>7P.2</td>
</tr>
<tr>
<td>Soderstrom, T.</td>
<td>4A.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>4D.10</td>
</tr>
<tr>
<td>&quot;</td>
<td>6D.5</td>
</tr>
<tr>
<td>Sohn, C.</td>
<td>6E.7</td>
</tr>
<tr>
<td>Solaiman, B.</td>
<td>4Pb.6</td>
</tr>
<tr>
<td>Sommen, P.C.W.</td>
<td>5D.3</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------</td>
</tr>
<tr>
<td>Soraghan, J.J.</td>
<td>4Pb Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>4Pb.9</td>
</tr>
<tr>
<td>&quot;</td>
<td>5P.9</td>
</tr>
<tr>
<td>&quot;</td>
<td>6B.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>6B.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>8C.9</td>
</tr>
<tr>
<td>Sorensen, J.A.</td>
<td>6A.1</td>
</tr>
<tr>
<td>Sorhus, R.</td>
<td>6B.5</td>
</tr>
<tr>
<td>soulah, N.</td>
<td>8B.1</td>
</tr>
<tr>
<td>South, C.R.</td>
<td>8A Chair</td>
</tr>
<tr>
<td>Spaggiari, S.</td>
<td>6P.1</td>
</tr>
<tr>
<td>Speirs, C.A.</td>
<td>5P.9</td>
</tr>
<tr>
<td>Sreenivas, T.V.</td>
<td>2A.10</td>
</tr>
<tr>
<td>Srithar, G.</td>
<td>4E.4</td>
</tr>
<tr>
<td>Stanovic, L.</td>
<td>5P.5</td>
</tr>
<tr>
<td>Stanovic, S.</td>
<td>5P.5</td>
</tr>
<tr>
<td>Stasinski, R.</td>
<td>2P.6</td>
</tr>
<tr>
<td>&quot;</td>
<td>3B.3</td>
</tr>
<tr>
<td>Stathaki, P.T.</td>
<td>3C.8</td>
</tr>
<tr>
<td>Stefanoiu, D.</td>
<td>3A.8</td>
</tr>
<tr>
<td>Stewart, R.W.</td>
<td>5P.9</td>
</tr>
<tr>
<td>&quot;</td>
<td>7A.8</td>
</tr>
<tr>
<td>Stogioglou, A.G.</td>
<td>2C.9</td>
</tr>
<tr>
<td>Stoica, P.</td>
<td>4A.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>4D.10</td>
</tr>
<tr>
<td>&quot;</td>
<td>6D.5</td>
</tr>
<tr>
<td>Sureti, M.</td>
<td>1B.3</td>
</tr>
<tr>
<td>Sun, H.W.</td>
<td>8A.3</td>
</tr>
<tr>
<td>Sun, T.</td>
<td>8B.4</td>
</tr>
<tr>
<td>Sun, X.</td>
<td>8B.6</td>
</tr>
<tr>
<td>Surtees, D.P.</td>
<td>3E.4</td>
</tr>
<tr>
<td>Sweeney, D.</td>
<td>4Pb.4</td>
</tr>
<tr>
<td>Swierczynski, R.</td>
<td>7A.8</td>
</tr>
<tr>
<td>Sys, V.</td>
<td>6B.2</td>
</tr>
<tr>
<td>Szymanski, J.</td>
<td>5E.5</td>
</tr>
<tr>
<td>Tabas, I.</td>
<td>3D.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>2P.1</td>
</tr>
<tr>
<td>Taguchi, A.</td>
<td>5B.2</td>
</tr>
<tr>
<td>Takahashi, K.</td>
<td>8B.4</td>
</tr>
<tr>
<td>Takahashi, S-I.</td>
<td>7D.7</td>
</tr>
<tr>
<td>&quot;</td>
<td>2P.19</td>
</tr>
<tr>
<td>&quot;</td>
<td>2P.20</td>
</tr>
<tr>
<td>Tan, T.N.</td>
<td>6P.3</td>
</tr>
<tr>
<td>Tanik, Y.</td>
<td>4D.9</td>
</tr>
<tr>
<td>Tarassenko, L.</td>
<td>4E.6</td>
</tr>
<tr>
<td>Taylor, D.</td>
<td>8E.5</td>
</tr>
<tr>
<td>Teixeira, A.</td>
<td>4E.5</td>
</tr>
<tr>
<td>Terre, M.</td>
<td>6C.2</td>
</tr>
<tr>
<td>Theodoridis, S.</td>
<td>4C.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>5E Chair</td>
</tr>
<tr>
<td>&quot;</td>
<td>7C.6</td>
</tr>
<tr>
<td>Theys, C.</td>
<td>5P.14</td>
</tr>
<tr>
<td>Thibaut, P.</td>
<td>3E.6</td>
</tr>
<tr>
<td>Thiel, J.</td>
<td>7E.1</td>
</tr>
<tr>
<td>Name</td>
<td>Paper</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Vincent, I. *</td>
<td>3D.2</td>
</tr>
<tr>
<td>Vlahakis, M.A.*</td>
<td>6P.16</td>
</tr>
<tr>
<td>Volant, P.</td>
<td>1A.1</td>
</tr>
<tr>
<td>Voutekelatos, S.P.</td>
<td>2D.6</td>
</tr>
<tr>
<td>Wada, S.</td>
<td>1A.5</td>
</tr>
<tr>
<td>Wadehra, V.</td>
<td>6B.4</td>
</tr>
<tr>
<td>Wagner, M.</td>
<td>2P.9</td>
</tr>
<tr>
<td>Wang, X.</td>
<td>4Pb.4</td>
</tr>
<tr>
<td>Ward, S.</td>
<td>1P.11</td>
</tr>
<tr>
<td>Warner, E.S.</td>
<td>6A.4</td>
</tr>
<tr>
<td>Watanabe, S.</td>
<td>8B.7</td>
</tr>
<tr>
<td>Weber, J.H.</td>
<td>5P.1</td>
</tr>
<tr>
<td>Wen, W.</td>
<td>2P.16</td>
</tr>
<tr>
<td>Wensink, H.E.</td>
<td>6B.6</td>
</tr>
<tr>
<td>Westall, F.</td>
<td>8D.7</td>
</tr>
<tr>
<td>Westall, F.</td>
<td>8B.2</td>
</tr>
<tr>
<td>White, P.R.</td>
<td>8D.7</td>
</tr>
<tr>
<td>Wilde, D.</td>
<td>7C. Chair</td>
</tr>
<tr>
<td>Wilinski, P.</td>
<td>3A. Chair</td>
</tr>
<tr>
<td>Wilkinson, R.</td>
<td>5P.7</td>
</tr>
<tr>
<td>Wolf, D.</td>
<td>6E.9</td>
</tr>
<tr>
<td>&quot;</td>
<td>4Pb.6</td>
</tr>
<tr>
<td>&quot;</td>
<td>4Pb.10</td>
</tr>
<tr>
<td>&quot;</td>
<td>2A.6</td>
</tr>
<tr>
<td>&quot;</td>
<td>5A.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>7P.19</td>
</tr>
<tr>
<td>&quot;</td>
<td>7P.20</td>
</tr>
<tr>
<td>Woodham, C.A.</td>
<td>2D.3</td>
</tr>
<tr>
<td>Woodhams, F.W.D.</td>
<td>2A.6</td>
</tr>
<tr>
<td>Wuest, H.</td>
<td>5C.2</td>
</tr>
<tr>
<td>Wulich, D.</td>
<td>8D.4</td>
</tr>
<tr>
<td>Xin, J.</td>
<td>3P.2</td>
</tr>
<tr>
<td>Xu, C.</td>
<td>7D.1</td>
</tr>
<tr>
<td>Xydeas, C.S.</td>
<td>7D.3</td>
</tr>
<tr>
<td>Yamazaki, O.</td>
<td>6A.7</td>
</tr>
<tr>
<td>Yang, B.</td>
<td>2A.1</td>
</tr>
<tr>
<td>Yang, G.</td>
<td>5C.1</td>
</tr>
<tr>
<td>Yang, R.</td>
<td>6A.7</td>
</tr>
<tr>
<td>&quot;</td>
<td>2A.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>4A.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>4A.8</td>
</tr>
<tr>
<td>&quot;</td>
<td>3P.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>7D.1</td>
</tr>
<tr>
<td>&quot;</td>
<td>1B.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>7D.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>3A.10</td>
</tr>
<tr>
<td>&quot;</td>
<td>8A.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>6P.8</td>
</tr>
<tr>
<td>&quot;</td>
<td>8B.5</td>
</tr>
<tr>
<td>&quot;</td>
<td>7A.4</td>
</tr>
<tr>
<td>&quot;</td>
<td>7B.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>6P.2</td>
</tr>
<tr>
<td>&quot;</td>
<td>1C.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>3P.20</td>
</tr>
<tr>
<td>&quot;</td>
<td>7C.3</td>
</tr>
<tr>
<td>&quot;</td>
<td>2P.2</td>
</tr>
</tbody>
</table>

Zhang, X-D.          | 3C.10 |
Zhang, Y.            | 3E.5  |
Zielinski, T.P.      | 4A.5  |
Zinke, J.            | 7P.20 |
Zivojnovic, V.       | 7E.3  |
Zyoute, M.           | 6E.3  |
d'Alessandro, C.     | 7P.12 |
da Silva, E.A.B.     | 6B.3  |
den Brinker, A.C.    | 5E.7  |
den Dekker, A..      | 6P.14 |
vander Kolk, A.       | 3P.20 |
vanden Brander       | 4A.6  |
Advances in Video Technology

H. Gharavi

Dept. of Electronic and Electrical Eng., University of Technology
Loughborough, Leicestershire, LE 11 3TU, England, UK,
Tel/Fax: +44 (0)15 222 830/222 891, Email: H.Gharavi@lut.ac.uk

Abstract. This paper presents an overview of the most recent technological advances in video technologies. Two important aspects, transportation medium and compression have been the main focus. As far as transportation is concerned, the emphasis has been on residential broadband access via copper wires, coaxial cable, and wireless. As a result technologies such as Asymmetrical Digital Subscriber Line (ADSL), fiber/coax upgrades, and wireless transmission with their associated architectures have been discussed. Finally, a brief overview of recent video coding standards is presented.

1. Introduction
In the past two decades digital video technology has undergone tremendous advancement in both compression and transportation. With the emerging digital network technology, visual communications is considered an unparalleled revenue driven service by major cable operators and telephone companies. Video, with applications in entertainment, education, business, and its unlimited quality representation nature, is pushing all transport mediums to their limits. Compressed video is expected to be transmitted at bandwidths as low as tens of kb/s up to hundreds of Mb/s. Despite significant growth in digital compression technology, future video communications will face new problems mainly due to its diversity in terms of application and the transportation medium in which they will be transmitted. For example, digitally compressed video is expected to be transported via wireless, fiber, coax, and copper wires for applications such as entertainment, personal communications, and teleconferencing with a point to point and multipoint to multipoint connections. With regards to service quality, it should cover a wide range from small-screen/low-bandwidth for portable video to large-screen/high-bandwidth HDTV or possibly super HDTV. This, together with recent coding standards such as H.261 [1] for video telephone and MPEG [2] for storage medium, are examples of such diverse service applications. As video communications continue to advance the integration and distribution of various video services will be a new challenge for future video telecommunications. In addition, the transportation of video over public networks has already added new dimensions in how and when video compression should be deployed. To enhance our understanding of the overall spectrum of future trends in video communication, we will summarize the most recent advances in transport as well as compression technologies.

2. Video Transport Technology
Affordability is becoming a key word for major players in the competition to bring broadband residential access primarily for video applications. In the US, competition between telephone companies, cable operators and wireless has intensified as a result of a recent regulatory policy in 1992 [3] which permitted the telephone companies to offer video-based information. The new ruling provided them with a unique opportunity to become a common carrier to all video providers by means of video dial tone (VDT). The race has already started to establish a dominance in the residential broadband market. The main motivation behind the telephone companies’ wish to enter video delivery is to share in the business of cable and movie
rentals which gross multi billions of dollars annually. At the same time cable operators, currently supporting 67% of US homes, are planning to expand their market very rapidly by the introduction of interactive digital video services to their operation. Having the advantage of a coax distribution backbone, they have instigated massive investments to upgrade their current distribution architectures so that they can offer services such as switched video, quasi video on demand, data and telephone. The following sub-sections describe some of the video transport strategies employed by the major competitors for home video delivery.

2.1 Video Via Copper Lines
For telephone companies the vision of video delivery is based on offering a wide range of interactive and customer controlled video services. While they believe that the broadband local exchange can support a video on demand (VOD) service, a bottleneck remains in the copper feeder; fiber to copper cannot create broadband access. Nevertheless, as an initial phase towards residential broadband access, the telephone companies are pursuing a plan to utilize the copper feeder and distributions by introducing a new technology called the Asymmetric Digital Subscriber Loop (ADSL). ADSL technology, is aimed at providing an affordable video dial tone (VDT) to residential areas using existing copper wires. As shown in Figure 1, it can provide downstream and upstream links of 1.5 Mb/s and 16 kb/s, respectively. The uplink channel is for control information as well as an analogue voice channel. The 1.5 Mb/s downstream link is for transmission of VHS quality video applications using the MPEG [2] coding standard. ADSL will be supported by ATM broadband networks for video on demand (VOD) applications. Therefore, information can be retrieved from the information provider libraries and transmitted in 150 Mb/s bursts to the customer's local central office RAM in less than a second.

There are concerns however; i.e. the POTS subscribers further away than 18 kf from the central office (CO) switches will be unable to subscribe to the ADSL service. In addition, the quality of the VOD service (limited by the 1.5 Mb/s) will be insufficient for programs such as sporting events. Furthermore, all television receivers have to be tuned into the same program.

The above concerns are expected to be largely overcome by the introduction of a next generation ADSL (called ADSL-III) which is currently in the planning stage. The ADSL-III is expected to enhance the service capacity to 6 Mb/s at the expense of reducing copper distance to less than 6kf. With its 6 Mb/s capacity, the ADSL-III can deliver up to two broadcast quality channels for live sporting events or four VCR quality channels for home movies. The main obstacle with the ADSL service provisioning is that it needs a data base

![Figure 1. Asymmetrical Digital Subscriber Line System](image)

630
and operational support systems which may
give cable operators an upper hand in early
market positioning by providing quasi video on
demand services to the subscribers.

2.2 Fiber/Coax Alternatives
Fiber/Coax systems are currently considered
the most cost effective alternative in providing
broadband access to residential areas. There
are various architectures which generally focus
on the fiber to the curb (FTTC) configuration
[4] via the “Fiber In The Loop” (FITL) system.
The deployment of the narrowband FITL
system is intended not only to deliver the
existing telecommunication services but also to
accommodate future upgrades for
broadband/video services to residential areas
(i.e. “Fiber To The Home”, FTTH) [5]. With
FITL upgrades it is possible to cover a range
of per subscriber bitrates from DS1-rate (1.5
Mb/s) to OC-3-rate (150 Mb/s).
In the all-digital fiber/coax architecture the
fiber feeder extends from the central office to
remote distribution nodes consisting of a
remote digital terminal and serving area
interface. From the feeder to the distribution
area, the signal is carried via bidirectional
coaxial bus to and from the subscribers. Digital
signals such as SONET OC-12 are transported
at the baseband level when travelling through
fiber. It is then digitally modulated before
being lunched on coaxial bus.
In the analogue/digital system (also known as
hybrid) a combination of both analogue and
digital video channels are included in the
transmission [6]. In the present proposed
hybrid upgrade FITL system [5], the network
can provide more than 100 channels including
basic, premium, pay per view, as well as
interactive video channels (including VOD).
The broadcast channels are transmitted in
analogue using the standard AM/VSB
modulation in accordance with current cable
ready television sets. The switched video
channels are digitally transmitted by employing
16-QAM. The carrier spacing of 6 MHz is
assigned for both analogue and digital carriers.
For the digital channels the video is
compressed to 4Mb/s using the MPEG-2
coding standard [14]. This can allow four
time-division-multiplexed videos per digital carrier
which leaves an additional 2 Mb/s to be
assigned for control overhead and other
purposes. The upstream link is 16 kb/s per
video customer which is allocated for
signalling information. Using time division
multiple access the signalling information is
first transformed to the narrowband FITL in
the remote distribution node and is then
transmitted via fiber to the central office.
Contrary to the all-digital system, in which the
signal is transported through fiber at its
baseband, in the hybrid system an analogue
laser operating in its linear region is used to
transmit the modulated video signals. In each
distribution area the analogue optical signals,
after being converted to electrical signals
(O/E), are amplified for distribution over
coaxial networks

2.3 Video Via Wireless
Independent of recent progress towards
personal communication systems/networks
(PCS/PCN), digital wireless video distribution
is emerging as a new technology providing
broadband access to the public. As a result,
wireless video is becoming a formidable
alternative in providing entertainments and
interactive home video services. In particular,
we are referring to the emerging microwave
based technology systems capable of providing
video, data, and telephone services. The
Cellular Vision system, using microwave
technology called multichannel local
distribution services (MLDS) is one example
which can simultaneously transmit up to 50
video channels as well as data information
[16]. Its technology is based on cellular radio
which divides a residential area into small cells
where each cell is covered by a separate
transmitter but all operating at the same
frequency. As shown in Figure 2, this is
achieved by transmitting the signal using
horizontal and vertical polarization for the
neighbouring cells. The cross polarization has
been the main reason for the spectral efficiency
of the Cellular Vision system due to its full
frequency reuse. The system operates at the
28 GHz UHF band and is claimed by its
developer to be the most cost effective video
service alternative, particularly in unwired
residential areas. The major impact of the
Cellular Vision system is in providing switched
video as well as telephone services at a
relatively low cost. It is also claimed that the
system will soon be capable of offering HDTV

631
services.
Another wireless alternative is the Hughes communications Inc.'s system called DirecTv. This DBS system can provide over 150 video channels for standard channels with over 60 channels assigned for pay per view movies and sporting channels. The main drawback with DirecTv is the cost of purchasing the initial equipment consisting of satellite antenna, compact dealer box, and a universal remote [16].

2.3.1 Wireless Access
The emergence of personal communication systems/networks PCS/PCN is expected to bring a new challenge for person to person video, audio, and data communications with a broadband ATM backbone. Although there is not yet any international standard for PCS/PCN, the general consensus is that video telephony will be one of its viable service applications. For example, the coding of video/image signals at very low bitrates has already been initiated by the Moving Pictures Export Group under MPEG-4 activities for applications such as person to person communications. Currently, a major challenge for portable/mobile video transmission is how to protect a sensitive signal (such as video) against hostile multipath fading environments. Bear in mind that to comply with the current narrowband channel requirements, it is essential that the video is compressed at a very high rate. This will make the compressed video signal extremely vulnerable to channel errors. This issue has been addressed in recent publications [8,9] in which the transmission of video signals for portable applications was investigated over the Rayleigh fading channels. In this study, the subband coded video signal was divided into two layers of differing

![Figure 2. Cellular Vision System](image)

![Figure 3. Subchannel-Splitting / QAM video-system schematic](image)
Table 1. Specifications of four proposed over-the-air HDTV systems in the U.S.

<table>
<thead>
<tr>
<th>System</th>
<th>Aspect ratio</th>
<th>Bandwidth MHz</th>
<th>Sampling rate MHz</th>
<th>Interlace system</th>
<th>Frame rate</th>
<th>RF Modulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSC-HDTV AT&amp;T, Zenith</td>
<td>16:9</td>
<td>34 (L) 17 (C)</td>
<td>75.3</td>
<td>1:1</td>
<td>59.94</td>
<td>2-level 4-level VSB</td>
</tr>
<tr>
<td>ADC-TV NBC, Philips, Thompson, Sarnoff</td>
<td>16:9</td>
<td>24.5 (L) 12.25 (C)</td>
<td>54</td>
<td>2:1</td>
<td>29.97</td>
<td>Spectrally Shaped QAM</td>
</tr>
<tr>
<td>CCDC-HDTV NBC, MIT, GI</td>
<td>16:9</td>
<td>34 (L) 17 (C)</td>
<td>75.5</td>
<td>1:1</td>
<td>59.94</td>
<td>16 QAM or 32 QAM</td>
</tr>
<tr>
<td>DigitCipher MIT, GI</td>
<td>16:9</td>
<td>21.5 (L) 5.4 (C)</td>
<td>53.65</td>
<td>2:1</td>
<td>29.97</td>
<td>16 QAM or 32 QAM</td>
</tr>
</tbody>
</table>

intelligences. As shown in Figure 3, the first and second layers containing the basic and contributory information were sent through subchannels C1 and C2 of a 4-bit interleaved 16 QAM. The bit interleaving provided the two most significant bits (signified as subchannel C1) with a far better protection against error than the two least significant bits (i.e., subchannel C2).

There are, however, other important issues of wireless which remain to be addressed before wireless hand held video telephony can be accepted by industry as a viable service technology. The most important issue should be compatibility with existing digital video telephony standards such as H.261 to make it a part of the integrated video telephony.

2.4 HDTV Transmission

High Definition Television (HDTV) transmission has been the center of worldwide debate since the early eighties. The main issue was the form in which HDTV signals should be transmitted. The first HDTV system, intended for direct broadcast satellite, was introduced by NHK of Japan. The system, due to its excessive bandwidth of 30 MHz, was modified and introduced in 1984 as a new system called MUSE [10]. The new system had a bandwidth reduced by a factor of four and was mainly developed for transmission over a satellite channel and for studio recording. Since then a family of MUSE systems has been introduced to deal with the compatibility and transmission aspects of HDTV signals [11]. The NHK system, which was initially backed by the US, was never received favourably by the Europeans who decided to launch their own program under the Eureka project called HDMAC [12]. The U.S., however, after a long debate, decided to have an all digital over-the-air simulcast HDTV system. Subsequently, following governmental guidelines, an alliance consisting of four groups of proponents was formed to compete with each other; the winner "taking all" for the best proposed system. Having four proposals on the table, the proponents were engaged in developing their prototype systems for the final shoot out test held in January, 1994. Although it is still not clear which system will be the winner, it seems that the Zenith-AT&T system, based on vestigial side band (VSB) transmission, has the winning edge [17]. With regards to the video compression aspect of the proposals, they are all based on hybrid DPCM/DCT structures similar to the MPEG-2 standard. Table I summarizes the important attributes of the four proposed HDTV systems. Europeans after failed efforts in continuing support for HD-MAC, have now refocused on terrestrial digital broadcasting instead of satellite for HDTV application. In terms of transmission, it seems that they are strongly in favour of Coded Orthogonal Frequency Division Multiplexing (COFDM).

3. Compression Techniques

Video compression has reached new heights in its technology after more than two decades of active research. It is now recognized that digital video will be the engine of the driving force in providing broadband access to consumer homes. A sudden rise in the number of video/image coding standards and availability of various transportation media options (see section 2) is a clear indication of a rapid growth in the visual communication industry.

While it is believed that standards can facilitate industrial growth, conversely it will slow down the pace of further expansion for many years to
come. For instance, continuing advances in video equipment technology will create new demands for higher quality video services now that broadband access will soon become a reality. It should therefore be expected that future video telecommunications will face a new dilemma in the interoperability between various digital video source formats and coding standards. To examine this briefly, the following provides a general overview of existing video coding standards.

Recently there have been a number of video coding standards developed for differing applications. Despite the fact that all of them are based on the same coding structure, compatibility and future expandability among them have not been fully resolved.

The first video coding standard, which led the way for more standardization, was the CCITT H.261. This standard was primarily developed for video teleconferencing applications in support of narrowband ISDN with a bitrate for combined audio and video in a range of 564 kb/s where p=1,2,...,32. P=1 is suitable for limited face to face communications using the ISDN basic access which is often referred to as videophone. However, higher quality video conferencing can also be provided by the support of the ISDN primary access (i.e. p=30) [15].

The video compression technique adopted for CCITT H.261 is a hybrid Transform/DPCM, (see Figure 4), where DPCM exploits the interframe redundancies. Referring to Figure 4, the incoming frame is divided into macroblocks of 16x16 for motion estimation where, after motion compensation, the prediction is further divided into four blocks of 8x8 for transform coding. The transform coefficients are quantized, entropy coded and together with the overhead information are sent to the buffer for transmission.

Almost four years have passed since the CCITT H.261 was recommended and despite tremendously high expectations and a drastic reduction in cost, videotelephony has not yet managed to position itself in the market place as a popular service/product. Nevertheless, its initial impact was to introduce another standard based on the same coding principal but with different applications. The new standard, which emerged from the activities of the International Standards Organization (ISO) known as Moving Picture Expert Group (MPEG), was targeted at video stored on CD ROM (i.e. CD-I) [2]. It uses the same compression technique as the H.261 (Hybrid Transform/DPCM) but with some added features such as fast-forward/reverse capabilities. This was achieved by incorporating an additional interframe mode called bidirectional interpolation. Figure 5 shows the overall coding modes consisting of intraframe (I), interframe prediction (P), and bidirectional interpolation (B). The bitrate of the MPEG-I standard with the compressed high quality audio may not exceed the 1.5 Mb/s of today's digital storage media.

For further expansion on applications and data rate, a second phase of standardization was followed under MPEG-II standard. The MPEG-2 [14], was drafted at the end of last year, and is an extension of the MPEG-I by an explicit way of handling video interface.

4. Final Remarks

The main thrust of this paper was to illustrate how visual communications are becoming the enabling technology for matching network transport capabilities to service-driven needs. Digital video and digital optical fibers are two emerging technologies which are becoming ideal complements for providing a broad range of applications such as video entertainment, video on demand, video on the desk top, interactive video, video games, interactive distance learning, etc.

In this paper the technological advances related to both the transportation and compression aspects of video were discussed. It has been shown how the telecommunication industry is gearing towards broadband digital networks by means of providing video-based services to residential areas. While the vision of future broadband networks remains that of fiber backbone, the challenge of the so-called "last mile" remains. For affordable broadband access we discussed various transport architectures which are based on copper, coax, and wireless technologies. These architectures are considered as interim steps towards the full implementation of "Fiber To The Home" technology. However, while the core technology is becoming feasible, there remains many issues for customer equipment to interoperate a variety of access network configurations. As far as video compression is
concerned, compatibility, expandability, and efficient bitrate sharing distribution will be the main tasks.

![Figure 5. Frame Coding Modes](image)

REFERENCES

Figure 4. Hybrid Transform/DPCM Encoder
High-Resolution Bearing Estimation by Fourier Methods

Jyrki Joutsensalo

Helsinki University of Technology, Laboratory of Computer and Information Science, Rakentajanaukio 2 C, FIN-02150 Espoo, Finland. e-mail: Jyrki.Joutsensalo@hut.fi, Fax: +358-0-4513277

Abstract. Eigendecomposition-based methods, especially Multiple Signal Classification (MUSIC) may be computationally too demanding in practice, and their performance degrades clearly when signal-to-noise ratios are small. We have recently introduced Fourier-based methods for estimating the signal subspace; these possess several advantages over standard MUSIC. In this work, these methods are improved and generalized into several directions. In some cases the Fourier-based Pisarenko method works approximately as well as the MUSIC, or even better.

1 Introduction

Directions-of-arrival (DOAs) estimation is a problem arising naturally in several application areas such as radar, sonar, or seismology. Eigendecomposition-based techniques, especially the MUSIC method [8, 9] have been studied extensively over the past years in these problems. Unlike most classical Fourier-based approaches, these methods correctly exploit the underlying subspace formalism typically assumed for these problems. Consequently, they provide superior resolution when the number of sensors is small or the angles of arrival to be estimated are close to each other. However, the cost of estimating the eigenvectors of the data covariance matrix spanning the signal or noise subspace is sometimes too high, and MUSIC does not utilize a priori information on the data. In addition, the performance of MUSIC rapidly degrades when the signal-to-noise ratio (SNR) is small. We have recently studied Fourier-based approaches for approximating the signal subspace [1, 3, 4]. In this work, the methods are improved and extended in several directions. First, versions based on fourth order statistical information, especially on cumulants [6] are developed. Second, versions estimating one dimensional noise subspace are introduced. This leads to the use of Pisarenko method which is a simpler estimator than MUSIC. Although Pisarenko method usually yields worse estimates than MUSIC in context with the eigenvector approach, and is often unreliable or inefficient, this does not necessarily hold in context with DFT based methods. Third, multidimensional transforms utilizing very efficient recursive DFT [7] are presented. These new methods are discussed more thoroughly in [2] where we present mathematical analysis and derive the fast algorithm for computing the signal subspace estimate.

2 Data model

Let us assume that the sensors of the array are uniformly spaced. Then the data value \( x_k(n) \) measured by the sensor \( k, k = 0, \ldots, L-1 \) at the snapshot \( n, n = 0, \ldots, N-1 \) can be represented as

\[
x_k(n) = \sum_{m=1}^{M} a_m(n)e^{j\omega_m} + w_k(n),
\]

where \( M, L (L > M) \), and \( N \) denote the number of signals, sensors, and snapshots, respectively. Amplitudes \( a_m(n) \) are here assumed to be independent non-Gaussian random variables. \( \omega_m \)'s are distinct angular frequencies normalized to the interval \([\pi, -\pi] \), and they correspond to the \( M \) directions. \( \omega_m \)'s depend on the known center frequency of the narrowband signals, sensor spacing, the plane-wave propagation velocity, and the unknown angles of arrival. If the estimates of \( \omega_m \)'s are available, one can straightforwardly calculate the corresponding estimates of the angles of arrival. The noise \( w_k(n) \) is either white or colored Gaussian. If the noise is white, then the autocorrelation function of the process has the form

\[
r(t) \triangleq E\{x_k(n)x_{k+\tau}(n)\} = \sum_{m=1}^{M} \beta_m e^{j\omega_m \tau} + \sigma^2 \delta(t),
\]

where \( \beta_m = E\{|a_m(n)|^2\} \), \( \sigma^2 \) is the variance of the noise, and \( \delta(\cdot) \) is the Kronecker delta. * denotes complex conjugate. In this case, the traditional eigenvector methods, such as MUSIC can be used to estimate the directions-of-arrival [8, 9]. Standard MUSIC procedure is performed as follows:

1. Construct the data vectors

\[
x(n) = [x_0(n) \quad x_1(n) \quad \cdots \quad x_{L-1}(n)]^T,
\]

637
2. Estimate the autocorrelation matrix
\[ \hat{R} = \frac{1}{N} \sum_{n=0}^{N-1} x(n)x(n)^H. \] (4) 
\( H \) denotes complex conjugate transpose.

3. Estimate the \( M \) first principal eigenvectors (corresponding to \( M \) largest eigenvalues) \( \hat{u}_i, i = 1, \ldots, M \) of \( \hat{R} \) and substitute them to the estimator
\[ P_{\text{MUSIC}}(\omega) = \frac{1}{L - \sum_{i=1}^{M} |e(\omega)^H \hat{u}_i|^2}, \] (5)
where \( e(\omega) = (1, e^{-i\omega}, \ldots, e^{(L-1)i\omega})^T \).

If the noise is Gaussian and colored, the correlation of the noise part is no more \( \sigma^2\delta(t) \), and the model becomes incorrect for eigenvector type methods. This drawback can be avoided by "smoothing" the data by fourth order cumulants, which are blind to Gaussian processes. Thus the noise can be filtered out, at least in theory, whereas the non-Gaussian signal part remains. The fourth order contracted quadricovariance \[ c(t) \triangleq \sum_{i=0}^{L-1} \left[ \mathbb{E}[x_k(n)x_{k+i}(n)]x_i(n)^2 \right] \]
\[ - \mathbb{E}[x_k(n)x_i(n)]\mathbb{E}[x_k(n)x_{i+k}(n)] \]
\[ - \mathbb{E}[x_k(n)x_{i+k}(n)]\mathbb{E}[x_i(n)^2] \], (6)
which results from the matrix notation of the contracted quadricovariance matrix \[ C = \mathbb{E}[|x(n)|^2 x(n) x(n)^H] - R^2 - \text{trace}(R)R. \] (7)
The contracted quadricovariance is like a noise-free covariance [5] having the expression
\[ c(t) = \sum_{m=1}^{M} \gamma_m e^{j\omega_m t}, \] (8)
where \( \gamma_m \) are real numbers. The coefficients can be used in context with MUSIC method by estimating the principal eigenvectors of the matrix \( \hat{C} \) which has theoretically the form
\[ C = \begin{bmatrix} c(0) & c(-1) & \cdots & c(-L+1) \\
 c(1) & c(0) & \cdots & c(-L+2) \\
 \vdots & \vdots & \ddots & \vdots \\
 c(L-1) & c(L-2) & \cdots & c(0) \end{bmatrix}. \] (9)

3 New Fourier-based methods

In the new DFT-based algorithms the frequencies \( \omega_m, m = 1, \ldots, M \) are first estimated crudely by using some simple classical method, e.g. Bartlett method. Let the test frequencies \( \hat{\omega}_m, m = 1, \ldots, M \) denote these crude estimates. The \( d \)-dimensional array to be transformed is denoted by \( y(n_1, \ldots, n_d) \), and it is defined in terms of the second or fourth order statistics of the data:
\[ y(n_1, \ldots, n_d) = y\left( \sum_{k=1}^{d} n_k \right) \triangleq \begin{bmatrix} r(\sum_{k=1}^{d} n_k - L + 1) \\
r(\sum_{k=1}^{d} n_k - L + 1) \end{bmatrix}, \]
(10)
where \( 0 \leq \sum_{k=1}^{d} n_k \leq 2L - 2 \). The array \( y \) has a lot of redundancy since it is obtained from the scalar sequence. The correlation coefficients \( r(t) \) are estimated as follows:
\[ \hat{r}(t) = \frac{1}{(L-t)N} \sum_{n=0}^{N-1} \sum_{k=0}^{L-t-1} x_k^{*}(n)x_{k+t}(n), \] (11)
f\( t = 0, \ldots, L - 1 \) and \( \hat{r}(t) = \hat{r}^*(-t) \) for \( t = -L + 1, \ldots, -1 \). The cumulant coefficients are estimated by exploiting the equation (6): We do not represent it here due to the lack of space; see [2]. The Fourier transform corresponding to the frequency \( \omega \) is defined as follows:
\[ v_n(\omega) \triangleq \sum_{n_2=0}^{A_2-1} \cdots \sum_{n_d=0}^{A_d-1} y(n_1, \ldots, n_d)e^{-j\omega \sum_{k=2}^{d} n_k}, \] (12)
where \( \sum_{k=1}^{d} (A_k - 1) = 2L - 2 \). In practice the Fourier transform is two- or three-dimensional. The transform can be implemented by fast algorithm using recursive DFT and FFT [2]. It can also be interpreted as a weighted DFT [2]. The signal subspace is estimated by computing the terms \( v_n(\hat{\omega}_m), n_1 = 0, \ldots, A_1 - 1, \) \( p = 1, \ldots, M \), and grouping them to the matrix
\[ \hat{V} \triangleq \begin{bmatrix} v_{0}(\hat{\omega}_1) & \cdots & v_{0}(\hat{\omega}_M) \\
 \vdots & \ddots & \vdots \\
 v_{A_1-1}(\hat{\omega}_1) & \cdots & v_{A_1-1}(\hat{\omega}_M) \end{bmatrix}. \] (13)

Hence the signal subspace is estimated in terms of \( d \)-dimensional DFT. It can be shown, that under mild conditions, \( V \) spans theoretically exactly the true signal subspace [2].

If the MUSIC estimator is used in context with the DFT approach, the columns of the matrix \( V \) must be orthonormalized before substituting them to the estimator (5). Computationally simpler Pissarenko method can be used by selecting the size of \( V \) to be \( (M + 1) \times M \), i.e. \( A_1 = M + 1 \). Then one solves \( \rho \) from the equation
\[ V^H \rho = 0, \quad \rho_{M+1} = 1, \] (14)
and substitutes it to the Pissarenko estimator
\[ P_{\text{Pissarenko}}(\omega) \triangleq \frac{1}{|e(\omega)^H \rho|^2}. \] (15)

3.1 Summary of the new method

The general Fourier method is summarized as follows:
1. Construct the estimate of $y(n_1, \ldots, n_d)$ in Eq. (10) from the available autocorrelation coefficients or from the cumulant coefficients.

2. Estimate the test frequencies $\omega_1, \ldots, \omega_M$ crudely by using e.g. some classical method.

3. Choose the values for $d$ and $A_1, \ldots, A_d$, and estimate the signal subspace $\bar{V}$ in (13) by using the DFT (12).

4. If $A_1 = M + 1$, compute the noise subspace vector $\rho$ by using the formula (14), and substitute it into the Pisarenko estimator (15). Otherwise, orthonormalize $\bar{V}$, and use MUSIC (5).

4 Experimental results

Experiment 1. 256 snapshots are available from the 14-dimensional narrow-band spatial data samples (measured by 14 sensors) arriving from the normalized directions 0.50 and 0.52 ($\omega_1 = 2\pi \cdot 0.50, \omega_2 = 2\pi \cdot 0.52$). The noise process is white and Gaussian. 200 simulations were performed with different phase (and thus signal) and noise processes. The methods are:

1. Standard MUSIC based on the two 14-dimensional principal eigenvectors of the data covariance matrix $\hat{R} = \sum_{n=0}^{255} x(n)x(n)^H / 255$.

2. Two-dimensional DFT applied to the correlation coefficients:

$$v_{n_1}(\tilde{\omega}_p) = \sum_{n_2=0}^{24} \hat{r}(n_1 + n_2 - 13)e^{-j\tilde{\omega}_pn_2},$$

where $n_1 = 0, 1, 2, p = 2$. The resulting $3 \times 2$ matrix $V$ is used to solve $\rho$, and this vector is substituted to the Pisarenko estimate.

3. Three-dimensional DFT applied to the correlation coefficients:

$$v_{n_1}(\tilde{\omega}_p) = \sum_{n_3=0}^{6} \sum_{n_2=0}^{18} \hat{r}(n_1 + n_2 + n_3 - 13)e^{-j\tilde{\omega}_pn_3+n_3},$$

where $n_1 = 0, 1, 2, p = 1, 2$. Pisarenko method is applied.

Pisarenko and MUSIC spectra were evaluated at 2048 points. In Fig. 1, three mean-square error curves corresponding to these methods are represented. One can see that even if the Fourier methods use the Pisarenko estimator, they operate better than MUSIC at lower SNR's. A simple explanation is that if we select $A_1$ to be small, then we can select the other $A_i$'s (i.e. the size of the Fourier transform (12)) to be large. The minimum value of $A_1$ is $A_1 = M + 1$, which corresponds to the Pisarenko method. Notice that the three-dimensional DFT performs better than the two-dimensional DFT. At high SNR's, the Fourier methods do not perform well compared to the standard MUSIC. The reason is that the estimate $\hat{r}(\tilde{\omega})$ does not approach asymptotically the correct sinusoidal model when SNR increases, but in the standard MUSIC, the estimate $\hat{R}$ does. This problem can be overcome by estimating the coefficients in such a way that they obey pure sinusoidal model when no noise exists. In this case, the algorithms become more complicated, but they yield better results.

Experiment 2. Conditions are otherwise similar as in the previous experiment, but now 512 snapshots are available, and the colored Gaussian AR noise having the center frequency 0.15 (pole at 0.9 exp[±j2π · 15]) is disturbing the process. The methods are:

4. Standard MUSIC based on the two 14-dimensional principal eigenvectors of the contracted quadricovariance matrix $C$. It is obtained by replacing the expectations in the equation (7) by direct calculations with vectors.

5. Two-dimensional DFT applied to the cumulant coefficients:

$$v_{n_1}(\tilde{\omega}_p) = \sum_{n_3=0}^{24} \hat{c}(n_1 + n_2 - 13)e^{-j\tilde{\omega}_pn_3},$$

where $n_1 = 0, 1, 2, p = 2$. The resulting $3 \times 2$ matrix $V$ is used to solve $\rho$, and this vector is substituted to the Pisarenko estimate.

6. Three-dimensional DFT applied to the cumulant coefficients:

$$v_{n_1}(\tilde{\omega}_p) = \sum_{n_3=0}^{6} \sum_{n_2=0}^{18} \hat{c}(n_1 + n_2 + n_3 - 13)e^{-j\tilde{\omega}_pn_3+n_3},$$

where $n_1 = 0, 1, 2, p = 1, 2$. Pisarenko method is applied. The only difference between the methods 1-3 and 4-6 is in the selection of the coefficients. One can see from the Fig. 2 that the Fourier methods perform better than the standard MUSIC at low SNR's. At very low SNR's, the estimates given by all the methods are poor.
5 Discussion

Here, conclusions from the experiments and theoretical considerations are presented.

- Multidimensional arrays can be transformed in an efficient way to get the basis of the signal subspace. In this work, the arrays were constructed from the one-dimensional sequences to save the memory and computation, but the similar techniques can be used to contract e.g. $L \times L \times L \times L$ cumulant tensor containing all the fourth order statistical information to the set of vectors spanning the signal subspace.

- A fast algorithm has been developed for computing a special type of multidimensional transform. This result may be useful in other connections, too. It can also be interpreted as a special case of weighted transform

\[ v_{n_1} (\omega) = \sum_{t=0}^{2L-1-A_1} p_t y(n_1 + t)e^{-j\omega t}, \]

where $n_1 = 0, \ldots, A_1 - 1$, $p_t$ is some integer, and $t = \sum_{k=2}^{d} n_k$. When $d$ is small enough, e.g. $d = 2$ or $d = 3$, the algorithm based on the recursive DFT is faster than the computation (16) directly.

- It has been shown that when the Fourier transform is used for estimating the signal subspace, then Pisarenko method can – exceptionally – be successfully exploited, and it yields similar or even better results than the standard MUSIC. If eigenvectors are used, MUSIC is generally better than Pisarenko method. A simple explanation is that when we select the minimal size to the signal subspace i.e. $M + 1$, the size of the Fourier transformable data vector set has the maximal size.

- The Fourier method can also be used in context with the time series data [1, 2, 3, 4]. E.g. the pure data samples can be then transformed so that the algorithm becomes very simple.

- In high SNR's, the correlation- and cumulant-based Fourier methods do not perform well compared to the standard MUSIC. This problem can be overcome by estimating the coefficients in such a way that they obey pure sinusoidal model when no noise exists. In this case, the algorithms become slightly more complicated, but they yield better results.

6 Conclusions

In this work, efficient methods for estimating the signal subspace have been proposed. They are expansions and improvements of the methods introduced in [1, 3, 4]. In [2], the methods are investigated more thoroughly. The algorithms are simpler than the standard eigenvector approach due to the existence of the fast algorithm based on the recursive DFT and FFT [2, 7]. Perhaps the most surprising result in this study is that the Pisarenko method is reliable and efficient in context with Fourier transform.

References

A Robust and Efficient Algorithm for Tracking Source Locations

Frank GERSEMSKY, Bin YANG and J.F. BÖHME

Dept. of Electrical Eng., Ruhr University Bochum 44780 Bochum, Germany
Fax.: +49 234 709 4261 Tel.: +49 234 700 7685 E-Mail: ger@stth.ruhr-uni-bochum.de

Abstract. In this paper we address the problem of estimating source parameters from narrow-band signals impinging an array of sensors. A new criterion function based on the well known MUSIC spectrum estimator is proposed. We present a new algorithm for tracking source parameters which is based on this criterion function. This method works with an arbitrary but known antenna geometry and can also handle partly correlated signals. The performance of this algorithm is comparable or better than other well known location estimators especially in scenarios with low SNR and closely spaced sources. The computational complexity is only linear with respect to the number of sensors. The algorithm exhibits a highly regular structure that allows an implementation on parallel architectures as multi-processor computers or systolic arrays.

1 Introduction

Array processing algorithms for estimating source parameters such as distance, direction of arrival (DOA) or the signal frequency using an array of sensors are widely used in many fields (e.g. radar, sonar, seismic applications). A frequently used algorithm for estimating locations of narrowband sources in white noise is the MUSIC method. Its high computational cost restricts its use to applications where online processing is not required. We propose a new algorithm based on a new criterion function which exhibits a significantly reduced computational cost compared to batch high resolution methods as MUSIC or Min-Norm. Additionally, its performance especially in scenarios with closely spaced sources and/or low SNR is preferable compared to other parameter tracking algorithms (e.g. WSF [4]).

The first step in estimating the unknown source parameters is the computation of the so called signal subspace which is spanned by the first r dominant eigenvectors of the sample covariance matrix.

Many well performing and efficient algorithms for this task like PAST [5] and ROSA [2] have been developed in the past. These algorithms only require O(nr) operations, where n is the number of sensors and r the number of sources.

In this paper we focus on the second part of the present problem, the computation of source parameters from the signal subspace.

2 Problem Formulation

2.1 Signal model

Consider r narrow-band waves received by an array of n sensors (n > r) corrupted by spatially white noise\(^1\). Here we restrict ourselves to the case where only one parameter per source is to be estimated (e.g. direction). The extension to multiple parameters is straightforward. The base-band signal model reads

\[
x(t) = A(\theta)s(t) + n(t),
\]

(1)

with the complex amplitudes \(\mathbf{a} = [s_1, \ldots, s_r]^T\), the array response vectors \(A(\theta) = [a(\theta_1), \ldots, a(\theta_r)]\) and white noise \(n(t)\) with white noise power \(\sigma^2\). The vector \(a(\theta_i)\) describes the unit response of the array to a source with parameter \(\theta_i\).

The real parameter vector \(\theta\) contains the unknown source parameters. We do not impose any constraints on the geometry of the array, instead we assume the geometry to be known, i.e. the array is calibrated. Additionally, we make the assumption that the array is unambiguous, which means that the matrix \(A\) has full rank for any distinct set of parameters. The amplitudes \(s_i\) are assumed to be zero mean with second order moments

\[
E[s(t_1)s^H(t_2)] = C_s \delta_{t_1,t_2},
\]

(2)

\[
E[s(t_1)s^T(t_2)] = 0,
\]

(3)

\(^1\)In the case of non-white noise the signal can be prewhitened if the noise covariance matrix is known.
where $T$ and $H$ denote transposition and conjugate transposition, respectively.

With these assumptions, the signal covariance matrix $E[x x^H]$ is given by

$$
C_x = A(\theta) C_x A_H(\theta) + \sigma^2 I
$$

(4)

2.2 Signal Subspace

We now look at the eigen-decomposition of $C_x$:

$$
C_x = [u_1, \ldots, u_n] \begin{bmatrix}
\lambda_1 \\
\vdots \\
\lambda_n
\end{bmatrix} [u_1, \ldots, u_n]^H,
$$

with the eigenvalues $\lambda_i$ ($i = 1, \ldots, n$) sorted in non-decreasing order. The eigenvalues exhibit the following property

$$
\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_n.
$$

(5)

It follows that

$$
\text{range}(U_s) = \text{range}(A(\theta))
$$

(6)

holds, where the column vectors of $U_s = [u_1, \ldots, u_r]$ span the so-called signal subspace of $C_x$.

3 The New Criterion

The idea behind this new criterion function is the same as in MUSIC or WSF [3]. We are looking for a parameter set $\theta$ such that the subspaces spanned by $U_s$ and $A(\theta)$ match each other as close as possible in a least squares sense. If we can assure that both matrices have full rank $r$, the following criterion provides this feature

$$
\min_{\theta} f(\theta) = \| P_{U_s} A(\theta) \|_F^2,
$$

(7)

where $P_{U_s}$ is the projection matrix onto the null space of $U_s$, and $\| \cdot \|_F$ denotes the Frobenius norm of a matrix. Eq. (8) can also be written as

$$
\min_{\theta} f(\theta) = \text{tr} \left( A_H(\theta) P_{U_s} A(\theta) \right).
$$

(8)

The matrix of estimated eigenvectors $U_s$ is always of full rank because of the orthogonality of the eigenvectors. To ensure that $A$ will have rank $r$, we modify $f(\theta)$ by multiplying it with a function $g(\theta)$ that introduces poles at values where $\theta$ contains non-distinct values. The resulting, modified function is denoted as $h(\theta) = f(\theta) \cdot g(\theta)$. The poles of $g(\theta)$ will cancel the minimum points of $f$ at these values. Thus only the minimum solution will be reached where all elements in $\theta$ are distinct.

We suggest to use the following function

$$
g(\theta) = \frac{n}{r} \text{tr} \left( (A_H(\theta) A(\theta))^{-1} \right).
$$

(9)

(10)

Because we stated the sensor array of being unambiguous, the matrix $A(\theta)$ will have rank $r$ for all values of $\theta$ with distinct elements. It follows that $g(\theta)$ reaches infinity if and only if two elements of $\theta$ are equal. In this case, the matrix $A_H(\theta) A(\theta)$ becomes singular.

A second reason for this choice of $g(\theta)$ is that this function equals approximately 1 for all values of $\theta$ with distinct $\theta_i$ ($i = 1, \ldots, r$). This means that the location of the minimum solution is almost not affected.

This choice leads to a criterion function that retains the high resolution properties of the MUSIC method and additionally ensures that no source parameter will be detected twice which would result in the loss of another.

4 The New Algorithm

Consider again the criterion function to be minimized

$$
h(\theta) = f(\theta) \cdot g(\theta) \quad \text{with}
$$

$$
f(\theta) = \text{tr} \left( A_H(\theta) P_{U_s} A(\theta) \right)
$$

(11)

$$
g(\theta) = \frac{n}{r} \text{tr} \left( (A_H(\theta) A(\theta))^{-1} \right).
$$

(12)

To find the minimum solution of this function, one could apply a gradient or Newton method to iteratively descend to the desired solution. In the case of a linear array we suggest to use a gradient technique, because Newton’s method did not show better results compared to the gradient method. The reason is that the Hessian matrix of $h(\theta)$ is approximately diagonal and constant over the entire parameter space. This means that a gradient method with a properly chosen stepsize will have similar properties as Newton’s method. For other antenna geometries however a Newton method might be preferable.

If we define $D(\theta)$ as the matrix of derivatives of the steering vectors with respect to $\theta$,

$$
D(\theta) = \left[ \frac{\partial g(\theta_1)}{\partial \theta_1}, \ldots, \frac{\partial g(\theta_r)}{\partial \theta_r} \right],
$$

(13)

(14)

the resulting gradient algorithm can be written as

$$
\theta(t) = \theta(t - 1) - \delta \nabla h(\theta)
$$

(15)

$$
\nabla h(\theta) = \nabla f(\theta) g(\theta) + \nabla g(\theta) f(\theta)
$$

(16)

$$
\nabla f(\theta) = 2 \Re \left\{ \text{diag}(D_H P_{U_s} A) \right\}
$$

(17)

$$
\nabla g(\theta) = -2 \frac{n}{r} \Re \left\{ \text{diag}(A_H A)^{-2} A_H D \right\},
$$

(18)

where the matrices $A$ and $D$ have to be evaluated at $\theta(t - 1)$. $\Re$ extracts the real part of its argument and diag defines a column vector build by the diagonal elements of the argument matrix.

For this algorithm to converge to the global solution, it is required, that the starting value lies in the
neighborhood of the desired solution. Otherwise the iteration will converge to a local minimum only. This can be ensured if the source parameters change slowly and a solution at time $t - 1$ has already been achieved. But for $t = 1$ or when sudden changes happen, no such solution is available. Therefore a different approach has to be taken.

In some very special cases (linear array with equispaced sensors and plane waves), solutions can be achieved via the Root-MUSIC or ESPRIT algorithm. If no such method can be applied, starting values cannot be obtained via a stochastic gradient algorithm or natural algorithms e.g. simulated annealing or genetic algorithms.

5 Experiments

In our simulations we used a linear array with $n = 9$ equispaced sensors. The waves are planar with Gaussian signal and the noise is spatially and temporally white with power $\sigma^2 = 1$. All algorithms use the same forgetting factor $\beta = 0.97$ for estimating the covariance matrix (Batch ED) or eigenvectors (PASTd).

We compare our method with the exact eigen-decomposition of the sample covariance matrix followed by a Root-MUSIC approach [1].

Figure 1 shows the behaviour of the different methods when the sources are located close to each other. Three sources located at angles with $0, 6$ and $12$ degree move to $-3, 3$ and $9$ degree respectively within 1000 samples. The signal to noise ratio is 5dB. One should note that the distance between the sources is closer than half the beamwidth. Although the computational effort of the exact eigen-decomposition and the Root-MUSIC method is much larger than that of our method, the result shows poor performance. The reason is, that the PASTd algorithm has better resolution performance than an exact eigen-decomposition [5]. Our algorithm does not have any problems in tracking the source parameters. The second simulation shown in Figure 2 contains three sources, one of them at a constant direction with 0dB SNR, the two others initially located at 10 and 20 degrees, crossing at sample 500 and finally reaching their starting positions again. The batch ED and Root-MUSIC suffers from the close spacing of the sources and is not able to estimate the signal subspace and the corresponding directions correctly. Figure 3 depicts the third scenario which is similar to the second one but the signal to noise ratio has dropped to -10dB and the third constant source has been left out. Our method still shows reasonable results. The batch ED is not able to find or track the signals at all.

The last simulation (Figure 4) shows five sources,
with four of them constant and the fifth travelling through all of them within 5000 samples. The batch ED only yields reasonable results when the distance between two sources is not too close. Our method tracks all five signals without problems.

6 Conclusion

We present a new criterion function whose global minimum, given an estimate of the signal subspace supplies, supplies the locations of narrowband sources. Based on this function we introduce a gradient based algorithm for tracking the global minimum and hence the parameter estimates.

It computes parameters for all sources simultaneously at a computational cost of $O(nr^2) + O(r^3)$. The structure of the operations is highly regular, allowing an implementation on a parallel structure (e.g. systolic array). Using simulations we compared this new algorithm with the exact eigen-decomposition of the sample covariance matrix, followed by a Root-MUSIC approach.

References


MUSIC Estimation of Real-Valued Sinewave Frequencies

Petre STOICA, Anders ERIKSSON, and Torsten SÖDERSTRÖM

Systems and Control Group, Uppsala University, P O Box 87, S-751 08 Uppsala, Sweden. E-Mail: se@SysCon.uu.se

1 On leave from the Department of Automatic Control, Polytechnic Institute of Bucharest, Romania.

Abstract. The main focus of this paper is on the estimation of the frequencies of real-valued sinusoidal signals by using the MUSIC methodology. The MUSIC problem associated with real-valued sinewaves turns out to be more intricate than the similar problem corresponding to complex-valued sinusoidal signals. A whole class of weighted MUSIC estimates which solve the real-valued problem is shown to exist. The paper discusses the selection of an appropriate estimate in that class, from both computational and statistical performance standpoints. The unweighted MUSIC estimate is identified as the choice with a good trade-off between computational complexity and statistical accuracy.

1. Introduction

Estimation of the frequencies of complex-valued sinusoidal signals by using the MUSIC (Multiple Signal-Classification) algorithm ([4]) is well-documented in the literature (see, e.g., [3, 7, 8]). However, the literature is scarce when it comes to consider the application of the same methodology to frequency estimation of real-valued sinusoidal signals. This may look like a serious omission, as all real-life signals are real-valued. However, there exists at least one simple reason for this omission: The complex MUSIC algorithm can be applied to real-valued data by pretending that the data consists of a double number of complex-valued sinewaves. The so-obtained frequency estimation method is referred to as MUSIC-COM in what follows.

In this paper it is shown that careful application of the MUSIC methodology to a real-valued sinusoidal signal leads to a frequency estimation problem which has a whole class of solutions. The MUSIC-COM algorithm is just one possible solution in that class. Since MUSIC-COM is an ad-hoc solution, it might be expected that its use would be associated with an accuracy degradation, compared with the best achievable accuracy in the class. Indeed, there is no a priori reason why MUSIC-COM should be preferred to other possible solutions in the aforementioned class. Nevertheless, the statistical analysis of this paper reveals the fact that the accuracy degradation corresponding to MUSIC-COM is not significant. Combining this finding with the observation that MUSIC-COM is the most convenient algorithm in the class, from a computational standpoint, identifies MUSIC-COM as the method of choice for most applications.

2. Problem Statement

Let

\[ z(t) = \sum_{k=1}^{n} x_k(t) + \varepsilon(t), \quad t = 1, \ldots, N \]  

(1)

where

\[ x_k(t) = \alpha_k \sin(\omega_k t + \phi_k) \]  

(2)

denote the \( N \) available observations of a signal consisting of a real-valued sinewaves in noise. In (1), \( \{\alpha_k > 0, \{\omega_k \in [0, \pi]\} \), and \( \{\phi_k\} \) are assumed to be independent random variables, uniformly distributed over the interval \([0, 2\pi]\). The measurement error \( \varepsilon(t) \) is assumed to be zero mean Gaussian white noise with variance \( \sigma^2 \) and independent of \( \{\phi_k\} \).

For \( m > 2n \), let

\[ y(t) = (z(t) \ldots z(t + m - 1))^T. \]  

(3)

Since

\[ x_k(t + l) = \cos(l \omega_k) \alpha_k \sin(\omega_k t + \phi_k) + \sin(l \omega_k) \alpha_k \cos(\omega_k t + \phi_k) \]  

(4)

it readily follows that

\[ y(t) = Bx(t) + \varepsilon(t) \]  

(5)

where

\[ B = (A(\omega_1) \ldots A(\omega_n)) \]  

(6)

\[ x(t) = (\alpha_1 \sin(\omega_1 t + \phi_1) \alpha_1 \cos(\omega_1 t + \phi_1) \ldots \alpha_n \sin(\omega_n t + \phi_n) \alpha_n \cos(\omega_n t + \phi_n))^T \]  

(7)

\[ \varepsilon(t) = (\varepsilon(t) \ldots \varepsilon(t + m - 1))^T \]  

(8)

and where

\[ A(\omega) = \begin{pmatrix}
1 & 0 & \cdots & 0 \\
\cos(\omega) & \sin(\omega) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\cos((m - 1) \omega) & \sin((m - 1) \omega) & \cdots & \cos((m - 1) \omega)
\end{pmatrix} \]
\[ \Delta \triangleq \begin{pmatrix} a_2(\omega) & a_1(\omega) \end{pmatrix}. \]

By using (5), the covariance matrix of \( y(t) \) is easily shown to be

\[ R \triangleq E\{ y(t)y^T(t) \} = BB^T + \sigma^2 I \]

where

\[ P = \frac{1}{2} \begin{pmatrix} \alpha_1^2 I_{(2 \times 2)} & 0 \\ 0 & \alpha_2^2 I_{(2 \times 2)} \end{pmatrix}. \]

The eigenelements of the matrix \( R \) above have some interesting properties. Let \( \{ \lambda_k \}_{k=1}^{2n} \) denote the 2n largest eigenvalues of \( R \), arranged in a decreasing order, and let

\[ \Lambda = \text{diag} (\lambda_1 \ldots \lambda_{2n}). \]

It is well known that the eigendecomposition of a matrix of the form of \( R \) can be written as (see e.g. [2, 3])

\[ R = \begin{pmatrix} 2n & m-2n \\ S & G \end{pmatrix} \begin{pmatrix} \Lambda & 0 \\ 0 & \sigma^2 I \end{pmatrix} \begin{pmatrix} S^T \\ G^T \end{pmatrix} \]

and that the matrices \( S \) and \( G \) satisfy the following range properties:\footnote{Here \( \mathcal{R} \) stands for the "range space", and \( \mathcal{N} \) stands for the "null space".}

\[ \mathcal{R}(S) = \mathcal{R}(B), \quad \mathcal{R}(G) = \mathcal{N}(B^T). \]

Note from (13) that \( \lambda_{2n} \geq \sigma^2 \). In what follows it is assumed that \( \lambda_{2n} > \sigma^2 \) holds strictly. Also note that the matrices \( S \) and \( \Lambda \) appearing in (13) can be uniquely defined, whereas \( G \) is not unique (any post-multiplication of \( G \) by an orthogonal matrix gives another valid \( G \) matrix). The non-uniqueness of \( G \), however, does not affect the discussion to follow as we will deal with the range space of \( G \) (which is unique) rather than with \( G \) itself.

From the second property in (14) we obtain the following equation which uniquely determines the frequencies:

\[ G^T A(\omega_k) = 0, \quad k = 1, \ldots, n \]

or equivalently

\[ Q(\omega_k) = 0 \quad \text{for} \quad k = 1, \ldots, n \]

where

\[ Q(\omega) = A^T(\omega)\Pi A(\omega), \quad \Pi = GG^T. \]

The matrix equation (16) looks different from the scalar equation that holds in the complex-valued case (see [3, 4, 8] and (22) below). The thrust of this paper is a discussion on how to use (15) for frequency estimation in a statistically and computationally sound way. The number of sinewaves, \( n \), in the data is assumed to be known.

3. A Class of MUSIC Estimates

Let \( \hat{G} \) denote an estimate of \( G \) in (15), obtained from the eigendecomposition of the sample covariance matrix:

\[ R = \frac{1}{M} \sum_{i=1}^{M} y(t)y^T(t) \]

\[ = \begin{pmatrix} \hat{\Sigma} & \hat{\mathbf{G}} \end{pmatrix} \begin{pmatrix} \Lambda & 0 \\ 0 & \Sigma \end{pmatrix} \begin{pmatrix} \hat{\mathbf{G}}^T \\ \hat{\Sigma}^T \end{pmatrix} \]

where \( M = N - m + 1 \). In view of (15), consistent estimates of the frequencies can be derived as the arguments which minimize\footnote{\( || \cdot ||_F \) denotes the Frobenius norm, \( ||A||_F = \text{tr}(A^T A) \). Hereafter, "tr" stands for the trace operator.}

\[ f(\omega) \triangleq || \hat{G}^T A(\omega)W^{1/2} ||_F^2 = \text{tr}[W \hat{Q}(\omega)] \]

where \( W \) is a non-negative definite weighting matrix, and

\[ \hat{Q}(\omega) = A^T(\omega)\hat{\Pi} A(\omega), \quad \hat{\Pi} = \hat{G}\hat{G}^T. \]

Different choices of \( W \) in (19) lead to different MUSIC estimates. In this paper we only consider column weighting of \( \hat{G}^T A(\omega) \) in (19). The study in [7] of optimally weighted MUSIC for complex-valued signals showed that there is no significant performance improvement by weighting the rows of the complex-valued vector counterpart of \( \hat{G}^T A(\omega) \) in (19); and this type of behaviour is expected to carry over the real-valued case as well.

Next we show that the MUSIC-COM algorithm, outlined in the Introduction, corresponds to (19) with \( W = I \). In MUSIC-COM, the frequency estimates are determined as the minimizers of the following function:

\[ f_{\text{COM}} (\omega) = a^H(\omega)\hat{\Pi} a(\omega) \]

where \( a(\omega) = a_2(\omega) + ia_1(\omega) \), and the superscript "\( ^H \)" denotes the conjugate transpose. However, it is easy to see that (22) can be written as follows,

\[ f_{\text{COM}} (\omega) = (1 - i) A^T(\omega)\hat{\Pi} A(\omega) \begin{pmatrix} 1 \\ i \end{pmatrix} \]

\[ = \text{tr} \left[ \begin{pmatrix} 1 & -i \\ i & 1 \end{pmatrix} \begin{pmatrix} 1 & -i \end{pmatrix} \hat{Q}(\omega) \right] \]

\[ = \text{tr} \left[ \begin{pmatrix} 1 & -i \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix} \hat{Q}(\omega) \right] \]

\[ = \text{tr} \left[ \hat{Q}(\omega) \right] \]

which is the sought result.

There is no obvious statistically-oriented reason why one should prefer the choice \( W = I \) corresponding to MUSIC-COM, over other possible choices of the weighting matrix. The next section establishes the asymptotic variance of the generic MUSIC frequency estimate
derived from (19). The minimum asymptotic variance is then determined and compared with the asymptotic variance of MUSIC-COM. The degradation of accuracy associated with MUSIC-COM is found to be minor in many cases. By combining this finding with the computational convenience of MUSIC-COM, it is concluded that MUSIC-COM is the method of choice for most applications.

4. Statistical Analysis

For either $N$ or the SNR (signal-to-noise ratio) sufficiently large, the function $f(\omega)$ has a minimum point $\hat{\omega}_k$ close to the true frequency value $\omega_k$ ($k = 1, \ldots, n$), by the consistency properties of the minimizers of (19). This observation makes it possible to use a Taylor series expansion technique to prove the following result (see [6]).

**Theorem 4.1.** Under the previously made assumptions, the asymptotic (in either $N$ or SNR) variance of the frequency estimates obtained from the minimization of the MUSIC criterion function (19) is given by

$$
\text{var}(\hat{\omega}_k) = \frac{1}{M^2} \left( \sum_{l=-(m-1)}^{m-1} (M-|l|) \text{tr}[W\hat{U}_k(l)W\hat{U}_k(-l)] \right) / (\text{tr}[W\Delta_k])^2
$$

(24)

where

$$
\Delta_k = D_k^T \Pi D_k,
$$

(25)

$$
D_k = A'(\omega_k).
$$

The other matrices appearing in (24) are defined as follows. Let

$$
\hat{\Lambda} = \Lambda - \sigma^2 I
$$

(26)

$$
R_l = E[y(t) y^T(\tau - t)],
$$

(27)

where

$$
\Omega_l = \text{diag}(\cos(l\omega_1) \sin(l\omega_1) \ldots \cos(l\omega_n) \sin(l\omega_n))
$$

and where $J_l$ is a square matrix of dimension $m \times m$ with ones on the $l$th diagonal ($l > 0$ is above the main diagonal) and zeros elsewhere. Then, in (24),

$$
\hat{U}_k(l) = A(\omega_k)^T S \hat{\Lambda}^{-1} S^T \hat{R}_l S \hat{\Lambda}^{-1} S^T A(\omega_k)
$$

(28)

$$
\hat{V}_k(l) = \sigma^2 D_k^T \Pi J_l \Pi D_k
$$

(29)

$$
\hat{U}_k(l) = \sigma^2 A(\omega_k)^T S \hat{\Lambda}^{-1} S^T J_l \Pi D_k.
$$

(30)

We can now pose the problem of determining the weight $W$ that minimizes the variance of the frequency estimation errors. Let

$$
\hat{w} \triangleq \text{vec}(W) \triangleq \begin{pmatrix} W_{1,1} & W_{1,2} & W_{2,1} & W_{2,2} \end{pmatrix}^T
$$

(31)

denote the vector obtained by stacking the columns of $W$ on top of each other; and let

$$
\delta_k \triangleq \text{vec}(\Delta_k).
$$

(32)

Also, let $\otimes$ denote the Kronecker matrix product. Then we can write

$$
\text{tr}[W\Delta_k] = w^T \delta_k
$$

(33)

and

$$
\text{tr} \left[ W(\hat{U}_k(l) W(\hat{U}_k(-l)) + W(\hat{U}_k(l)) W(\hat{U}_k(-l)) \right]
$$

$$
= w^T \left( \hat{V}_k(l) \otimes \hat{U}_k(l) + \hat{U}_k(-l) \otimes \hat{U}_k(l) \right) w
$$

(34)

(see, e.g., [5]). Hence, if we define

$$
C_k = \frac{1}{M^2} \sum_{l=-(m-1)}^{m-1} (M-|l|) \left( \hat{V}_k(l) \otimes \hat{U}_k(l) + \hat{U}_k(-l) \otimes \hat{U}_k(l) \right)
$$

(35)

then the variance expression (24) can be rewritten in the following more compact form:

$$
\text{var}(\hat{\omega}_k) = \frac{(w^T C_k w)}{(w^T \delta_k)^2}.
$$

(36)

We want to minimize the above variance with respect to the weighting matrix elements in $w$. This can be done in the following steps.

**Step 1.** By the Cauchy-Schwartz inequality, $w^T C_k w / (w^T \delta_k)^2 \geq 1$ with the equality being attained for

$$
w_0 = C_k^{-1} \delta_k.
$$

(38)

Let $W_0$ denote the weighting matrix corresponding to $w_0$ above (if $W_0$ is not symmetric, one can use $W_0 + W_0^T$ in lieu of $W_0$ without affecting the frequency estimates). If $W_0$ is non-negative definite, then the optimal weight has been found. Otherwise, go to the next step.

**Step 2.** If $W_0$ determined in Step 1 does not belong to the set of non-negative definite matrices, the minimizer of (36) with respect to $w$ should belong to the boundary of the aforementioned set. Since a scaling of $W$ does not affect the frequency estimates, there is no restriction to set $W_{1,1} = 1$. By making use of this observation, the minimization problem under discussion becomes:

$$
\min_{W_{2,2} > 0} \frac{(w^T C_k w)}{(w^T \delta_k)^2}
$$

(39)

$$
W_{1,1} = W_{1,2} = \pm \sqrt{W_0, 0}
$$

Here $C_k^{1/2}$ denotes a (symmetric) square root of the positive definite matrix $C_k$. 

647
Similarly small gains have been observed in several other examples whose results are reported in [6]. This minor difference in statistical performance along with the computational attractiveness of MUSIC-COM lead to the conclusion that MUSIC-COM should be the method of choice for estimation of the frequencies of real-valued sinusoidal signals in white measurement noise in most cases.

Acknowledgements

This work has been supported by the Swedish Research Council for Engineering Sciences under contracts no 91-676 and 93-669.

References


Computation of Affine Wigner-Ville Distributions by Means of Chirp Z-Transform

Tomasz P. ZIELINSKI

Abstract. The paper describes in detail an novel application of the chirp z-transform to the computation of a wide family of affine Wigner-Ville distributions, including tomographic Bertrand, Flandrín as well as active and passive Unterberger representations. This family represents a wide-band counterpart of the narrow-band Cohen class of joint bilinear time-frequency signal representations (standard Wigner-Ville distribution, Choi-Williams, etc.) and it is used in signal analysis performed in wide-band Doppler echography. The proposed method is compared with the Mellin transform based approach in the paper and experimental examples of its usage are also presented.

1. Introduction

Mixed linear and bilinear time-frequency signal distributions (MTFSDs) have been proved recently to be very powerful tool for analysis and processing of nonstationary signals [1,2]. The short-time Fourier, Wigner-Ville and wavelet signal representations are the most important examples.

Wide application of the standard Wigner-Ville distribution (WVD) results from the fact that for a linear frequency modulated signal the WVD spectrum energy is concentrated in the time-frequency domain exactly along the straight line of its modulation law, so it is perfect "observable". Since not only the LFM signals are of interest in engineering practice, a significant research effort has been made recently for defining a new MTFSD that is capable of ideal "tracking" of an arbitrary modulation law in the TF plane. Until now two important solutions to the problem have been proposed, namely polynomial [3] and affine [4-7] versions of the classical WVD.

The main goal of this report is to propose an efficient algorithm for the computation of the affine WVDs (AWVDs) from the definition given in [4]. The chirp z-transform (CZT) [8] is used in our approach for the calculation of signal scaled spectra as a effective numerical number cruncher. For the first time such application of the CZT was proposed in [9] and then further elaborated in [10] but in both cases only computation of the wavelet transform was discussed.

The AWVDs can be also calculated via fast Mellin transform [11,12]. We will compare our present approach with the previous one [12] as well as give some experimental examples of the time-varying signal analysis in the paper.

2. Affine Wigner-Ville Distributions

A diagonal subclass of affine Wigner-Ville distributions (AWVDs) is defined with the following formule [4]:

\[
\text{AWVD}(t,f) = \int \frac{e^{j2\pi f u}}{u} X(f_l,\lambda(u))X^*(f_l,-\lambda(u))\mu(u)du
\]

(1)

where \( t, f \) and * denote time, frequency and complex conjugation, respectively, \( X(f) \) is the Fourier transform of an analytic time signal \( x(t) \), \( \lambda(u) \) and \( \mu(u) \) are functions that are specific for different representations (see table 1), and \( \phi(u)=\lambda(u)-\lambda(-u) \). Functions \( \lambda(u) \) and \( \mu(u) \) determine features of the resultant AWVDs calculated from eq. (1) [13]. For narrow-band signals eq. (1) can be transformed to the classical definition of the WVD [4].

After introduction of new variables \( \nu=\phi(u) \) and \( \gamma=\lambda(u) \) eq. (1) can be rewritten into the following form

\[
\text{AWVD}(\gamma,\nu) = \int_{-\infty}^{\infty} e^{j2\pi \gamma \nu} X(f_l,\nu)X^*(f_l,-\nu)\mu'(\nu)d\nu,
\]

(2a)

where

\[
\mu'(\nu) = \frac{\mu(u)}{(\lambda(u)\lambda(-u))^{1/2}} \frac{d\phi(u)}{du}
\]

(2b)

and interpreted as the inverse Fourier transform performed over variable \( \nu \) on

Functions \( \lambda(u), \phi(u), \mu(u) \) and \( \mu'(u) \) for four particular AWVDs defined with eq. (1).

<table>
<thead>
<tr>
<th>Representation</th>
<th>( \lambda(u) )</th>
<th>( \phi(u)=\lambda(u)-\lambda(-u) )</th>
<th>( \mu(u) )</th>
<th>( \mu'(u) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bertrand [4]</td>
<td>( u/(1-\exp(-u)) )</td>
<td>( u )</td>
<td>( u/(2\sinh(u/2)) )</td>
<td>1</td>
</tr>
<tr>
<td>Flandrín [13]</td>
<td>( (1+0.25u)^2 )</td>
<td>( u )</td>
<td>( (0.25u)^2 )</td>
<td>1</td>
</tr>
<tr>
<td>Unterberger active [4,13]</td>
<td>( \exp(u/2) )</td>
<td>( \exp(u/2)-\exp(-u/2) )</td>
<td>( \cosh(u/2) )</td>
<td>1</td>
</tr>
<tr>
<td>Unterberger passive [13]</td>
<td>( \exp(u/2) )</td>
<td>( \exp(u/2)-\exp(-u/2) )</td>
<td>( 1/2 )</td>
<td>( 1/(2\cosh(u/2)) )</td>
</tr>
</tbody>
</table>

Table 1
The final result is scaled by appropriate frequency values $f$.

3. Scalograms Calculation via CZT

There are numerous functionalities used in signal processing area that are based on the scaled version of a signal or on its scaled spectrum (for example wavelet transform, wide-band ambiguity function, affine Wigner distributions, time-frequency Q distribution of Altes, etc.) [1, 2]. When it is possible to transform these functionalities into frequency domain the CZT can be used for calculation of the dilated signal spectra $X(qf)$, where $a$ stands for a scaling coefficient. For the first time this approach was reported in [9].

Let assume that a discrete scaled spectrum $X(2\pi ak/M)$ is to be calculated for $k=M/2,...,0,...,M/2-1$. After change of variables $9 <= k = p+M/2$ the following relations hold:

$$X\left(\frac{2\pi}{M} ak\right) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi\frac{a}{M}p}\cdot e^{-j2\pi\frac{a}{M}n} = \sum_{n=0}^{N-1} x(n)e^{-j2\pi\frac{a}{M}p} = \sum_{n=0}^{N-1} x(n)A^pW^n = \sum_{n=0}^{N-1} x(n)A^pW^n =$$

$$= \sum_{n=0}^{N-1} x(n)A^pW^n.$$  

(3a)

where

$$A = e^{j\frac{2\pi}{M}} \quad \text{ and } \quad W = e^{-j\frac{2\pi}{M}}.$$  

(3b)

Since eq.(3a) has a form of the CZT [8], the spectrum $X(2\pi ak/M)$ can be calculated by means of the standard CZT routine (three FFTs after appropriate change of constants $A$ and $W$) (eq.(3b)).

For comparison $A$ and $W$ are defined in [9] as: $A = \exp(-jka)$ and $W = \exp(j2\pi km/M)$. It results from the following parameterization of $X(2\pi ak/M)$:

$$X\left(\frac{2\pi}{M} ak\right) = \sum_{n=0}^{N-1} x(n)A^pW^n = \sum_{n=0}^{N-1} x(n)A^pW^n.$$  

(3a)

where $p = k+M/2$.

Computer source code of the modified CZT is presented in table 2. It is a transformed version of a Fortran program published in [8] (lx=1=M+1-1, ly=2=N).

4. Calculation of Affine WVD via CZT

Any affine WVD can be computed via modified CZT after suitable discretization of its continuous definition (2)

$$AWVD(x,k) = \frac{k}{M} \sum_{n=0}^{N-1} x\left(\frac{2\pi}{M} n\right)A^{-n}W^n,$$

(4)

The algorithmic formula has been given in table 2.

Step 1: Appropriate discretization of variable $u$:

$$u^d = u^i \forall i = 0,...,N-1.$$  

Step 2: Calculation of the functions values $\lambda^d(u)$, $\lambda^{-d}(u)$, and $\mu^d(u)$ (see table 1).

Step 3: Finding scaled discrete versions $X^1(k) = X(2\pi k\lambda^d(u)/M)$ and $X^{-1}(k) = X(2\pi k\lambda^{-d}(u)/M)$ ($k = 0,M/2,...,N/2-1$) of the spectrum $X(f)$ by means of the modified CZT given by eq.(3).

Step 4: Calculation of real physical time-coordinates of the resultant representation $AWVD(u,k) \rightarrow AWVD(u,k)$.

5. Calculation of Affine WVD via Mellin Transform Method

An alternative method for discretization and efficient numerical computation of the integral (1) has been proposed in [11]. It makes use of the discrete fast Mellin transform and utilizes its features in order to rewrite eq.(1) into a form more suitable for the computer calculation. It is assumed that the analysed analytic signal $x(t)$ is limited both in time (for $t \in [-T, T]$) and in frequency ($X(f) = 0$ for $f \in [-B/2, B/2]$). An outline of the computational algorithm consists of the following operations:

- Geometrical sampling of $X(f)$ in the interval $[f_1,f_2]$.
- $M[X(B) = MellinTransform(X(f))].$
- $Z(\alpha,\beta) = MellinTransform(Z(\alpha,\beta)).$
- $F_k(l,\beta,\alpha) = F_k(l,\beta,\alpha).$
- $P(u,\alpha,\beta) = P(u,\alpha,\beta).$
- $P(t,\alpha,\beta) = P(t,\alpha,\beta).$

The detailed computer program implementing above algorithm is given in [12].

5. Experimental results

As an example we have computed tomographic Bertrand representation (eq.(1) and table 1) of a synthetic signal consisting of two gaussian components using both chirp z and Mellin transform algorithms. The resultant time-frequency matrices are shown in figure 1. Since the analysed signal has been narrow-band both achieved affine WVDs are equivalent to the standard one.

6. Conclusions

Computation of the affine WVDs by means of discrete chirp z-transform described in the paper is more simple but less efficient than the fast Mellin transform method [11,12]. Calculation of an exemplary AWVD matrix with dimensions $N \times N$ by means of the proposed algorithm is approximately equivalent to performing the $25N$-point FFT procedures while only $N/2$-point FFTs are required in the Mellin method.
References


Table 2

Source code of the modified discrete chirp z-transform.

Calling: ModChirpZ(x, lx1, lx2, a)
Sequence of operations:

\[
\begin{align*}
\text{ly} &= \text{x1} * \text{lx2} + 1, \quad \text{A} = \text{exp}(-\text{jx1/ly);}
\text{for k} = 1, \ldots, \text{lx1};
\text{kk} &= k - 1;
\text{if (kk)} \text{then work}[k] = x[k] * (A_k * W^{kk})^{kk};
\text{if (k)} \text{then work}[k] = 0;
\text{if (kk)} \text{then x[k] = W^{kk} * x[k];}
\text{if (ly)} \text{then x[k] = W^{ly} * (lx1-lx2)^2} \text{work = FFT(work, lx1);}
\text{x = FFT(x, lx1);}
\text{for k = 1, ..., lx1}: 
x[k] = x[k] * work[k];
\text{x = IFFT(x, lx1); }
\text{for k = 1, ..., ly ; kk = k - 1, x[k] = W^{kk} * (kk * kk);}
\text{ibord = NearInteger(ly2/2^a), ly2 = ly2/2;}
\text{for k = ly2-1, ..., ly2-2:}
\text{if (Abs(k) > ibord) x[ly2+k]} &= 0;
\end{align*}
\]

Table 3

Source code for the computation of affline Wigner-Ville distributions by means of the chirp z-transform.

1. Parameters initialization:

\[
\begin{align*}
\text{analysed signal c[n]: NsFirst, NsSamples, f_0, B = ?}
\text{N = NsSamples, } f_1 = f_0 - 0.5B, f_2 = f_0 + 0.5B, 
\text{a} = f_1/0.5B, \text{a} = f_2/0.5B; 
\text{M = NearPower2(NsSamples);}
\text{M2 = M/2; ly1 = 2^M, ly2 = M, df = 0.5/M2;}
\end{align*}
\]

2. Calculation of N, \(u_{max}\) and scaling coefficients \(\lambda(+u), \lambda(-u)\) and \(\mu(u)\) for values of u:

\[
\begin{align*}
\text{2.1 Finding such values u > 0 i u < 0, that:} 
\lambda(-u) &= a_1, \lambda(+u) = a_2; 
\text{2.2 u_{max}^2 = min(u_{max}, u_{min});}
\text{2.3 N - 2, N = NearPower2(2^10 x_{max} + a_2/(-a_2 - 1));}
\text{2.4 Discretization of variable u:}
\quad \text{for Bertrand and Flandrin:}
\quad \text{for i = 1, ..., N: u_{i} = (i - 1)_{max}/N;}
\quad \text{for active and passive Unterberger:}
\quad \text{phi_{max} = lambda_{max} -lambda_{min}, Aphi = phi_{max}/(N - 1);}
\quad \text{for i = 1, ..., N:}
\quad \text{find such u > 0, that lambda(u) - lambda(u_{i}) = (i - 1) \times phi_{max};}
\text{2.5 Setting work vectors:}
\quad \text{for i = 1, ..., N:}
\quad \text{rwi[1] &= lambda(u_{i}), rwi2[i] = lambda(u_{i}) + 1, rwi3[i] &= lambda(u_{i});}
\end{align*}
\]

3. Main loop:

\[
\text{for i = 1, ..., N:}
\text{3.1 Computation of the spectrum X(2n(lambda_{i} + u_{i}))/M):}
\quad \text{si_{plus} = rwi1[i];}
\quad \text{for j = 1, ..., NsSamples :}
\quad \text{cxj[i] &= cs(NFirst - 1 + j);}
\quad \text{for j = NsSamples + 1, ..., M : cxj[i] = 0;}
\quad \text{cx = ModChirpZ(cx, ly1, ly2, si_{plus});}
\quad \text{3.2 Computation of the spectrum X(2n(lambda_{i} - u_{i}))/M:}
\quad \text{ai_{minus} = rwi2[i];}
\quad \text{for j = 1, ..., NsSamples :}
\quad \text{cyj[i] &= cos(NFirst - 1 + j);}
\quad \text{for j = NsSamples + 1, ..., M : cyj[i] = 0;}
\quad \text{cy = ModChirpZ(cy, ly1, ly2, ai_{minus});}
\quad \text{3.3 Multiplication of the spectra:}
\quad \text{for k = 1, ..., M2:}
\quad \text{ctf[i, k] = rwi3[i] * cx[M2 + k] * cy[M2 + k];}
\end{align*}
\]

4. Inverse Fourier transform:

\[
\text{for k = 1, ..., M2:}
\quad \text{ctf[k, i] = cwork[i] = ctf[i, k];}
\quad \text{cwork[1] &= cwork[1] * 2;}
\quad \text{rwork = 2 \times Real(IFT(cwork, N));}
\quad \text{for i = 1, ..., N:}
\quad \text{AWVD[n, k] = kdf + cwork[n];}
\]

5. Calculation of the representation real co-ordinates (x - time, y - frequency):

\[
\text{for k = 1, ..., M2:}
\quad \text{rtemp = kdf;}
\quad \text{for n = 1, ..., N:}
\quad \text{x[n, k] = n(u_{max} + temp), y[n, k] = rtemp;}
\]

651
Figure 1. Affine Wigner-Ville distributions (tomographic Bertrand representations) of a synthetic narrow-band signal consisting of two gaussian components computed via: a/ b/ - chirp z-transform method, c/ d/ - fast Mellin transform method (a/, c/ - unscaled, b/, d/ - scaled co-ordinates).
Wavelet Packets for the Estimation of Parameters of Localized Sinusoids

Christian van den Branden Lambrecht and Mohsine Karrakchou

Signal Processing Laboratory
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
Tel: (+41 +21) 693-4714
EMail: vdb@lts.de.epfl.ch

Abstract. In this paper, a new approach to high resolution spectral estimation is presented. It is based on an adaptive subband decomposition performed with wavelet packets and it minimizes a new criterion designed for the problem of complex exponential estimation. The Tufts and Kumaresan method is then applied on each of the selected subbands. The new criterion consists in the evaluation of the number of modes contained in each subband, using the minimal description length criterion. Simulations performed on synthetic signals confirm the gain in performance of the proposed method.

1 Introduction

One problem often encountered in spectral estimation is that of resolving very sharp peaks that are closely located. This is the case in many applications like radar, sonar and passive arrays. Moreover, such applications require the estimation to be performed on short data records, yet low variance, low bias and high resolution estimates are desired. This demand gave birth to several high resolution techniques that are based on autoregressive modeling and robust parameterization. Tufts and Kumaresan [1] have developed a robust method (TK) combining linear prediction and low rank approximation of the autocorrelation matrix using singular value decomposition. The approximation is then used to estimate the coefficients of the prediction vector.

In a recent paper, Rao and Pearlman [2] demonstrated the superiority of spectral estimation performed on an M-band decomposition compared to that performed directly on the original signal. They assumed however the use of an ideal analysis filter bank. Such a filter bank does not exist and a major problem appears due to aliasing that can lead to the attenuation of some modes or even their complete disappearance. Hence the need for the use of a signal–dependent subband decomposition arose. Such a decomposition will not only avoid to attenuate these modes but also tend to isolate as far as possible each mode in a separate frequency band. This will lead to a better quality of the estimation due to the reduction of the interference between different modes.

Wavelet packets [3] are a time–frequency analysis tool that allows such an adaptive subband decomposition of the signal. It is possible to choose in a binary decomposition tree the optimal subtree optimizing a given criterion.

In this paper, the combination of wavelet packets decomposition, subband spectral estimation and high resolution estimation method is presented. For this purpose, a new decomposition criterion is introduced. The latter is directly adapted to the addressed problem since it consists in the counting of the number of modes contained in each subband.

This paper is organized as follows: Section 2 reviews the Tufts and Kumaresan spectral estimation method. The advantages of a subband decomposition for spectral estimation are presented in section 3. A brief summary of wavelet packets is given in section 4 while section 5 presents the proposed method. Finally, an experimental result is given in section 6 that shows the efficiency of the algorithm.

2 The Tufts and Kumaresan Method

The problem of interest is the estimation of the p frequencies of a linear combination of p complex exponentials embedded in zero mean white noise:

\[ y[n] = \sum_{k=1}^{p} c_k e^{j\omega_k n} + w[n], \]  

where \( c_k \) and \( \omega_k \) respectively represents the complex amplitude and radial frequencies to be estimated. The
additive zero-mean white noise is denoted by \( w[n] \).

The uncorrupted signal \( y_p[n] = \sum_{k=1}^{p} c_k e^{j\omega_k n} \) can be modeled as an autoregressive process according to the following forward and backward prediction equations:

\[
y_p[n] = \sum_{k=1}^{p} a_k y_p[n-k] \quad (2)
\]

\[
y^*_p[n] = \sum_{k=1}^{p} a_k y^*_p[n+k]. \quad (3)
\]

The prediction vector, made up of the \( a_k \) coefficients can be used, once known, to extract the frequency information by finding the zeros of the polynomial \( A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} = \prod_{k=1}^{p} (1 - c_k e^{j\omega_k} z^{-1}) \). This model has been at the origin of a large number of estimation methods. One of them, widely known, has been introduced by Tufts and Kumaresan [1]. It solves the system

\[
A \mathbf{a} = \mathbf{b}, \quad (4)
\]

where

\[
A = \begin{bmatrix}
y[L-1] & y[L-2] & \cdots & y[0] \\
\vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
\end{bmatrix},
\]

\[
\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_L \end{bmatrix}^T,
\]

\[
\mathbf{b} = \begin{bmatrix} y[L] & y[L+1] & \cdots & y[N-1] \\
y[N-L-1] & y[N-L] & \cdots & y[0] \end{bmatrix}^T,
\]

and \( L > p \) is the chosen prediction order.

The method is based on the resolution of the system of equation (4) by a low rank approximation of the matrix \( A \). The low rank approximation is performed with a singular value decomposition and is followed by a mean square resolution. The performance of the estimation is directly related to the robustness of the singular values of \( A \) with respect to perturbations in the matrix elements. Moreover, this allows an accurate estimation of the number of modes present in the signal without any \( a \ priori \) knowledge.

3 Subband Spectral Estimation

Subband signal processing techniques have received much attention from researchers in the last few years and stand as promising solutions to many problems encountered in various fields. Image coding is the field where such techniques have been most used. However, researchers recently started to investigate other applications for subband techniques [4]. Very recently, Rao and Pearlman investigated the use of subband decomposition for spectral analysis in a DPCM (differential pulse coded modulation) framework.

They theoretically showed, based on information theory criteria, that under the assumption of the use of an ideal filter bank, subband spectral estimation offers several advantages, namely:

- The minimum prediction error achievable in full band exceeds the sum over the subbands of the minimum prediction errors.
- The subband spectra are whiter than the full band spectrum. The analysis filters act as whitening filters.
- The composite entropies of the subbands are closer to the source entropy than the entropy in full band for the same global prediction order.
- The subband prediction error spectrum is whiter than the full band prediction error spectrum.

The key motivation for using subband decomposition techniques for sinusoid parameter estimation is twofold: not only the above advantages are gained but, in addition, subband decomposition can isolate the various modes in separate subbands, which allows a much better estimation. It is indeed well known that the lower the number of modes, the better the Cramer–Rao bound.

However, the non-ideal nature of any real-life filter bank will introduce spectral aliasing around the cutoff frequencies of the analysis filters. Such aliasing can severely attenuate some modes close to the transition regions and weaken the quality of estimation.

4 Adaptive Subband Decomposition

Wavelet packets are a generalization of the wavelet transform concept where arbitrary time–frequency resolution can be chosen according to the signal. This will be done, of course, within the bounds of Heisenberg uncertainty principle. The idea is to obtain an adaptive partitioning of the time–frequency plane depending on the signal of interest.

From the filter bank point of view, the wavelet packets decomposition algorithm can be summarized as follows:
Let $z[n]$ be an $L$-sample discrete–time signal and $h[n]$, $g[n]$ be the $N$–tap impulse responses of two FIR filters compliant to the conditions set forth in [5]. A tree of depth $\log_2(L)$ is generated by cascade filtering and downsampling as depicted in figure 1. The merging of some nodes still guarantees the orthonormality of the expansion while offering several decomposition bases.

Among the collection of bases, the one optimizing a given criterion is kept. The search for the best decomposition is made by a split and merge process locally applied on every node of the tree (see figure 1), starting from the leaves and going to the root. In this process the criterion is computed for a node and his two children. The nodes optimizing the criterion are conserved.

5 Algorithm and Criterion

It is desirable for the decomposition to be as deep as possible. Indeed, the more a subband is decimated, the more resolution is gained. In addition, the more subbands are present, the more likely it is to isolate each component in a subband. However, as the depth of the decomposition tree increases, the aliasing regions become wider. Hence the decomposition criterion has to find the tradeoff corresponding to deepest tree with no component aliased.

For the problem addressed here, the selection of the optimal decomposition is made by maximizing the number of modes over the whole decomposition tree [6]. This corresponds to stop the decomposition as soon as one mode would disappear due to further decompositions. The estimation of the number of modes is performed using the minimum description length criterion (MDL). It has been shown [7] that an analytical expression of the MDL criterion exists for complex exponential signals, which is:

$$ MDL(k) = -2\log \left( \frac{\left(\prod_{\lambda_i > 0} \lambda_i\right)^N}{\left(1/(L-k)\sum_{i=h+1}^{L} \lambda_i\right)^{(L-k)N}} \right) + \frac{1}{2} k (2L - k) \log(N), $$

where the $\lambda_i$'s are the $L$ eigenvalues of the autocorrelation matrix, $N$ is the number of samples. The number of modes is then the number $k$ that maximizes the criterion. During the decomposition, the criterion is evaluated each time on each node and its children.

If the number of modes remains constant, the decomposition is accepted (figure 2, left hand side). On the contrary, if a mode is severely attenuated due to aliasing, this mode will not be detectable by the criterion (figure 2, right hand side). In this case the process of maximization of the criterion will stop the decomposition. Once a decomposition is chosen, the TK method is applied on each of the subbands in the same manner it would be applied on the original signal. The estimated values of the frequencies are then converted to their full band values. The subband prediction orders are chosen so that their sum is equal to the optimal prediction order in full band [2].

![Figure 2. Illustration of the decomposition criterion effect: on the left hand side, the number of modes is conserved, the decomposition is kept. On the right hand side, one mode is lost, the decomposition is refused.](image)

Since the decomposition is no longer uniform, the prediction order in each subband is also proportional to the width of that subband. Note also that the use of the
The TK method is only a particular choice that does not exclude the use of any other method. The subband selection favors the estimation process regardless of any particular method. Its purpose is actually to increase the Cramer–Rao bound of every single mode contained in the signal. The purpose is to gain the advantages of subband spectral estimation, namely an increase of spectral resolution as well as attenuation of some inter-modes interferences, without having the drawbacks of loosing modes due to aliasing.

6 Experimental Results

The experiment presented in this section consists in the frequency estimation of a signal composed of two exponentials embedded in a zero–mean white noise. The normalized frequencies are 0.11 and 0.15. Monte Carlo simulations have been performed at different signal-to-noise ratios (SNR). 250 trials have been performed for every SNR value. For each trial, three estimations are performed: The first one is performed on the original signal (plain line in figure 3), the second uses the proposed adaptive decomposition (dashed line) and the third is based on a rigid decomposition (dot-dashed line). The rigid decomposition corresponds to a uniform filter-bank of eight bands generated by a binary decomposition tree without any merging. The analysis filters used for both decompositions are Daubechies $D_{10}$ filters [5]. The performance measures chosen are the variance of the estimation and the miss ratio versus SNR. The miss ratio represents the percentage of cases where no estimated frequency falls in a reasonable neighborhood of the actual frequency. In figure 3.a, are depicted the variance curves of the estimate of the lower frequency. The figure 3.b represents the miss ratio for the same estimation.

![Graph](image)

Figure 3. Performance of the estimation for the lower frequency of the test signal. The inverse of the variance is depicted above in a logarithmic graph. Miss ratio curves are shown below.

It appears that the proposed method allows to obtain lower miss ratios as well as better accuracy in the estimates. In the case of rigid decomposition, since one of the frequencies to be estimated lies in the transition region of the filter bank, the miss ratio is 100%. This is why no variance curve is reported for this case. These results illustrate the importance of the adaptivity of the decomposition. The proposed criterion has indeed stopped the decomposition at the adequate levels before a complete attenuation of the considered modes occurs.

7 Conclusion

This paper presents a novel scheme that combines wavelet packets and high resolution spectral estimation algorithms for the estimation of very localized spectral lines. The advantages of a subband decomposition are presented. Wavelet packets allow the creation of subband decompositions that are adapted to the spectral content of the signal. A criterion that is appropriate for the considered application is introduced. Finally Monte Carlo simulations are presented and show the gain in performance brought by the new method.

References


Improvements of NMR Data Quantitation by Using the Cadzow Signal Estimator Prior to Linear-Prediction Methods

Aboubakar DIOP, Sophie CAVASSILA, André BRIGUET and Danielle GRAVERON-DEMILLY

Laboratoire de Résonance Magnétique Nucléaire, Université LYON 1, 69622 Villeurbanne Cedex, France,
Tel/Fax: +33-72431049 / 72448199, E-Mail: graveron@muzelle.univ-lyon1.fr

Abstract. Signal analysis of in vivo Magnetic Resonance Spectroscopy (MRS) data modelled by a sum of exponentially damped sinusoids is considered. To improve the quantitation, the noisy data are pre-processed by using the Cadzow Enhancement Procedure (EP). Used up to convergence EP results in near optimum signals in the least-squares sense. These estimated signals involve purely exponentially damped sinusoids, then any linear-prediction methods can be used indifferently in the subsequent quantitation step. A Monte Carlo simulation using three different SVD-based quantitation methods shows that amplitude estimates have negligible bias and standard deviations nearly equal to their Cramer-Rao lower bounds even for low SNR. EP used prior to SVD-based methods results in near-optimal automatic quantitation methods.

1. Introduction
For medical diagnosis, reliable estimates of the NMR parameters have to be obtained from low SNR signals. To obtain such information, non-interactive algorithms such as linear-prediction methods (LP) based on SVD have been reported. The Kumaresan-Tufts method [1,2], the Kung's state-space approach [3-5] (respectively known as LPSVD and HSVD in NMR) and the total-least-squares method (TLS) [6-9] lead to biased estimates of parameters for low SNR signals. To overcome these drawbacks, we propose to pre-process the data using the Cadzow enhancement procedure [10] (called EP in NMR) up to convergence. The efficiency of this method followed by the Kumaresan-Tufts algorithm (EP-LSVD method in NMR) has already been reported [11, 12].

2. Method
The Cadzow Enhancement Procedure works directly on the linear-prediction matrix \( X \) of Hankel structure:

\[
X = \\
\begin{pmatrix}
x_1 & x_2 & x_3 & \ldots & x_m \\
x_2 & x_3 & x_4 & \ldots & \ldots \\
x_3 & \ldots & \ldots & \ldots & \ldots \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
x_{N-m+1} & \ldots & \ldots & \ldots & x_N \\
\end{pmatrix}
\] (1)

where \( x_1, \ldots, x_N \) are the \( N \) considered samples of the signal \( x \) modelled by a sum of \( K \) complex damped sinusoids and \( m \) is the prediction order. The matrix rank is equal to \( K \) in absence of noise and to \( \min(n, m) \) for noisy signals \( (n = N-m+1) \).

The Cadzow procedure is based on singular value decomposition (SVD) of the matrix \( X \) and consists in finding a matrix \( X_K \) of rank \( K \), having a Hankel structure and so that

\[
\| X - X_K \|_2^2 \rightarrow \text{minimum}
\]

This is reached in an iterative procedure including three steps:
1. SVD of the matrix \( X \) so that \( X = USV^* \).
2. truncation : only the \( K \) largest singular values are kept in \( S_k \).
3. \( X_T = US_KV^* \) and restoration of the Hankel structure by averaging all elements for each antidiagonal leading to \( X_K \).

At convergence the matrix \( X_K \) possesses two relevant properties. It has a rank \( K \) and a Hankel structure. Consequently, the estimated signal \( \hat{x} \), read from the entries of the matrix \( X_K \) corresponds exactly to a sum of \( K \) exponentially damped sinusoids so if we apply any quantification method to this pre-processed signal, the results must be exactly the same.

As the EP signal estimator is strongly non-linear, a systematic numerical study was performed on simulated signals for different values of the noise standard deviation \( \sigma \) to determine the best parameters \( \hat{m}, \hat{N} \) to be used. We found that at convergence, the condition

\[
\frac{\hat{N}}{\sqrt{2}} \sum_i^n |x_i - \bar{x}_i|^2 \leq \sigma^2
\]

is reached even for low SNR, if the matrix order \( \hat{m} \) is in the range \( \hat{N}/4 \leq \hat{m} \leq \hat{N}/2 \) and if the number \( \hat{N} \) of samples used is such that \( \hat{N}_T \approx 3T_2^*, T_2^* \) being the largest relaxation time. Four or five iterations are generally needed. EP then leads to a near optimum signal in the least-squares sense. Then for quantitation, the signal estimator EP will allow to use indifferently any quantitation method...
methods (e.g SVD-based methods such as LPSVD, HSVD, TLS) to get reliable estimates of the parameters. As the estimated signal is a purely exponentially decaying signal, the final quantitation step requires only a few samples of the estimated signal (e.g. $N = 2K$ can be used). So the quantitation step does not require a long computation time. The full computation time is nearly equal to the time needed by the EP procedure that can be very reduced if the orders are shrewdly chosen during the iterative procedure (e.g. if $m = N/4$ or if $m$ is chosen equal to $N/2$ during the first iteration and to $N/4$ after).

3. Results

The efficiency of the Cadzow signal estimator is clearly demonstrated in Fig.1 for an in vivo NMR $^{31}$P spectrum. A Monte Carlo study performed on simulated time-domain signals with a gaussian noise added, clearly shows the necessity of pre-processing. For that, a set of 100 FIDs corresponding to $^{31}$P NMR spectra of a muscle (see Fig.1) was simulated for each value of the noise standard deviation.

To compare the previously mentioned methods, we estimate for each SNR the mean amplitude values $\bar{A}_j$ and the corresponding standard deviation $\sigma_{A_j}$ from the $N_j$ successful fits. We focused our interest on the inorganic phosphate (Pi) amplitude because in $^{31}$P in vivo experiments its intensity is often slightly larger than the noise. It can be seen in Fig.2a that LPSVD, HSVD, TLS estimators lead to biased values for low SNR and that the corresponding amplitude standard deviations do not fit the Cramer-Rao lower bounds (Fig.2a). Then the same estimators have been applied to the pre-processed signals (see Fig.2b). EPLPSVD, EPHSVD, EPTLS give exactly the same good unbiased results for all SNR that means that EP leads to the near optimum signal. Moreover, the amplitude standard deviations perfectly fit the Cramer-Rao lower bounds (Fig.2b).

Figure 1. a) FFT of a $^{31}$P signal obtained from a calf muscle at 4.7 Tesla. b) FFT of the same signal obtained after the EP procedure (PCr = phosphocreatine, Pi = inorganic phosphate, ATP = adenosine triphosphate, MDP = methylene diphosphonic acid used as external reference).

Figure 2. Monte Carlo simulation results, the mean amplitude of Pi peak and the corresponding standard deviation are given as a function of the noise standard deviation $\sigma$.

The true Pi amplitude value is equal to 640.

a) LPSVD, TLS, HSVD ($N = 256$, $m = 192$).

b) EPLPSVD, EPTLS, EPHSVD ($N = 256$, $m = 128$ for EP, $N = 256$, $m = 50$ for the quantitation step).
Figure 3. Monte Carlo simulation results. The MC standard deviation of the Pi amplitude is displayed as a function of the noise standard deviation $\sigma$. The full-line represents the Cramer-Rao lower bounds.

a) LPSVD, TLS, HSVD (N = 256, m = 192).
b) EPLPSVD, EPTLS, EPHSVD (N = 256, m = 128 for EP, N = 256, m = 50 for the quantitation step).

4. Conclusion

EP used prior to any linear-prediction methods leads to an important reduction of the threshold SNR and results in near-optimal automatic quantitation methods. Moreover if the prediction orders are shrewdly chosen during the EP procedure and since the quantitation step needs only a few samples of the estimated signal, the computational cost is very similar to that of the linear-prediction methods using square matrices.

This work shows that these methods (EPLFSVD, EPHSVD, EPTLS) are well suited for NMR in vivo quantification. The Monte Carlo simulation proved that amplitude estimates have negligible bias and standard deviations nearly equal to their Cramer-Rao lower bounds even for low SNR.

Acknowledgements

The authors used the HSVD program kindly provided by R. de Beer and D. van Ormondt and the TLS program of the Netlib library written by S. Van Huffel. Thanks are due to these authors.

References

A Design Method For The Dyadic Wavelet Representation: Multiresolution Signal Decomposition

Gao Yang*,**, G. Zanellato*, H. Leich**

*Lernout & Hauspie Speech Products n.v., Belgium
**Lab. T.C.T.S., Bd Dolez 31, 7000 Mons, Faculté Polytechnique de Mons, Belgium

Abstract. In this paper a design method for dyadic wavelet representation and multiresolution signal decomposition is proposed, the properties of which are completely determined by a low-pass discrete filter \( H(\omega) \) and a high-pass discrete filter \( G(\omega) \). With the proposed method, a desired wavelet orthonormal basis, in the space \( L^2(\mathbb{R}) \) of measurable, square-integrable one-dimensional functions, can be obtained; the filter \( H(\omega) \) is easily controlled to make the scaling function \( \phi(x) \) and the wavelet \( \psi(x) \) have good localization properties in both the frequency and spatial domains. This design method leads to a faster pyramidal algorithm than the current one in [1] for the multiresolution decomposition and wavelet transform.

1. Introduction

Wavelets, filter banks, and multiresolution signal analysis, which have been used independently in the fields of applied mathematics, signal processing, image coding and computer vision, have recently converged to form a single theory. Main researches in this field consist in wavelet design (or design of filter banks), fast computation, error estimation, and their applications, respectively for continuous and discrete signals. The dyadic wavelet representation and the corresponding multiresolution decomposition is an important subject in this theory.

This paper is based on the work described in [1]. The similar notations as in [1] will be used here. \( Z \) and \( R \) denote the set of integers and real numbers respectively. For \( f(x) \in L^2(R) \) and \( g(x) \in L^2(R) \), the inner product of \( f(x) \) with \( g(x) \) is written as \( \langle g(x), f(x) \rangle \). The continuous Fourier transform of \( f(x) \) is \( \hat{f}(\omega) \). \( \hat{f}(Z) \) denotes the space of square-summable real sequences. The discrete Fourier transform of \( h(n) \in \hat{f}(Z) \) is indicated by \( H(\omega) \). \( H(\omega) \) means taking the conjugate operation of \( H(\omega) \).

Section 2 recalls the main research results for the dyadic wavelet transform and multiresolution representation. The detailed description of these results can be found in [1-2]. Section 3 proposes a design method, utilizing the results in Section 2. A fast pyramidal algorithm is then presented in Section 4. The paper is concluded in Section 5.

2. Dyadic Wavelet Transform and
Multiresolution Representation

A dyadic wavelet \( \psi(x) \) is designed using a scale function \( \phi(x) \) which satisfies the following equations:

\[
\hat{\phi}(2\omega) = H(\omega) \hat{\phi}(\omega)
\]

or,

\[
\hat{\phi}(\omega) = \sum_{k=1}^\infty H(2^{-k}\omega)
\]

where \( H(\omega) \) is an ideal or approximate halfband low-pass filter with impulse response \( h(n) \),

\[
|H(0)| = 1, \quad |H(\omega)| \neq 0 \quad \text{for} \quad \omega \in [0, \pi/2],
\]

and \( \phi(x) \) is a "low-pass" signal. The wavelet \( \psi(x) \) is determined from

\[
\hat{\psi}(2\omega) = G(\omega) \hat{\phi}(\omega)
\]

where \( G(\omega) \) is a discrete high-pass filter. An example of such function \( G(\omega) \) is given by

\[
G(\omega) = e^{-i\omega} \overline{H(\omega + \pi)}
\]

or,

\[
g(n) = (-1)^{1-n} h(1-n)
\]

If the low-pass filter satisfies

\[
|H(\omega)|^2 + |H(\omega + \pi)|^2 = 1
\]

it is ensured that \( \left( \sqrt{2}^j \phi(2^j x - n) \right)_{n \in Z} \) constitute an orthonormal basis of a space \( V_{2^j} \) of multiresolution
approximation at the resolution $2^j$ of $L^2(R)$ (the proof is given in [1][2]).

Let $O_{2^j}$ denote a space of the orthogonal complement of $V_{2^j}$ in $V_{2^{j+1}}$, i.e.,

$$O_{2^j} \oplus V_{2^j} = V_{2^{j+1}}$$

With (1) $-$ (7), we can obtain

$$|G(\omega)|^2 + |G(\omega + \pi)|^2 = 1$$

(8)

$$\sum_{m=-\infty}^{\infty} \tilde{\varphi}(2\omega + 2n\pi) \varphi(2\omega + 2n\pi) = 0$$

(9)

As a consequence, (8) and (9) are necessary and sufficient to ensure that

$$\left\{ \sqrt{2^n} \varphi(2^n x - n) \right\}_{n \in Z}$$

is an orthonormal basis of $O_{2^j}$ and

$$\left\{ \sqrt{2^n} \psi(2^n x - n) \right\}_{(n, j) \in Z^2}$$

is an orthonormal basis of $L^2(R)$ (the proof is given in [1][2]).

If we set $\phi_{2^j}(x) = 2^j \phi(2^j x)$, the approximate signal $A_{2^j} f(x) \in V_{2^j}$ of the signal $f(x) \in L^2(R)$ at the resolution $2^j$ is given by

$$A_{2^j} f(x) = 2^{-j} \sum_{n=-\infty}^{\infty} \left( f(u), \phi_{2^j}(u - 2^{-j} n) \right) \phi_{2^j}(x - 2^{-j} n)$$

(10)

$A_{2^j} f$ is thus characterized by the set of inner products:

$$A_{2^j} f = \left\{ \left( f(u), \phi_{2^j}(u - 2^{-j} n) \right) \right\}_{n \in Z}$$

$A_{2^j} f$ is called a discrete approximation of $f(x)$ at the resolution $2^j$. Similarly, the detail signal $B_{2^j} f(x) \in O_{2^j}$ of $f(x) \in L^2(R)$ at the resolution $2^j$ is given by

$$B_{2^j} f(x) = 2^{-j} \sum_{n=-\infty}^{\infty} \left( f(u), \psi_{2^j}(u - 2^{-j} n) \right) \psi_{2^j}(x - 2^{-j} n)$$

(11)

$B_{2^j} f$ is characterized by the set of inner products:

$$B_{2^j} f = \left\{ \left( f(u), \psi_{2^j}(u - 2^{-j} n) \right) \right\}_{n \in Z}$$

$B_{2^j} f$ is called a discrete detail signal of $f(x)$ at the resolution $2^j$. It contains the difference of information between $A_{2^j} f$ and $A_{2^{j+1}} f$. $D_{2^j} f$ is related to the orthonormal wavelet transform.

In order to compute efficiently $D_{2^j} f$ and $A_{2^j} f$ with computers, [1] has described a pyramidal algorithm which can be repeated in cascade as shown in Fig. 1 and Fig. 2 where $\hat{H}$ and $\hat{G}$ are respectively the mirror filters with impulse responses $\hat{h}(n) = h(-n)$ and $\hat{g}(n) = g(-n)$. Fig. 1 describes the decomposition of a discrete approximation $A_{2^{j+1}} f$ into an approximation at a coarser resolution $A_{2^j} f$ and a discrete detail signal $D_{2^j} f$ at the resolution $2^j$. Fig. 2 shows the reconstruction of $A_{2^{j+1}} f$ from $D_{2^j} f$ and $A_{2^j} f$.

![Fig. 1 Decomposition of $A_{2^{j+1}} f$ into $D_{2^j} f$ and $A_{2^j} f$.](image1.png)

![Fig. 2 Reconstruction of $A_{2^{j+1}} f$ from $D_{2^j} f$ and $A_{2^j} f$.](image2.png)

**3. A Design Method of $H(\omega)$ AND $G(\omega)$**

The properties of wavelet representation and multiresolution signal decomposition are completely determined by the filters $H(\omega)$ and $G(\omega)$. The following properties are often desired: 1) the wavelets constitute an orthonormal basis of $O_{2^j}$ or $L^2(R)$; 2) the scaling functions constitute an orthonormal basis of $V_{2^j}$ in $L^2(R)$; 3) the wavelets and scaling functions have good localization properties in both the frequency and spatial domains, i.e., they are well band-limited in the frequency domain and rapidly energy-decayed in the spatial domain; 4) the computations are fast performed. With the conditions (5) and (7) it is not often easy to find a filter $H(\omega)$ which leads to the desired properties. This section proposes a design method of $H(\omega)$ and $G(\omega)$, with which the above
four properties can be easily obtained. The results are described by the following theorems.

**Theorem 1:** Let $H(\omega) = |H(\omega)| e^{i\theta(\omega)}$, if $H(\omega)$ satisfies

\[
\begin{align*}
|H(\omega)| &= \cos(\theta(\omega)) \\
\theta(0) &= 0, \quad \theta(\pi) = ± \pi/2 \\
\theta(\pi - \omega) + \theta(\omega) &= ± \pi/2, \quad \text{for} \quad \omega \in [0, \pi]
\end{align*}
\]  
(12) — (14) with (1) ensure that \( \left( \sqrt{2^j} \phi(2^j x - n) \right)_{n \in \mathbb{Z}} \) is an orthonormal basis of a multiresolution approximation space \( V_2 \) in \( L^2(\mathbb{R}) \).

The proof of Theorem 1 is easy. Since \( h(n) \) is required to be a real discrete low-pass filter, \( \theta(\omega) \) satisfies

\[
\begin{align*}
\theta(-\omega) &= -\theta(\omega) \\
\theta(\omega + 2k\pi) &= \theta(\omega), \quad \text{for} \quad k \in \mathbb{Z}. \\
\theta(\omega) &= ± \pi/2, \quad \forall \omega \in [0, \pi/2]
\end{align*}
\]  
(15) — (17) With (14)—(16), we have

\[
\theta(\omega + \pi) = \theta(\omega) ± \pi/2,
\]  
(18)

From (12) and (18), we can demonstrate that (7) is satisfied. Thus Theorem 1 is proved using the result in Section 2.

**Theorem 2:** Let

\[
G(\omega) = 1 - H(\omega)
\]  
(19)

(1), (4), (12)—(14) and (19) ensure that \( \left( \sqrt{2^j} \psi(2^j x - n) \right)_{n \in \mathbb{Z}} \) is a dyadic orthonormal basis of \( \mathcal{O}_{2^j} \) and \( \left( \sqrt{2^j} \psi(2^j x - n) \right)_{(n,j) \in \mathbb{Z}^2} \) is a dyadic orthonormal basis of \( L^2(\mathbb{R}) \).

From (12) and (18), we obtain

\[
H(\omega) + H(\omega + \pi) = 1
\]  
(20)

With (20), (19) can be expressed as

\[
G(\omega) = H(\omega + \pi)
\]  
(21)

hence, (8) is satisfied with (7) and (21). *Theorem 1* shows that \( \left( \phi(x - n) \right)_{n \in \mathbb{Z}} \) is an orthonormal basis of \( V_1 \). This results in

\[
\sum_{k=-\infty}^{\infty} |\phi(\omega + 2k\pi)|^2 = 1
\]  
(22)

Using the above equation with (1) and (4), we have

\[
\sum_{n=-\infty}^{\infty} \phi(2\omega + 2n\pi) \overline{\phi(2\omega + 2n\pi)}
\]  
(23)

By inserting (12), (18) and (21) into the above expression, we can demonstrate (9). Since (8) and (9) are satisfied, *Theorem 2* is proved. Of course, if in *Theorem 2* the definition (19) is replaced by (5), we can also obtain a wavelet orthonormal basis; however (19) leads to a simpler algorithm as will be described.

A typical example is to define \( \theta(\omega) \) as

\[
\theta(\omega) = -\left( \omega/\pi \right)^p \pi/4, \quad \omega \in [0, \pi/2]
\]  
\[
= -\theta(\pi - \omega) - \pi/2, \quad \omega \in [\pi/2, \pi]
\]  
(24)

The localization properties of the wavelet and the scaling function can be easily controlled by adjusting the parameter \( p \). Fig. 3 gives three examples for \( p = 1, 10, +\infty \), respectively. We can see that the example for \( p = 10 \) have good localization properties in both the frequency and spatial domains.
4. A Fast Pyramidal Algorithm

By using (19), the pyramidal algorithms shown in Fig. 1 and Fig. 2 for the decomposition and reconstruction of $A_{2j+1}^d f$ are modified into those illustrated in Fig. 4 and Fig. 5. There remains only one digital filter ($h(n)$ or $\tilde{h}(n) = h(-n)$) in Fig. 4 and Fig. 5; besides $h(2k) = 0$, for $k \in \mathbb{Z}$ and $k \neq 0$, due to the property of (20). Therefore, the new algorithms described by Fig. 4 and Fig. 5 are faster and simpler than the current ones shown in Fig. 1 and Fig. 2.

Fig. 4 Fast Decomposition of $A_{2j+1}^d f$ .

Fig. 5 Fast Reconstruction of $A_{2j+1}^d f$

5. Conclusion

This paper proposes a design method for the dyadic wavelet representation and multiresolution signal decomposition. With this method we can easily obtain a wavelet and a scaling function, which result in the desired properties.

References


Estimation of the order of the AR part of ARMA models with application to frequency estimation

David BRIE, Dominique DAL PONTE, Marc TOMCZAK, Alain RICHARD

Centre de recherche en automatique de Nancy - CNRS - URA 821 , BP 239, 54506 Vandoeuvre Cedex, France,
E-Mail : brie@cran.u-nancy.fr

Abstract: This paper sets forth an approach which enables to associate an order estimation procedure with any identification algorithm. The particular problem of order estimation of the AR part of ARMA model is then addressed. The performances of this method are assessed regarding the problem of frequency resolution of AR spectral estimation for sinusoids in white noise. It is shown by numerical simulations that, among three different penalty functions chosen, Akaike’s one provides the better results.

1. Introduction

The problem of order estimation has received considerable attention in the literature since the early works of Neyman Pearson ([11]) who solved it using an hypotheses testing approach. This involves the comparison of the Log-Likelihood Ratio Statistic (LRS) with some threshold to achieve a desired level. This approach supposes some arbitrariness which may be avoided by choosing the order which minimizes the LRS. This leads to minimize the well-known AIC criterion ([2]) which includes two terms, the Log-likelihood statistic and a penalty function preventing from overparametrization.

In this paper, an alternative approach which has been originally proposed in [3], is presented. Its main idea is to use the Local Test Statistic (LTS) in lieu of the Log-likelihood statistic. Section 2 recalls how LTS may be used for order estimation and in section 3 this method is applied to the problem of order estimation of the AR part of an ARMA model which may be of great interest for some situations such as frequency estimation of sinusoids in noise signals. The fourth section is devoted to the evaluation of the performances of the approach proposed regarding the particular problem of frequency resolution for sinusoids in white noise signals.

2. Local Test Statistic for order estimation

Let \( \theta^* \) be the true parameter of the stochastic process \( X = \{x_1, \ldots, x_N\} \). \( \overline{m} \) be a pre-specified upper bound on the order \( m \) and \( \hat{\theta}^{(m)} \) be the estimate of \( \theta^* \) under order \( m \). For simplicity the possible time dependence of \( \hat{\theta}^{(m)} \) and \( \theta^* \) has been omitted. Define now:

\[
\hat{\theta}^{(m)}_m = [\hat{\theta}^{(m)}_0^T, 0, \ldots, 0]^T \quad \text{for} \quad (\overline{m} - m) \text{ zeros}
\]

that is the vector \( \hat{\theta}^{(m)}_m \) extended with \( \overline{m} - m \) zeros so that \( \text{dim}(\hat{\theta}^{(m)}_m) = \overline{m} \). Then, the problem of order estimation can be stated as the following local hypotheses testing:

\[
H_0: \theta^* = \hat{\theta}^{(m)}_m \quad \text{against} \quad H_1: \theta^* = \hat{\theta}^{(m)}_m + \frac{8\theta}{\sqrt{N}}
\]

As shown in [4], it can be solved using the LTS:

\[
s(m) = D_N^T (m) R^{-1} D_N (m) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} H(\hat{\theta}^{(m)}_m, \xi_k)
\]

\[
R = \sum_{k \in \mathbb{Z}} \text{cov}(H(\hat{\theta}^{(m)}_m, \xi_k), H(\hat{\theta}^{(m)}_m, \xi_{k+n}))
\]

\[
H(\hat{\theta}^{(m)}_{m+1}, \xi_n)
\]

is the function used in adaptive algorithms of the form:

\[
\hat{\theta}_n = \hat{\theta}_{n-1} + \Gamma_n H(\hat{\theta}^{(m)}_{m+1}, \xi_n)
\]

where \( \hat{\theta}_n \) is the parameter vector at sample \( n \), \( \xi_n \) is a controlled markovian process, \( \Gamma_n \) is a gain matrix depending on the choice of the vector field \( H(\hat{\theta}^{(m)}_{m+1}, \xi_n) \) and on stationarity assumptions.

A first approach is to compare the LTS \( s(m) \) with some threshold to decide whether the model is good enough or not. A more systematic approach is to choose the model which gives the smaller bias \( \delta \theta \), that is the one which minimizes the LTS. However, it is known that the local likelihood statistic, i.e. the LTS associated to the Maximum Likelihood Estimate (MLE), is asymptotically equivalent to the LTS ([4]), and then has to be corrected by a penalty function as it is done for the AIC. This is necessary because the statistic is guaranteed to decrease or stay the same as the model order increases. The inconsistency of the AIC in the AR case ([5]) has led to the use of other penalty functions of the form "m f(N)" (see [6] [7] for example). Extending these remarks to the more general case considered here, the following procedure is suggested:

\[
\hat{m} = \arg \min_{m \in \mathbb{N}} (s(m) + m f(N))
\]

This approach allows to associate a specific order estimation procedure to a large class of estimation algorithms, provided that some estimate of the covariance matrix \( R \) is available. In [3], this approach has been mainly studied for the case of AR models. In particular, an order recursive form of the statistic \( s(m) \) has been derived. Let us now develop the reasons which motivated the derivation of a procedure for estimating the order of the AR part of an
ARMA model. It is well known (8) that noisy AR(m) are ARMA processes. If the noise is assumed to be white and gaussian, the MA part has the same order than the AR part but if the noise is colored, the order of the MA part becomes greater. However, all the spectral information of the original process still remains in the AR part of the ARMA process and thus, only this part has to be estimated. For this purpose, the Instrumental Variable Estimate (IVE) [9] is particularly adequate. But if, for example, the Akaike criterion is used for the order estimation task, a MLE of all possible order models is needed. This supposes the use of time-consuming non-linear optimization procedures to determine the ARMA(p,q) models. Moreover, without any a priori knowledge, m and q estimations have to be performed for exhaustive search. The approach proposed enables to estimate only the order of the AR part without having any knowledge of the MA part order and avoids estimation of the MA parameters.

3. Estimation of the order of the AR part of ARMA models

The following ARMA(m, q) process is now considered:

\[ x_t = \sum_{i=1}^{m} a_i x_{t-i} + v_t, \quad v_t = \sum_{i=1}^{q} b_i u_{t-i} \]  

(8)

where \( u_k \) is a white gaussian noise. The problem is to estimate the order of the AR part without any knowledge of the order of the MA part. This may be very useful for the spectral analysis of noisy sinusoidal signals, recalling that they may be seen as ARMA signals with all the sinusoidal spectral information contained in the AR parameters. Assuming that N observations are available, it can be written:

\[ X^{(m)} = X^{(m)} \theta^{(m)} + V = \begin{bmatrix} X^{(i)} \end{bmatrix} \theta^{(m)} + V \]  

(9)

where \( X^{(i)}, \theta^{(m)} \) and \( V \) are defined as:

\[ X^{(i)} = \begin{bmatrix} x_{i+1}, \ldots, x_{i+m} \end{bmatrix} \]  

(10a)

\[ \theta^{(m)} = \begin{bmatrix} a_1, \ldots, a_m \end{bmatrix} \]  

(11a)

\[ V = \begin{bmatrix} v_{i+1}, \ldots, v_{i+m} \end{bmatrix} \]  

(11b)

However, the noise \( V \) is now colored and thus the least-squares estimate of \( \theta \) is biased. This problem may be avoided by using the IVE with delayed observations which is known to be consistent if the delay \( M \) satisfies \( M \geq q \) (10b). So, provided that an upper bound \( \bar{q} \) for the order \( q \) of the MA part is available, the choice \( M = \bar{q} \) ensures the consistency. In the case of narrow band signals, it is preferable to use the Overdetermined Instrumental Variable Estimate (OIVE) (9). This method which is strongly connected to the overdetermined modified Yule Walker method enables to take into account high order correlation coefficients. The OIVE is given by:

\[ \hat{\theta}^{(m)} = (\Psi_C^T \Sigma^{m})^{-1} \Psi_C X^{(m)} \]  

(12)

where \( A'' \) denotes the pseudo inverse of the matrix \( A \) and \( \Psi_C \) is the instrumental matrix defined by:

\[ \Psi_C^T = \begin{bmatrix} \psi(0) & \ldots & \psi(C) \end{bmatrix} \]  

\[ \psi(i) = \begin{bmatrix} x_{N-M+i-1} \ldots, x_{M+i} \end{bmatrix}^T \quad i = 1, \ldots, C \]  

(13)

C is referred to as the overdetermination factor.

Note that the order recursive lattice version of the OIVE developed in [11] is used because it provides all the models of successive orders and exhibits good behaviours from a numerical point of view. Now considering:

\[ \hat{\theta}^{(m)} = \left( \begin{array}{c} \hat{\theta}^{(m)} T \\vdots \quad 0 \end{array} \right) \]  

(14)

where \( \hat{\theta}^{(m)} \) is the OIVE of \( \theta \) under order \( m \), delay \( \bar{q} \) and overdetermination factor \( C \), the corresponding local instrumental statistic \( s(m) \) ([(4) [12]) is equal to:

\[ s(m) = \frac{1}{N - \bar{q} - C} \Sigma_i X_i \Psi_C \Sigma R^{-1} \Sigma_i \Psi_C X_i^T \]  

(15)

\[ E = X^{(0)} - X^{(m)} \hat{\theta}^{(m)} = \left( \begin{array}{c} \epsilon_1 \ldots, \epsilon_{N-q-C} \end{array} \right) \]  

(16)

The matrix \( R \) may be consistently estimated as [(4)]:

\[ \hat{R} = \frac{1}{N - \bar{q} - C} \Sigma_i \sum_{j=1}^{N-q} \sum_{k=i}^{N-q} \epsilon_j \epsilon_{k-j} (Z_k^T Z_{k-j}) \]  

(17)

\[ Z_j = \begin{bmatrix} x_{j-q} \ldots, x_{j-q-C} \end{bmatrix}^T \]  

(18)

However, this estimate is not always positive definite and gives raise to numerical problems. So another estimator, proposed in [12], which is always positive definite, has been used:

\[ \hat{R} = \frac{1}{K(L+1)} \Sigma_{k=1}^{K} \left( \sum_{j=k(2L+L)}^{k(2L+L)-L} \Sigma_{j=k(2L+L)-L}^{k(2L+L)} \epsilon_j Z_j \right) \]  

(19)

with

\[ K = \frac{N - \bar{q} - C}{2m + L} \]  

(20)

L is a window length fixed in order to optimize the tradeoff between accuracy and numerical conditioning of the estimator.

The order is then chosen as:

\[ \hat{m} = \arg \min_{m \in \mathbb{N}} \left( s(m) + m f(N) \right) \]  

(21)

It should be noted that this approach offers the opportunity to estimate both orders, \( m \) and \( q \), of the ARMA process. Once \( m \) is known, the order of the MA part may be estimated as the one which minimizes the criterion corresponding to the method used to identify the MA part. Numerical simulations [(3)] have shown that the LTS statistic associated to the Rissanen penalty function yields consistent AR order estimates provided that the poles of the ARMA model are not too close. The next section deals with the performances of this approach with respect to the problem of frequency resolution.

4. Performance evaluation : the case of frequency estimation

Classically, the performances of parametric spectral estimators are assessed considering that the order is known
a priori. In this section, we proposed to investigate the performances of the OIVE associated to the LTS method for spectral analysis. Because it is known that parametric spectral estimators have a poor magnitude resolution, we only considered the performances with respect to the frequency resolution, for sinusoids in white noise signals. In this framework, our aim is to determine the penalty function which provides the best results. The simulated signals are generated from:

\[ y_n = A \sin(2\pi f_1 n) + A \sin(2\pi (f_1 + \delta f) n) + v_n \]  

(22)

where \( n = 1, 2, \ldots, N \). The noise \( v_n \) is normally distributed with zero mean and variance unity and the normalized frequency \( f_1 \) is uniformly distributed over the frequency range \([0.1, 0.4]\). This choice has been made in order to avoid boundary problems. The value of \( \delta f \) is varying from \(1/N\) to \(30/N\) by steps of \(1/N\). In all cases, the magnitude \( A \) is chosen so as to give a desired signal to noise ratio (SNR) (0 - 10 - 20 dB). The number of samples considered is \( N = 512 \). For each case, 50 Monte Carlo trials are performed. For each realization, the frequency estimator is computed with the order \( m \) estimated by the LTS criterion. Three penalty functions are used:

- \( m_f(N) = 2n \), \( m_f(N) = m_i \log(N) \), and \( m_f(N) = m_i \log(\log(N)) \) that correspond respectively to Akaike’s penalty function (Apf), Rissanen’s penalty function (Rp) and Hannan’s penalty function (Hpf) [2], [6], [7].

Following [13], the sinusoidal frequencies are estimated from the angular positions of the roots of the AR(\( m \)) polynomial rather than from maximum peaks of the PSD. Noting that a noisy sinusoid can be viewed as a limiting form of an ARMA(2,2) process with two complex conjugated poles infinitesimally close to the unit circle, we introduce the following Frequency Mean Square Error (FMSE) criterion to evaluate the accuracy of the proposed order estimation procedure and to determine the best penalty function:

\[ \text{FMSE} = \frac{1}{50} \sum_{i=1}^{30} \left( \frac{\hat{f}_i - f_i}{\hat{f}_i + f_i} \right)^2 \]  

(23)

where the estimated frequencies \( (\hat{f}_i)_{i=1,2} \) correspond to the two pairs of poles which are the nearest of the unit circle. Note that all trials where the estimated order was smaller than 4 have been ignored and thus have been performed again. However, they only represent less than 1% of the total number of trials. The results achieved are compared to the best results obtained by performing, for each trial, a minimization of the FMSE criterion with respect to all the different possible orders.

Before discussing the results of the simulations, some comments should be made about the choice of the order estimation procedure tuning parameters. All trials have been performed with \( m_i = 50, \ q_i = 50, \ C = 100 \). The overdetermination factor \( C \) should be chosen large enough to ensure a sufficient frequency resolution of the OIVE but small enough to avoid numerical problems, especially those induced by the inversion of the matrix \( R \) (eq. 19). The choice of the window length \( L \) is important. Small values will ensure accuracy while large ones improve the conditioning. Successive trials have shown that \( L = 50 \) was a good candidate.

The results of the numerical simulations are depicted on the different figures.

First, it appears that the FMSE criterion is a non increasing function of the SNR and that the difference between the optimal results and those provided by the proposed approach, decreases when the SNR increases. When the SNR increases, the three penalty functions tend to behave identically. For all the cases considered here, Apf provides better results than Hpf and Rp. However, note that Hpf performs similarly than Apf. This is not surprising because, for \( N = 512, \ m_f \log(\log(N)) = 2m \). However, in the case of a 20 dB SNR, it can be seen from the different histograms corresponding to the three penalty functions that for a modelling purpose, the Rp performs better than the two others because it provides the right order with a greater probability. This may be connected to the theoretical result of [5], which states that, in the case of AR models, Rp and Hpf provides consistent order estimates while Apf tends to overestimate the order. This behaviour of the Apf is not necessarily unfavourable in the case of spectral analysis where it is well known that an overparametrization is needed to ensure a sufficient resolution degree, especially for very nearby components. These results have to be interpreted carefully because the FMSE criterion only takes into account a number of spectral components that is known a priori. In addition, the problems of spurious peaks and dynamic resolution are not accounted for by the criterion.

5. Conclusion

This paper presents an application of the LTS to order estimation. It's main interest is to build, in a systematic way, an order estimation procedure well adapted to the identification method chosen. In particular, this approach associated with the OIVE enables to estimate the order of the AR part of an ARMA model. Numerical simulations have been performed to evaluate the performances of the procedure with respect to the problem of frequency estimation. They show that, among the three penalty functions chosen, Apf provides the best results from a frequency resolution point of view while Rp is better from a modelling point of view. At the present time, this approach is compared with the procedures proposed in [14]. Future works will be directed at investigating the case of short segments for which specific penalty functions [15] may be used.

References


figures (1), (2), (3) caption :
- FMSE with Akaike's penalty function
- FMSE with Hannan's penalty function
- FMSE with Risssannen's penalty function
- Optimal FMSE

figures (4), (5), (6) caption :
- Akaike penalty function
- Hannan's penalty function
- Risssannen's penalty function
Overlapped Block Structure for Fast Time-Varying Orthogonal Wavelet Packet Transform Algorithms

Andrzej DRYGAJLO

Signal Processing Laboratory, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland, Tel/Fax: +41 (21) 693 4328/693 4660, E-mail: drygajlo@elde.epfl.ch

Abstract. In this paper we describe a unified framework - overlapped block structure - where discrete orthogonal wavelet packet transforms can be realized efficiently by means of fast algorithms. Our prime interest is in the representation of nonstationary discrete-time signals in terms of wavelet packets used as time-varying discrete orthogonal systems, and we concentrate on the fast transform algorithms for such systems. We analyse lattice-polyphase and lattice-tree structures which are common for multiresolution multirate filter banks and wavelet packet transforms. The purpose of the present paper is to compare and contrast the wavelet packet based approach to the traditional techniques for fast orthogonal transform algorithms. We indicate some results from lattice-structured filter banks which are useful for the design of fast lapped wavelet packet transform algorithms. A time-varying structure that is based on fast algorithms of orthogonal transforms and their orthogonal sub-transforms is presented. The discrete approach proposed readily provides numerical algorithms and gives a new way of looking at multiresolution multirate filter banks; it describes them as a temporal multiresolution decomposition rather than as a subband frequency decomposition.

1. Introduction

The last five years have seen the development of a new class of signal representation - the wavelet packet transform - which has replaced multiresolution signal analysis on a rigorous footing and has opened up new possibilities for efficient multiresolution methods of nonstationary signal processing. In particular, this technique provides flexible combinations of time-support and spectral bandwidth, and represents a powerful generic tool for efficient time-spectral signal decompositions. Discrete block transforms (e.g. DFT, DCT, etc.) and uniform-bandwidth filter banks are well-understood and widely used analysis tools but they lack the flexibility for time-spectral analysis mentioned above [1].

Perfect reconstruction unequal-bandwidth multirate filter banks as well as orthogonal wavelet and wavelet packet transforms, are relatively newer structures with more flexible multiresolution features [2,3,4,5]. The orthogonal wavelet transform is an important breakthrough and has already had a significant impact in digital signal/image processing, numerical analysis, and other areas [6]. It is well known, however, that the wavelet transform gives just one of many possible time-spectral decompositions. The orthogonal wavelet-packet transform approach, especially if used in a time-varying framework, may therefore offer powerful forms of adaptive analysis in a more flexible way than a wavelet transform. What is still lacking, however, is a thorough understanding of what orthogonal transform algorithms can do for the computationally-intensive, but naturally appealing notions of a signal-adaptive transform or filter bank which preserves the perfect reconstruction property and realizes the most desirable time-spectral decomposition.

It is well known that orthogonal wavelet packet transforms can be designed by hierarchical association of perfect reconstruction paraunitary filter banks (tree-structured filter banks) [7]. Filter banks, however, were studied from the viewpoint of subband decomposition, which masks the time decomposition properties. These properties should be transparent in time-varying wavelet packet transform applications. On the other hand the suitability of an orthogonal transform or a filter bank in a given application depends on the existence of an efficient computational algorithm.

It should be clear that there are many more similarities than differences between signal processing with overlapped block transforms and with perfect-reconstruction multiresolution multirate filter banks. The similarity can very easily be seen with the help of the lattice structures of paraunitary filter banks.

2. Lattice Structures

The main characteristic of a lattice structure is that it is composed of a cascade of orthogonal operators and delays [3]. The orthogonal operator blocks can be combi-
nations of rotations, symmetries and permutations (e.g. as in Fig. 1) [8].

\[
\begin{array}{c}
\gamma_1 \\
\circlearrowleft \\
\gamma_1 \\
+ \\
\gamma_1 \\
\gamma_1 \\
+ \\
\gamma_1 \\
\end{array}
\]

clockwise rotation + symmetry + permutation

Figure 1

The lattice structure implements, among others, two-band maximally-downsampled paraunitary FIR filter banks of even length \( L \), having the perfect-reconstruction property, e.g. as in Fig. 2 \( (L = 6) \) [8,11].

It also has a hierarchical property, i.e., higher order paraunitary filter banks can be obtained from those of lower order by adding more lattice sections. Another important property of the lattice structure is that by changing the rotation parameters \( \gamma_i \) we can generate all paraunitary filter banks. They depend on the desirable filter characteristics and can be chosen in order to compute Daubechies filters [5], binomial filters [1], Malvar filters [9] and other paraunitary filter sets [3]. Lattice structures are computationally efficient and preserve perfect-reconstruction after coefficient quantization. They are also well-suited to VLSI implementations, such as CORDIC processors. These properties make them particularly important with reference to orthogonal wavelet packets. The rotation parameters \( \gamma_i \) can be directly computed by an optimization procedure, that minimizes a cost function like the stop-band energy of one of the two filters. They can also be indirectly computed from the impulse response coefficients of one of the filters, if it has been designed by another suitable paraunitary filter design method (e.g. spectral factorization [10] or eigenfilters [3]). Thus, if we constructed the tree structure (binary or M-ary) using lattices, we could generate all orthogonal wavelet packet bases by manipulating the lattice coefficients. To keep the illustration simple an example was chosen as in Fig. 3 \( (L = 4) \).

3. Overlapped Block Structure

The two-channel lattices can be re-arranged in an overlapping block transform lattice as depicted in Fig. 4 \( (L = 6) \) [8,11].

In this polyphase structure all downsamplers and delays can be moved to the left side, and the rotation parameters \( \gamma_i \) are the same as for the cascaded structure. In a similar way, we can build a polyphase structure for the filter bank of Fig. 3. The result appears in Fig. 5.

In this paper we indicate some results from lattice-structured filter banks which can be useful for the design of fast time-varying overlapping orthogonal wavelet packet transform algorithms. In this case, starting from a complete uniform tree or a polyphase structure a sub-tree or a polyphase sub-structure is selected, that matches at best the signal time-spectral characteristics following a given criterion. Such sub-structures allow the partitioning of the time-spectral domain into nonuniform tiles in connection with the time-spectral contents of the signal to be analyzed. By modifying the wavelet packet basis in time, adaptive tree structures can be implemented to process nonstationary signals. Examples of applications of such methods have been already presented and they show promising results [12,13,14].
4. Perfect Reconstruction Throughout Transitions

An obvious question is whether we can find an algorithm for fast implementation of time-varying wavelet packet bases maintaining perfect reconstruction throughout transitions. To obtain this algorithm it is necessary to know how adjacent time-invariant wavelet packet transform algorithms should overlap in order to avoid transition distortions due to segmentation. The overlapping polyphase structure is particularly useful to answer the above questions.

One major problem of tree structures is that the signal samples of the pruned lattice branch are lost, when switching abruptly from one structure to another (Fig. 6).

![Figure 6]

It is a natural consequence of suppressing the delay blocks and the samples they hold. Such a problem does not happen with the polyphase structure, since all samples are kept in the delay chain, which is never modified when different sub-structures are chosen (e.g. as in Figs. 5 and 7).

![Figure 7]

The procedure that generalizes the approach already presented above is illustrated in Fig. 8.

![Figure 8]

For example, we want to perform an overlapping block analysis according to the time-spectral tiling in Fig. 9 and the corresponding time-scale decomposition using first the non-overlapped time-varying Hadamard-Haar transform \( H_0(z) = 0.5 + 0.5z^{-1} \) (Fig. 10) and then the Daubechies length-4 filters (Fig. 11) [5,11,15,16].

![Figure 9]

In this second case the lowpass filter transfer function is given by \( H_0(z) = 0.4827 + 0.8365z^{-1} + 0.2241z^{-2} - 0.1294z^{-3} \), and the lattice parameters are \( \gamma_0 = 1.7321 \) and \( \gamma_1 = -0.2679 \).

![Figure 10]

![Figure 11]

Two polyphase lattice structures are necessary and at
n = 4 we switch from one structure to another. In such a
switching not a simple sample has been lost, but extra
data for describing the transition structure is inevitable.
For practical realization the complete structure from
Fig. 5 can be used where flow graph nodes provide
all possible time-spectral coefficients. In this case, to
calculate each block of time-spectral coefficients, it is
only necessary to realize non-overlapping and partially
overlapping butterflies of the flow-graph. The inverse
time-varying transform can be accomplished by inverting
the flow graph.

5. Conclusion

In this paper we presented a procedure for the design
of fast orthogonal wavelet, wavelet packet and
wavelet packet transform algorithms using lattice
structures which guarantee distortionless processing.
This procedure generalizes an approach presented
in previous papers and leads to overlapping polyphase
block transform structures. The advantage of using such
structures over tree structures was justified especially
in the context of time-varying systems.

From the presented examples we see that analysis struc-
tures based on the developed principle can provide an
extremely effective way to extract the multiresolution
time-spectral characteristics of nonstationary signals. It
is also worth mentioning that the scheme presented in
this paper can be applied to design other efficient mul-
tiresolution signal processing systems such as biorthog-
onal linear phase filter banks, orthogonal wavelet packet
transforms which cannot be generated using a tree struc-
ture, multiresolution multi-carrier modulators, trans-
multiplexers and time-spectral scramblers [17,18].

References

[1] A. N. Akansu, R. A. Haddad, "Multiresolution Sig-
Packet Algorithms," in "Ondelettes et Paquets
d'Ondes" (France), Jun 1991, pp. 31-99.
[5] I. Daubechies, "Ten Lectures on Wavelets", SIAM,
[6] M.B. Ruskai et al. (eds), "Wavelets and Their
Applications", Jones and Bartlett, Boston, 1992.
[7] A. K. Soman, P. P. Vaidyanathan, "On Or-
thonormal Wavelets and Paraunitary Filter Banks,
1993, pp. 1170-1183.
Orthogonal Wavelet Packet Implementations," Proc.
ECCOTD-93, Davos, Sept. 1993, pp. 1379-1384.
[9] H.S. Malvar, "Signal Processing with Lapped Trans-
[10] M. J. T. Smith, T. P. Barnwell, "Exact Reconstruc-
tion Techniques for Tree-Structured Subband Coders,
No. 3, June 1986, pp. 434-441.
rithms for Multiresolution Time-Sequence Signal Decompo-
sition and Processing," Proc. of SPIE, vol. 2034, Math-
ematical Imaging: Wavelet Applications in Signal and
Image Processing, San Diego, CA, July 1993, pp. 349-
358.
[12] E. Westfeld, M.V. Wickerhauser, "Adapted Lo-
lcal Trigonometric Transforms and Speech Processing",
No. 12, Dec.1993, pp. 3596-3604.
Vetterli, "Tilings of the Time-Frequency Plane: Con-
struction of Arbitrary Orthogonal Bases and Fast Tiling
Algorithms", IEEE Trans. Acoust. Speech, Signal Pro-
Speech, Signal Processing, Vol. 41, No. 12, Dec.1993,
pp. 3293-3305.
and Their Multiresolution Properties," in Signal Pro-
cessing VI: Theories and Applications, J. Vandewalle et
Packet Transmultiplexing Using Lattice Structures,
Proc. of GRETSI Conference, Juan-les-Pins, Sept. 1993,
pp. 431-434.
[18] A. Drygajlo, "Multiresolution Time-Sequence Speech
Processing Based on Orthogonal Wavelet Packet Pulse
Forms," Proc. of EUROASPEECH'93, Berlin, Sept. 1993,
pp. 147-150. 671
Adaptive Three-Dimensional Motion-Compensated Coding of Image Sequences for Multiresolution Applications.

J.-P. Leduc

IRISA/INRIA, Campus de Beaulieu 35042, Rennes Cedex, France, Tel: +33 99 84 74 25 / Fax: +33 99 88 38 98, E-Mail: leduc@iri.fr

Abstract. This paper presents a three-dimensional coding algorithm for digital image sequences. The scheme is based on adaptive wavelet transforms analyzing spatio-temporal domains. The video signal analysis aims at maximizing the coding compression ratio under the constraint of graceful degradations of the subjective quality. The analysis operation relates on adaptive filters built as lattices where the coefficients are estimated by a stochastic gradient algorithm to yield optimum decorrelation under the filtering constraints of quasi-linear phase and maximum regularity. Three-dimensional multiresolution coding schemes provide an elegant generalization to classic two-dimensional spatial coding schemes since they add abilities of easily performing format conversions and interpolations. In the scheme under investigation, the video signal is first decomposed into temporal subbands on which spatial filters are further applied to generate the final spatio-temporal subbands. Furthermore, a motion compensation at pixel accuracy is also applied on the lower temporal frequencies. The temporal filtering is adapted to the global or the local motions in the scene. The spatial filtering performs a time-varying analysis adapted according to the correlation structure within the images. The adaptation implies not only a signal-adapted frequency splitting but also adaptive dyadic filtering cells at each level of the decomposition. To complete the coding scheme, the coding process is further composed of a scalar quantizer which exploits the spatio-temporal three-dimensional psycho-visual properties of the Human Visual System and eventually of an entropy coder which performs the bit rate compression. An output buffer regulation enables eventually constant bit rate transmissions.

1 Introduction

Most of the coding algorithms designed for the bit rate compression of digital image sequences are usually built with hybrid two-dimensional spatial coding analysis using motion compensation prediction. The extension towards including a temporal analysis is a natural progression. Some pioneering research works have indeed already been devoted to three-dimensional (spatio-temporal) coding analyses which exploit the subband decomposition technique. Most of the researches in that field have restricted the temporal filtering to the Haar wavelet. Karlsson, Lewis and Knowles have studied 3-D structures without motion compensation. Kronander, Ohm and Woods have experimented the motion compensation in their coding schemes. Recently, Dubois has introduced the concept of motion-compensated filters. The three-dimensional (spatio-temporal) transform proposed in this paper steps towards a more advanced approach. Since the adjunction of an efficient motion compensation to subband coding performs a shift of the three-dimensional spectrum towards lower temporal frequencies, motion compensation turns out to be desirable to attain higher energy compaction (or decorrelation) and to substantially increase the compression ratios of the resulting algorithm. Incidentally, the motion compensation can also reduce possible temporal aliasing effects. The three-dimensional wavelet analysis implements a multiresolution decomposition. Moreover, since this coding scheme involves simultaneously a temporal and a spatial decomposition, it generalizes the concept of spatial scalability to embody the temporal axis and offers some elegant and efficient solutions to other problems related to high-definition television like the spatio-temporal prediction in predictive coding, the spatio-temporal interpolation for format conversions allowing direct access to any kind of intermediate image sequence formats, and the spatio-temporal smoothing for noise reduction.

The three-dimensional multiresolution algorithm makes use of adaptive time-varying filters. In the spatial dimensions, the filters are adapted locally to the correlation scene content. Adaptive dyadic para-unitary filter banks are intended to be used in the lattice form. This lattice structure guarantees structural perfect reconstruction and para-unitary properties and enables the use of adaptive signal processings, i.e. a gradient-type algorithm, to determine the filtering coefficients which minimize a cost function. The temporal signals behave somewhat differently from those of the spatial dimensions in the sense that the temporal frequencies depend on both the spatial details and the object velocities. Temporal filters can efficiently exploit the knowledge of the
estimated motion to perform an adaptation according to the local scene motions. This concept is referred to as the motion-compensated filtering, it makes the analysis more effective than a fixed filtering in many circumstances. The spatio-temporal filtering adaptivity aims at achieving optimum decorrelation under constraints of preserving graceful visual properties to the reconstructed image. Quasi-linear phase para-unitary wavelets or bi-orthogonal may yield that goal. The block-matching motion estimation and compensation of classical coding schemes can further be improved by refining the motion description and splitting the motion fields into object-oriented segments. Such motion-object-adapted prediction algorithm can efficiently enhance the decorrelation and the quantization processes to eliminate remaining block effects originating from the block matching algorithms and to allow selective quantization levels according to the objects in motion.

2 3-D wavelet decomposition

The wavelet filter banks presented in this paper are based on a three-dimensional linear and separable structures. As mentioned earlier, the spatial and temporal dimensions are treated differently as a result of their different spectral contents. Temporal decomposition is performed first and generates the temporal subbands of different velocities. Objects in motions are blurred in the low frequency and clearly delimited in the intermediate or high frequency subbands. The spatial decomposition is thereafter applied.

2.1 The motion-compensated filtering

Under the assumption of signals slowly varying along their trajectories which make efficient any operation of prediction, interpolation and smoothing, the motion-compensated filtering operation along the motion trajectory at point \((x,y,t)\) may be written as

\[ s(x,t) = F_i(i(x,t)) \] (1)

where \(s(x,t)\), \(F_i[\cdot]\), and \(i(x,t)\) are respectively the filtering output, the motion-compensated filtering operation and the input image along the motion trajectory. In the simple cases where all motion trajectories are supposed to be parallel, the impulse response run along the line of equation \(x - vt = 0\) and takes the form \(h(v,t) = h_1(t)\delta(x-vt)\) with \(h_1(t)\) the impulse response along the motion trajectory. The filtered output is given by

\[ s(X,t) = \int h_1(\tau)\delta(m - v\tau)u(x - s, t - \tau)dsd\tau \] (2)

Instead of considering one single dominant motion, two other methods can also be developed as improvements of the basic global-motion-based technique, namely the block-based motion-compensated temporal filtering which carry out the adaptation according to the motion vector defined on fixed-size blocks estimated through a block matching compensation algorithm and an object-based motion-compensated temporal filtering which is adapted along the motion trajectories of the different objects contained in the scene and generating tubes in the 3-D scenes. Of course, the latter coding technique requires a spatio-temporal segmentation of the scene performed so as to discriminate the regions of significantly different motions. One single motion descriptor (vector or affine transformation) per region needs to be transmitted on the channel. Moreover, the segmentation can be either performed once in the encoder on the incoming video signal and encoded and transmitted to the decoder or performed simultaneously on both the encoding and decoding sides on the reconstructed image on the basis of the same motion information.

2.2 The adaptive spatial filtering

The spatial adaptive filters are implemented as a cascade structure of two-band lossless/para-unitary lattices of the polyphase matrix

\[ H_i(z) = c \prod_{k=1}^{N-1} A_{N-k}D \] (3)

where \(c = \prod_{k=0}^{N-1} \left[ \frac{-1}{1+z^{-1}} \right]^{1/2} \),

\[ A_k = \begin{bmatrix} 1 & 0 \\ -\alpha_k & 1 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix} \] (4)

The coefficients \(\alpha_k\) are adaptively estimated to minimize the following cost function

\[ J = \sigma_H^2 + \varphi \sum_{n=0}^{N-1} [h(n) - h(2N - 1 - n)]^2 \] (5)

where \(\varphi\) is a weighting factor and \(\sigma_H^2 = H^T R_{XX} H\). The matrix \(R_{XX}\) is the covariance matrix of the input signal \(x_i\), \(H_i\) is the impulse response of the high-pass filter \(h_i = (-1)^{n+1}h_0(N - 1 - n)\), \(N\) is the length of the filter impulse response. The constraints of L degrees of regularity means L conditions for the low-pass filter to be expressed on the form of

\[ (-1)^r \sum_n n^r (-1)^n h_0(n) = 0 \quad r = 0, ..., L-1 \] (6)
This leads to develop a Lagrangean function and to implement a gradient algorithm.
The adaptive FIR-filter bank implemented by cells of two-band para-unitary lattices presents the properties which follows

1. orthogonality supports better decorrelation efficiency.

2. perfect reconstruction as required for image coding. Like the first property, perfect reconstruction property is structurally guaranteed by the lattice formalism.

3. regularity has been proved to be useful for image coding applications. The conditions of regularity are imposed by L constraints mentioned in 6, they express that \( \frac{dH_0(e^{j\omega r})}{d\omega} \Big|_{\omega=\pi} = 0 \). Those conditions are not sufficient to lead to the maximum of regularity as a result of the existence of other potential zeros not located at \( z = -1 \).

4. decorrelation or energy compaction which leads to minimize the ratio \( \frac{\sigma_x^2}{\sigma_L^2 + \sigma_H^2} \), expressing the coding gain. L and H stand respectively for the input signal, the low and high-pass output signals. It is clear that minimizing \( \sigma_H \) is sufficient. High decorrelative performances concentrate the energy into a minimum number of transform coefficients and lead to improve the quantization performances in the sense that well-decorrelated subbands coefficients induce smaller subjective degradation or distortion at similar quantization step size.

5. quasi-linear phase response improves the quality of reconstructed quantized images, especially when quantization noise spreads around edges. Then, minimizing the asymmetry of the impulse response leads to minimizing the phase non-linearity. Though exploiting linear phase analysis-synthesis filters does not significantly improve the coding gain, the major effect of the phase linearity in synthesis belongs to the perceptual realm.

6. multiresolution offers an elegant issue to a generalized scalability;

7. adaptivity and optimality are provided in the analysis operation and in the buffer regulation. The three-dimensional filter bank is made adaptive to face the time-varying characteristics of within the image sequences. The adaptation is performed spatially according to the local and/or the global statistical content of the scene and temporally according to the motion analysis or estimation. Adaptivity is related to optimality since it provides the framework to develop optimum search algorithms.

The optimum is defined according to the criteria 4 and 5 in the cost function \( J \) to be minimized under the constraints mentioned in 3. In this framework, the highest compression is ultimately searched in conjunction with graceful degradations of the subjective quality. The lattice structure for para-unitary filters is presented in Figure 1 and the gradient algorithm configuration is sketched in Figure 2.

![Figure 1: Cascading lattice structures in adaptive filters.](image)

![Figure 2: Cascading lattice structures in adaptive filters.](image)

The adaptation covers a time-varying choice of both the lattice filter coefficients and the tree splitting structure. The filter length is adapted at each step to maximize the coding gain. The length of the impulse response may be adapted while progressing in the splitting. In whole coherence with the correlative signal structure, with the balance of spatial and frequency localization, the first or highest decompositions may use longer filter lengths than the last or lowest decompositions. Each splitting cell is optimized independently of each other and takes only into account the local input signal.

### 3 Motion estimation and compensation

The motion estimation and trajectory analysis is performed on an image-by-image on a depth at least equal to the length of the temporal filters. The motion analysis is be performed prior to coding, structured in forms
of a multiresolution description to develop a whole framework for a three-dimensional motion structure with trajectories and displacement fields.

4 The coding scheme

The efficiency of the coding scheme is based on two optimum algorithms. The analysis operation and the control of the computation of the quantization step size which is optimally managed to preserve graceful degradations with a most uniform subjective image quality.

The coding schemes used in the study implement either in-band prediction or per-format prediction schemes to mean that the motion-compensated prediction is performed on a band-per-band basis or that the subband information is cumulated up to resolution levels defined in a format hierarchy. The subband coefficients are quantized and entropy coded. The entropy coder uses the UVLC entropy coder. Finally, a buffer regulation implements a feedback from the buffer to the quantizers in order to control the evolution of the buffer content and fit to constant transmission rates. The feedback reaction chain is steered by a controller which manages optimum graceful degradations and maintains as far as possible uniform subjective image quality. The quantization step size is up-dated at the horizon of three-dimensional parallelepipeded fields (for example, 8 or 16 lines and 8 images in depth).

The computation of the quantization step size uses 3-D frequency weighting factors according to Robson’s and Girod’s results [8], [9]. Girod states that the HVS tracks the moving objects. Therefore, it would be impracticable to exploit these results on the psychovisual behavior without motion compensation. Robson’s results presents the spatio-temporal response of the HVS; they can be exploited to compute 3-D weighting factors.

5 Conclusions

This paper has outlined an on-going research work. The coding scheme studied in this paper has derived a general optimized linear operator adapted to the signal to be analyzed and encoded. It is based on a general three-dimensional wavelet transform enhanced by a motion analysis in terms of trajectories leading to adapting the temporal analysis filters and to compensating the temporal subbands. The coding adaptivity results from the use of lattice-structured filter banks according to a gradient algorithm which determines the filter coefficients to be applied in the spatial decomposition. The adaptive search aims at optimizing the filtering performances. The motion characteristics of the scene under process. Eventually, such a scheme generalizes the multiresolution to embody the temporal axis and to provide an efficient substratum for video format conversions and interpolations.

Acknowledgements

The following paper present a research work supported by the Commission of the European Communities DG XII under a fellowship ERB4001GT930967 of the Human Capital and Mobility.

References


Selective Coding Method Based on Motion Information

Jong-Bae Lee†, Seong-Dae Kim† and Sang-Zee Lee†

† Dept. of Electrical Engineering, KAIST 375-1 Kusong-Dong, Yusong-Gu, Taejon 305-701, KOREA
Tel/Fax: +82 [42] 869 5430/ 869 5410, E-Mail: jbh@evision.kaist.ac.kr
‡ Agency for Defence Development, P.O.BOX 35, Taejon 500-600, KOREA,
Tel/Fax: +82 [42] 821 3611 / 829 3400

Abstract. A selective coding scheme is proposed that has different bit allocation between background and target region. The suggested method initially estimates global motion parameters and segments target regions from a given image. Then we coded background and target region by assigning more bits to the target region and less bits to background in order to reconstruct the target region with high quality. Simulations show that the suggested algorithm has better result than the existing methods, especially in the circumstances where background changes and target region is small enough compared with that of background.

1. Introduction

The tactical scenes we want to transmit often consist of small target regions and the capacity of transmission channel is very limited. Also we must compensate camera movement by estimating global motion parameters. In such a circumstances, very low coding scheme is required and also target region must be reconstructed with high quality. But the existing methods do not consider a relative importance of information between background and target region and it is almost impossible to reconstruct target region with required quality on the limited transmission channel. So we suggest a new very low bit rate coding scheme which can compensate camera movement well and reconstruct target region successfully. The suggested algorithm first estimates global motion parameters in order to compensate camera motion. Next we calculated local motion vectors between motion-compensated frame and current frame and extracts target region from image sequences. Finally we encoded image sequences by allocating more bits to target region than background in order to reconstruct the target region with required quality by somewhat lowering the quality of background. Fig.1 shows the simplified block diagram of selective coding scheme. Motion analysis and segmentation block extracts local motion parameters, global motion parameters and some shape information. These parameters are sent to the transmission channel and also fed back to the motion compensation(MC) block. At the target region, motion-compensated error(MCE) signals are compressed by DCT and quantization(Q) blocks and sent to the transmission channel. But in case of background region, MCE signals are not transmitted except that the power of MCE signals are greater than the prescribed level determined by the background mode detection(BMD) block. In the background, quantization step size is controlled by regulator in order not to exceed the required channel capacity.

2. Global motion parameter estimation

It is not easy to estimate the motion vectors correctly by block matching algorithm[1], especially in the tactical image where background has some complexities and camera zoom happens. In this situation, global motion compensation must be required because it is possible to represent background motion easily by some global parameters. Some researches have been done on global motion parameter estimation[2,3,4]. Here we measure global motion parameters based on the global zoom/pan estimation method[2]. Because our algorithm must be applied well in the tactical scenes, let's assume that image satisfies the characteristics listed below.

Assumption 1. Background region has only global motion by camera movement and its velocity is different from that of target regions.

Assumption 2. Target regions are sufficiently small compared with background region.

It is necessary to define some values before estimating global motion parameters. Let \((X',Y')\) be the corresponding image space coordinates of the point before and after camera motion respectively. Also let's define \(f_z\) the zoom factor and denote panning parameters of \(x\)- and \(y\)-direction \(\Delta X, \Delta Y\) and rotation center \(g_x, g_y\) respectively. Then \((X', Y')\) is measured by Eq.(1).

\[
X' = f_z X + \Delta X, Y' = f_z Y + \Delta Y. \tag{1}
\]

Also we can have a dual equation

\[
X = f_z' X' + \Delta X', Y = f_z' Y' + \Delta Y'. \tag{2}
\]

by defining \(f_z' = \frac{1}{f_z^2}, \Delta X = -\frac{\Delta X}{f_z^2}\) and \(PY' = \frac{\Delta Y}{f_z^2}\).
Here we can put the estimated displacement vectors $ar{u}_i = X_i - X'_i = (f'_i - 1)X'_i + PX'$ and $ar{v}_i = Y_i - Y'_i = (f'_i - 1)Y'_i + PY'$. Also assume that we have an available set of local displacement vectors $u_i, v_i$ respectively at locations $(X'_i, Y'_i), i = 1, \cdots, P$ in the current frame. In order to estimate global motion parameters, let’s define $l_2$ norm of displacement error $E(f'_i, PX', PY')$

$$E(f'_i, PX', PY') = \sum_i ||(f'_i - 1)X'_i + PX' - u_i^2 + (f'_i - 1)Y'_i + PY' - v_i^2||$$

Then differentiating $E(f'_i, PX', PY')$ and setting the derivatives to zero, we can find out optimal global motion parameters like below.

$$f'_i = \frac{\sum_i (X_iX'_i + Y_iY'_i) - \frac{1}{3}(\sum_i U'_i)(\sum_i U'_i)}{\sum_i (X'_i^2 + Y'_i^2) - \frac{1}{3}(\sum_i X'_i)^2(\sum_i Y'_i)^2}$$

$$PX' = \frac{1}{P}(\sum_i X_i - f'_i \sum_i X'_i)$$

$$PY' = \frac{1}{P}(\sum_i Y_i - f'_i \sum_i Y'_i)$$

where $X_i = u_i + X'_i, Y_i = v_i + Y'_i, U_i = X_i + Y_i$ and $U'_i = X'_i + Y'_i$.

Because target region affects global motion parameters, we iterate estimation process two or three times excluding local motion vectors which do not match the previously estimated global motion vectors. Then we can find out the motion compensated frame $U(X, Y)$ using these global motion parameters.

3. Region segmentation

3.1 Background detection

Let’s first estimate local motion vectors $(u_i, v_i)$ for each 16 x 16 block using full search BMA algorithm between the motion-compensated frame $U(X, Y)$ and current frame $C(X, Y)$. Next, we measure the similarity between the velocity vectors of each block and define groups. Group is a collection of blocks which have similar velocities. In order to assign the group number to each block, we used the hierarchical clustering algorithm based on the agglomerative clustering method[5].

After counting the number of groups for the P blocks, let’s initially define the background region by putting the largest group to it and the other groups undefined region. In case of undefined region, we segmented more finely by using quadtree algorithm[6] because 16 x 16 block size is not small enough to represent the boundary between background and target region.

3.2 Region Segmentation by quadtree scheme

Stobach[7] suggested the quad-tree based coding scheme which uses the homogeneity of error signals.

In this paper, we segmented undefined regions by quad-tree scheme using homogeneity of motion vector information as well as that of error signals.

< quadtree algorithm >

Step 1 : Let the value of level be 0 for undefined region.

Step 2 : Divide the undefined region by four regions and estimate the velocities for each region. Increase the level of each divided region.

Step 3 : Find out the group number of divided regions from predefined groups. If we can not find out group number for each region, we assign a new group number to this region.

Step 4 : If the group numbers for four divided regions are same or the power of MCE signals for the undefined region is smaller than the sum of power of MCE signals for four divided regions, we finish quad-tree scheme by merging these regions. Otherwise, go to step 2 and iterate again.

Fig.2 shows the example of quad-tree scheme between motion-compensated frame and current frame. Here, region I, II and III are defined as background region and region IV and V target region respectively in the quad-tree algorithm.

4. Selective coding and data transmission

There exist MCE signals through all regions even if we compensate the image by global motion parameters especially in the tactical imagery which has complex background and happens large motion. Therefore in order to reconstruct the target region with high quality, we must minimize quantization error of motion-compensated frame difference signal by making quantization step size small. Let’s assume that frame difference signals for background and target region are first normalised by the factor $G_B$ and $G_T$ respectively. Also if we define $E_{max}$ as the maximum admissible error for target region and $\theta_T(B_i)$ the average quantization error of the $i$-th block which belongs to target region, then we must assure bit $B_i$ in order to satisfy Eq.(5).

$$\frac{1}{B_T} \sum_i G_B^2 \theta_T(B_i) \leq E_{max}$$

where $B_T$ is the number of blocks which belong to target region. Also Eq.(6) must be satisfied if $R$ indicates the overall system rate.

$$B_i \leq R - \frac{1}{B_T} \sum_i G_T^2 \theta_T(B_i) - P_a$$

where $B_i$ means the bit rate that can be used to transmit the background signal and $P_a$ several parameter information which includes global motion parameters, velocities of groups, position information.
Generally we do not transmit the MCE signals at the background region if the power of error signal is greater than the predefined threshold value.

5. Simulation results and concluding remarks

Fig. 3 shows a real image obtained from the image processing system designed by our laboratory. This image is taken as the previous image and current image is synthesized by moving the previous image globally with \( f_x = 1.03, P_X = 2.0 \) and \( P_Y = 2.0 \). But the big van is moving differently with velocity \( u_1 = 4 \) and \( v_1 = 4 \). The global motion parameters of proposed algorithm are \( f_x = 1.03, P_X = 1.92 \) and \( P_Y = 1.97 \), which shows that estimated values are very similar to true parameters. Fig.4 indicates the target region and fig.5 shows reconstructed image. Table 1 shows the PSNR of image and target regions and bit allocations for three methods and table 2 indicates components consisting of transmission bits. Here, our algorithm has higher PSNR gain and smaller transmission bits compared with other two algorithms because local motion vector information is very little by grouping similar motion vectors through background/target region clustering and quad-tree segmentation. Also in our algorithm target region has much higher PSNR values than background region because sufficient bits are allocated to target region in order to reconstruct the target region with required quality.

If we consider rotation factor as well as zoom factor, then this global parameter algorithm in section 1 must be modified by four or six parameters.

Reference


7 P. Strobackh, (1990), Tree-structured Scene Adaptive Coder, IEEE Tr. on Comm. 38.
Table 1 PSNR of image and target regions and bit allocation for three methods

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR of reconstr. image (dB)</th>
<th>PSNR of reconstr. target (dB)</th>
<th>required bits for transmission</th>
</tr>
</thead>
<tbody>
<tr>
<td>full search BMA</td>
<td>16.5</td>
<td>16.5</td>
<td>1544</td>
</tr>
<tr>
<td>full search BMA after motion comp.</td>
<td>27.9</td>
<td>27.9</td>
<td>1194</td>
</tr>
<tr>
<td>suggested algorithm</td>
<td>29.9</td>
<td>35.1</td>
<td>480</td>
</tr>
</tbody>
</table>

Table 2 Components of transmission bits for three methods

<table>
<thead>
<tr>
<th>Method</th>
<th>global para.</th>
<th>local motion vector</th>
<th>quad-tree infor.</th>
<th>error signal (target)</th>
</tr>
</thead>
<tbody>
<tr>
<td>full search BMA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>full search BMA after motion comp.</td>
<td>30</td>
<td>1074</td>
<td></td>
<td></td>
</tr>
<tr>
<td>suggested algorithm</td>
<td>30</td>
<td>36</td>
<td>108</td>
<td>306</td>
</tr>
</tbody>
</table>
Advanced Spatiotemporal Propagation Strategy for Multiresolution Motion Estimation

Lilla BÖRÖCZKY, Péter CSILLAG and Kálmán FAZEKAS*

KFKI Research Institute for Measurement and Computing Techniques
P. O. Box 49, H-1525 Budapest, Hungary, Fax: +36 [1] 169-5532
E-mail: boroczky@sun60.mszki.kfki.hu, o-csillag@nov.mht.bme.hu

*Technical University of Budapest, Department of Microwave Telecommunications,
E-mail: t-fazekas@nov.mht.bme.hu

Abstract. In this paper a predictive multiresolution motion estimation scheme is proposed. It is based on a spatiotemporal propagation strategy used for refinement of motion fields through the levels of an image pyramid. Consequently, the initial estimates of motion field vectors are predicted from the previously estimated motion vectors in the current image pyramid level, from the motion vectors estimated at the previous levels of the current image pyramid and from the motion vectors of the previous frames. Then, these initial estimates are updated within the actual level of the image pyramid by a chosen motion estimation algorithm. Application of the developed algorithm to motion-compensated wavelet video coding (estimating motion fields for all subbands) is also presented. Experimental results showed, that the proposed algorithm significantly outperforms the ones using only full resolution images and its performance in wavelet coding environment is also promising.

1. Introduction

Estimation of motion between successive frames of a time-varying image sequence has received a lot of attention in different fields of image processing including digital video and dynamic scene analysis. High temporal correlation of intensities along the motion trajectory can successfully be utilised for motion-compensated (MC) video coding, MC image sequence restoration, MC frame interpolation, etc. Applications of dynamic scene analysis include robot vision, traffic monitoring, biomedical imagery and remote sensing.

Frame-to-frame object motion in an image sequence can be represented by a 2-D time-varying vector field consisting of the local displacement vectors. This vector field is often referred as motion field, which is unknown in general and has to be estimated based on the given image sequence. Several motion estimation algorithms such as smoothness based, parametric, etc. were developed [1]. However, uniresolution motion estimation algorithms using two (or more) full resolution frames generally have the property of slow convergence rate even with complex methods. In the worst case, they are not at all suitable to estimate fast motion (large displacement vectors). To overcome this problem, multiresolution motion field estimation schemes have been proposed [2-4].

The multiresolution motion estimation framework consists of three basic components: construction of multiresolution image representation, motion estimation and propagation strategy for refinement of the motion fields estimated in coarser resolutions. Obviously, different
for all subbands is also presented. After a brief discussion of multiresolution motion estimation, the advanced spatiotemporal propagation strategy is outlined in Section 3. A predictive multiresolution motion estimation scheme for wavelet video coding is discussed in Section 4. Experimental results obtained for real-life video sequences are given in Section 5.

2. Multiresolution motion estimation

The motivation for using multiresolution motion estimation is twofold: to eliminate false matches using large scale structures, and to achieve computationally efficient algorithms. For example, large motion can cause aliasing of high spatial frequencies which results in obtaining false minima in the criterion function used for estimation of the motion field. Minimisation carried out in a multiresolution framework helps to eliminate this problem, because the non-convexity of the objective function is not stable with respect to the scale. On the other hand, false displacement vectors can be computed using low resolution image information, low computational complexity can be achieved. Then, higher resolution information is used to improve the accuracy of the motion field.

In the context of multiresolution motion estimation one of the most important problems is to design a sophisticated propagation strategy, which helps to prevent the propagation of false estimations into the final result. This is independent of the pyramid transform creating multiresolution image representation and the motion estimation algorithm applied in the image pyramid levels.

In the conventional and widely used coarse-to-fine propagation strategy, the estimated motion field obtained at a coarser level of the image pyramid is considered as the initial estimate for the next level, where it is refined. Its main disadvantage is the error propagation, i.e., the estimation error introduced to the displacement vector in the starting or coarser level propagates through the levels of the pyramid and it can also appear in the final motion field of full resolution. In order to judge the accuracy of the estimated motion field in the actual level and to control the estimation strategy a criterion function has been proposed [2]. Unfortunately, in this oscillatory approach some additional control parameters have to be tuned, which increase the computational complexity of the estimation algorithm.

3. Spatiotemporal propagation strategy

Design of sophisticated propagation strategy for multiresolution motion estimation can be considered as developing an advanced prediction strategy for the initial estimate of the motion field. Thus, it has to be updated within the actual level of the image pyramid by a chosen motion estimation algorithm. A good initial estimate will reduce the magnitude of the unknown update vector, consequently a simple estimation algorithm can be applied for the correction vector field and less iterations are required to get an accurate motion field.

For this prediction strategy it is worth to consider the spatiotemporal smoothness of the motion field [7, 8], which increases mainly the robustness of the motion estimation algorithm. Temporal smoothness of the motion field allows to track the object motion in time, thus only temporal changes in motion vectors have to be estimated. However, motion of an object – even along the motion trajectory – does not always vary smoothly. In order to overcome this problem a MC spatiotemporal prediction scheme was proposed in [7] for uniresolution pel-recursive motion estimation. This vector predictor benefits from the spatial smoothness of the motion field, as well as from the temporal one. Therefore, in our approach the initial estimates of motion field vectors can be predicted from the previously estimated motion vectors in the current level \( \mathbf{d}_c \), from the motion vectors estimated at the previous levels of the current image pyramid \( \mathbf{d}_{p1} \) and from the motion vectors of the previous frame \( \mathbf{d}_q \): \[
d_0 = f(d_c, d_{p1}, d_q) \tag{1}
\]
The most natural choice for \( f \) is a linear combination containing two parameters, \( P \) and \( Q \). The parameter \( P \) determines the ratio of vector \( d_{p1} \) in calculation of the motion vector \( d_c \) using only two consecutive frames from the image sequence:

\[
d_c = P d_{p1} + (1-P) \mathbf{d}_q \tag{2}
\]
The parameter \( Q \) is the ratio of the temporal component when the previous motion field pyramid is considered as well:

\[
d_0 = Q d_q + (1-Q) d_c \tag{3}
\]

Therefore, eq.(1) is:

\[
d_0 = Q d_q + (1-Q)(P d_{p1} + (1-P) \mathbf{d}_q) \tag{4}
\]

Parameters, \( P \) and \( Q \) of the linear equation can be determined experimentally for a given image sequence (See Figure 1 and 2), and their typical values are around 1/3.
motion fields for all frequency bands. The basic idea of this algorithm is that the initial estimate of a subband can use the result of another band. Consequently, for the low-pass filtered (LL) subbands a spatiotemporal prediction strategy given by eq.(1) is used. For other subbands (LH, HL, HH) a linear combination of motion vectors - obtained for the corresponding LL bands and for the spatial neighborhood of the actual pixel - is applied:

\[ d_0 = Rd_1 + (1-R)d_s \]  \hspace{1cm} (5) 

where \( d_0 \) is the initial estimate for the iteration, \( d_s \) is the spatial motion vector and \( d_1 \) is the motion vector calculated for the lowest frequency band in the same frame and the same level of the pyramid. The parameters \( R \) for different bands \( (R_{lh}, R_{hl}, R_{hh}) \) can be determined experimentally.

5. Experimental results

The proposed spatiotemporal propagation strategy was tested and verified on a real-life image sequence in an MC wavelet image coding environment.

The chosen motion estimation algorithm used in the above described multiresolution framework is a Wiener-based pel-recursive motion estimator [10]. It operates in a scanwise causal way through the images and is based on a prediction/update principle. At each pixel an iteration is performed and three strategies need to be specified: determination of the initial estimate to start the iterations, computation of the update vector at every iteration and control of the iteration process. The extension of this algorithm to multiresolution motion estimation is carried out by determination of the initial estimate using the spatiotemporal propagation strategy given by eq.(4) and eq.(5).

For the experiments the odd fields of the well-known image sequence “Trevor White” were selected. 3-level wavelet pyramid was created using filter coefficients given in Table 1 [5].

<table>
<thead>
<tr>
<th>n</th>
<th>h(n)</th>
<th>n</th>
<th>h(n)</th>
<th>n</th>
<th>h(n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.542</td>
<td>4</td>
<td>0.023</td>
<td>8</td>
<td>0.006</td>
</tr>
<tr>
<td>1</td>
<td>0.307</td>
<td>5</td>
<td>0.030</td>
<td>9</td>
<td>0.006</td>
</tr>
<tr>
<td>2</td>
<td>-0.035</td>
<td>6</td>
<td>-0.012</td>
<td>10</td>
<td>-0.003</td>
</tr>
<tr>
<td>3</td>
<td>-0.078</td>
<td>7</td>
<td>-0.013</td>
<td>11</td>
<td>-0.002</td>
</tr>
</tbody>
</table>

Table 1. Coefficients of the filter. \( h(-n) = h(n) \)

Three algorithms estimating motion fields for the low frequency subband were compared. The first algorithm uses only full resolution frames, the second is using...
multiresolution image representation with P=0.3 and the last one is utilizing the proposed spatiotemporal propagation strategy P=0.3, Q=0.35). The mean square errors obtained by the three algorithms are given in Figure 3.

![Graph showing MSE vs fields for Full res., Pyramid, and Temporal algorithms.](image)

Figure 3. Mean square error obtained by the three algorithms.

It is clearly visible, that the multiresolution approaches (pyramidal, temporal) significantly outperform the full resolution algorithm. At some frames, where the motion is smooth enough, the spatiotemporal algorithm results in a smaller MSE, then the pyramidal one.

6. Conclusions

An advanced propagation strategy based on the spatiotemporal smoothness of the motion field was proposed for multiresolution motion field estimation. The experimental results showed, that the proposed predictive multiresolution algorithm outperforms significantly the ones using only full resolution images. The application of the proposed scheme to wavelet video coding is also promising.

Our future work will be focused on combination of the proposed motion estimation scheme and an efficient segmentation algorithm which can improve its performance further. The application of the proposed algorithm to motion-compensated interpolation for low-bit rate wavelet coding is also considered to be a topic of our future research.

References


Estimation of Dense 2-D Motion Based on the Constancy of Intensity Gradient

Janusz KONRAD and Paolo TREVES

INRS-Télécommunications
Institut National de la Recherche Scientifique
Île-des-Sœurs, Verdun, Canada H3E 1H6
E-mail: konrad@inrs-telecom.uquebec.ca

Abstract. This paper describes a new approach to the estimation of dense 2-D motion from image sequences. Unlike in many other approaches that assume the constancy of image intensity along motion trajectories, we propose to use a higher order model that permits a variation of such intensity. We define the structural model that is based on the assumption of intensity gradient constancy along motion trajectories. This model has been proposed before [7], however in formulations that require exact satisfaction of the intensity gradient constraint. Due to inherent noise, aliasing, etc. present in images such solution necessitates additional post-processing, for example smoothing. We propose a different approach that is based on simultaneous estimation and smoothing. We formulate the problem using regularization where the assumptions of gradient constancy and of motion smoothness are combined into a single cost function. We minimize this function by an iterative method. We demonstrate estimation results for the original and for the “regularized” approach on natural image sequences.

1 Introduction

Dense 2-D motion, often called apparent motion or optical flow, is an essential characteristic of a time-varying image. It can be used for compression and/or processing [3] or for recovery of other characteristics of the original 3-D scene, such as structure or 3-D motion parameters. Therefore, it is important to develop reliable and robust algorithms for the estimation of the apparent motion.

In this paper we are interested in motion estimation that uses explicit modeling of motion trajectories in conceptual zty space, as opposed to methods that use implicit modeling such as in block matching or pixel-recursive approaches. With the above assumption, a motion estimation problem is usually formulated based on two fundamental models [3]: a structural model that establishes a relationship between motion parameters and image intensity, and a motion model that describes the properties of underlying motion. In this paper we are concerned primarily with the first model. For a detailed discussion of the motion model see [3].

2 Formulation

The structural model most often used in image correspondence over short time intervals is the assumption that image intensity along motion trajectories is constant [4]. Let \( u \) be the true underlying image and let \( x = [zy]^T \) and \( t \) be spatial and temporal coordinates of a point in \( u \). We assume that \( u \) is sampled sufficiently densely in \( t \) so that linear trajectories can be used. Let \( d(x,t) = [d^x(x,t)d^y(x,t)]^T \) be a displacement vector with horizontal and vertical components \( d^x \) and \( d^y \), respectively, associated with a pixel at \((x,t)\).

The constancy of image intensity can be expressed via the directional derivative as follows

\[
\frac{du(x, t)}{d\phi} = 0,
\]

where \( \phi \) is a variable defined along a motion trajectory.

When the image \( u \) is sampled, this directional derivative is usually approximated by a finite difference. In the case of estimating motion at time \( t \) from the preceding image at time \( t_- \) and from the following image at time \( t_+ \), equation (1) is approximated by a displaced pixel difference (DPD)

\[
u(x + (1 - \alpha)d(x, t), t_+ - u(x - \alpha d(x, t), t_-) = 0,
\]

where \( \alpha = (t - t_-)/(t_+ - t_-) \). This structural model may lead to various matching errors and thus to various formulations. For example, in block matching it is assumed to hold over a block of image pixels.

Since the assumption about intensity constancy is often violated due to scene illumination effects, a different characteristic that is more robust to such effects should be used. Tretiak and Pastor [7] have proposed to solve a set of two equations that are based on second-order image derivatives. Later, Bertel et al. [1] have restated this formulation as the invariance of the intensity gradient along motion trajectories:

\[
\frac{d\nabla u(x, t)}{d\phi} = 0,
\]

where \( \nabla = [\frac{\partial}{\partial x} \frac{\partial}{\partial y}]^T \) denotes a spatial gradient. In both papers the authors have proposed to solve this vector
equation exactly at each image point and to smooth the noisy result by post-filtering.

In this paper, we propose to extend the above approach. First, the approach used in [7],[1] carries out estimation and smoothing in two separate steps. We believe that the two should be executed simultaneously. Secondly, the formulation presented in both papers is limited to the calculation of motion vectors pivoted exactly at pixel locations. This is a severe constraint since in motion-compensated coding and processing motion parameters at arbitrary locations are often needed.

Assuming that motion trajectories are linear and that two images are used in the estimation, we propose to use the following displaced gradient difference (DGD) as the structural model

\[ \nabla u(x + (1 - \alpha)d(x, t), t_+) - \nabla u(x - \alpha d(x, t), t_-) = 0. \]  

(4)

Equation (4) is assumed to hold exactly for the true underlying (but unavailable) image u. Since typically the observed image sequence g is a noise-corrupted, filtered and sampled version of u [6], the DGD is no longer zero but instead the following equation holds

\[ s(d, x, t) = \nabla g(x + (1 - \alpha)d(x, t), t_+) - \nabla g(x - \alpha d(x, t), t_-), \]  

(5)

where \( s = [s_x, s_y]^T \) is a 2-D vector with noislike components (see [3] for a similar development in the DPD model).

To find an optimal solution based on the above structural model we propose to use regularization with respect to the global motion field \( d \)

\[ \min_d \sum_i (\|s(d_i, x_i, t)\|^2 + \lambda \sum_j \|d_i - d_j\|^2), \]  

(6)

where \( \eta_i \) is a spatial neighborhood of a vector at \( x_i \) and \( d_i = d(x_i, t) \) is used for compactness. The first term in (6) is an error due to the mismatch of gradients at the ends of displacement vector \( d_i \); whereas the second term is an error due to the departure of \( d_i \) from the assumed spatial smoothness of motion vectors [6]. Note that the above formulation allows \( g \) and \( d \) to be defined over different sampling structures.

3 Solution

The cost function (6) is non-quadratic since \( s(d_i, x_i, t) \) is a non-linear function of \( d_i \) (via the intensity \( g \)). Therefore, we assume that an initial estimate \( d_i = [d_{1i}^T d_{2i}^T]^T \) of \( d_i \) is known. Using the Taylor series expansion and retaining only the first-order terms we have

\[ s(d_i, x_i, t) \cong s(d_i, x_i, t) + \nabla_d^T s(d_i, x_i, t)|d_i = \tilde{d}_i (d_i - \tilde{d}_i), \]  

(7)

where the gradient is with respect to \( d_i \) but evaluated at \( \tilde{d}_i \)

\[ \nabla_d^T s|\tilde{d}_i \triangleq \nabla_d^T s(d_i, x_i, t)|d_i = \tilde{d}_i, \]  

\[ = \nabla_d^T (\nabla g(x_i + (1 - \alpha)d_i, t_+) - \nabla g(x_i - \alpha d_i, t_-))|\tilde{d}_i. \]

By the definition of the gradient and by the chain rule, the operator \( \nabla_d^T (\nabla) \) can be rewritten as follows:

\[ \nabla_d^T (\nabla) = \left[ \begin{array}{c} \frac{\partial \nabla}{\partial x} \frac{\partial g}{\partial x} \\ \frac{\partial \nabla}{\partial y} \frac{\partial g}{\partial y} \end{array} \right]_d \]

\[ = \left[ \begin{array}{cc} \frac{\partial^2 g}{\partial x^2} & \frac{\partial^2 g}{\partial x \partial y} \\ \frac{\partial^2 g}{\partial x \partial y} & \frac{\partial^2 g}{\partial y^2} \end{array} \right]\tilde{d}_i. \]

where \( \nabla (\nabla^T g) \) is the Hessian and \( e = [\partial x/\partial d_x \partial y/\partial d_y]^T \) both evaluated at \( d_i = \tilde{d}_i \).

For all locations in the preceding image \( x = x_i - \alpha d_x^T, y = y_i - \alpha d_y^T \) we obtain \( e = -\alpha [1 \ 1]^T \) and for the following image \( x = x_i + (1 - \alpha) d_x^T, y = y_i + (1 - \alpha) d_y^T \) we have \( e = (1 - \alpha) [1 \ 1]^T \). Thus, the final form of the gradient in (7) is

\[ \nabla_d^T s_i|\tilde{d}_i = \alpha \nabla (\nabla^T g(x_i - \alpha d_i, t_-))|\tilde{d}_i + (1 - \alpha) \nabla (\nabla^T g(x_i - \alpha d_i, t_+))|\tilde{d}_i. \]  

(8)

Note that \( \nabla_d^T s_i|\tilde{d}_i \) is calculated as an average of motion-compensated second-order derivatives at times \( t_- \) and \( t_+ \). This is in contrast with the derivative calculation used in [7],[1] where non-compensated derivatives are calculated regardless of motion amplitude. Our approach is expected to provide more precise derivatives due to “tracking” incorporated into derivative calculation.

With the approximation (7), the cost function in (6) has become quadratic in \( d \). Therefore, we can establish necessary conditions for optimality. Assuming the Euclidean norm and the first-order neighborhood \( \eta [6] \), we differentiate the cost function in (6) with \( s \) defined by (7) and (8). For each \( d_i \) we obtain the following set of equations:

\[ (\Phi + 4\lambda I)d_i = 4\lambda \tilde{d}_i + \Phi \tilde{d}_i - \left[ \begin{array}{cc} s_1 g_{xx} + s_2 g_{xy} \\ s_1 g_{xy} + s_2 g_{yy} \end{array} \right], \]  

(9)

where \( \tilde{d}_i \) is an average vector calculated over the neighborhood \( \eta_i \), \( i \) is the identity matrix and

\[ \Phi = \left[ \begin{array}{cc} g_{xx}^2 + g_{yy}^2 \\ g_{xy}^2 (g_{xx} + g_{yy}) \end{array} \right], \]

\[ \nabla_d^T s_i|\tilde{d}_i = \left[ \begin{array}{c} g_{xx} \ g_{xy} \\ g_{xy} \ g_{yy} \end{array} \right]. \]

Note that \( s_1 \) and \( s_2 \) are defined in (5).

Equations (6) applied at each \( x_i \) result in a large linear system. Note that due to a local nature of the neighborhood system used, the matrix of this system is very sparse. Therefore, we solve the system iteratively using the Gauss-Seidel deterministic relaxation. Once the
convergence is attained for a given $d_i$, we use the last estimate $d_i$ as the new $d_i$ and repeat the process all over again. This is equivalent to carrying out the Gauss-Newton minimization.

In (9) first- and second-order derivatives compensated for motion ($s_1, s_2, g_{xz}, g_{xy}, g_{yy}$) are calculated using separable bicubic interpolation [5].

4 Experimental results

To test the proposed algorithm and to compare it with other approaches we have computed motion fields of several natural image sequences. Fig. 1 shows one field of an interlaced image sequence for which motion was calculated from the immediately preceding and following fields. To regularize differentiation [2] each field has been first smoothed by a 2-D separable Gaussian filter with the variance $\sigma^2=2.5$.

![Image](image_url)

Figure 1: One field (100 pixels by 80 lines) from an interlaced sequence.

In Fig. 2 raw motion field1 obtained from solving the vector equation (3) [2] is shown. The noise present in the estimate is due to solving equation (3) exactly without any a priori knowledge about the estimate. To smooth out the estimate, filtering is often performed [1],[2]. In Fig. 3 the same displacement field filtered by a 2-D separable Gaussian filter is shown. Note the regularity of the solution at the cost of spatial localization.

In Fig. 4 motion field obtained using the proposed approach is shown. The solution is very regular but simultaneously it is quite precise in describing the motion of the hand, arm and head. For comparison we have also calculated a motion field using the standard structural model of intensity constancy along motion trajectories as given by equation (2). Instead of $||s(d_i, x_i, t)||^2$ in minimization (6), we have used a square of the left-hand side of equation (2). For details, please consult [5]. Motion estimate based on this model (called DPD) is shown in Fig. 5. Note the similarity between the DGD- and DPD-based estimates. Perhaps even greater similarity could have been obtained had we experimented more with various values of $\lambda$.

5 Summary and conclusions

We have proposed a new approach to the estimation of motion by combining via regularization a structural model based on intensity gradient constancy and a motion model. We have shown how to find a solution of the regularized formulation. We have demonstrated experimental results for one image sequence.

The proposed approach has been shown experimentally to give superior results in comparison with the “two-step” method described in [2]. This may be due to (a) simultaneous estimation and smoothing, (b) motion-compensated differentiation, and/or (c) consistent differentiation derived from a 2-D separable cubic interpolator. Currently we are working on the application of the proposed approach to motion-compensated processing under severe time-varying illumination. The initial results are very encouraging; the DGD-based motion estimates subjectively outperform the DPD-based estimates for synthetic illumination added to an image sequence. Objective comparison within motion-compensated temporal interpolation remains to be completed. These results will be published shortly.

References


Figure 2: Raw displacement field calculated by the method proposed in [2].

Figure 3: Displacement field from Fig. 2 smoothed by 2-D separable Gaussian filter with variance $\sigma^2=16$.

Figure 4: Displacement field obtained using the proposed approach (DGD) with $\lambda=1$.

Figure 5: Displacement field obtained using the standard structure model (DPD) with $\lambda=10$ [5].
Multi-level motion estimation for image sequence coding

Henri NICOLAS and Fabrice MOSCHENI

Signal Processing Laboratory
Swiss Federal Institute of Technology
1015-Lausanne, Switzerland
Phone: (+41 21) 693 26 01
Fax: (+41 21) 693 46 00
Email: nicolas@tssg3.epfl.ch

Abstract.
In this paper a new block-based motion estimation method for image sequence coding applications is proposed. This method combines an adaptive motion representation which uses a hierarchy of motion models, a multi-level motion representation and a quadtree segmentation. The adaptive motion representation permits to optimize the number of motion parameters which should be transmitted. The multi-level representation allows to exploit efficiently the redundancy existing in the motion vector field. Furthermore, the motion compensation error is quantized and transmitted only in the smaller blocks of the quadtree segmentation where the quality of the motion compensation is not satisfactory. Experimental results have shown a reduction of around 20% of the bit-rate using the proposed multi-level motion representation in comparison with the mono-level motion representation.

1. Introduction

In recent years, there has been a growing interest in image data compression, related to the emergence of high definition television (HDTV), digital television and videoconference. Classically, image sequence coding techniques reduce the temporal redundancy using motion estimation and compensation methods, the latter relying on a translational model of the motion. An example of such a technique is given in MPEG-II which is based on a block matching motion estimation algorithm [1]. The motion estimation can be made more accurate by the use of a variable size block matching technique. Referred to as quadtree segmentation, it starts from an initial block-based segmentation. Each region where the Displaced Frame Difference (DFD) energy is higher than a predefined threshold $S$ is split. The quadtree segmentation is mainly used for its low transmission cost and its computational simplicity [2].

These classical methods however have limitations. Since different kinds of motion exist in the images (i.e., translation, rotation, zoom, ...), the use of the translational model alone is not optimal. Representing the multiple motions existing in a scene with this unique model, it will be, on one hand, too simple in some regions leading to an oversegmentation and, on the other hand, too complex in others entailing the transmission of useless motion parameters. In [3], an adaptive motion model representation has been introduced. The most suitable motion model for each region is selected based on a cost criterion. The latter aims at optimizing the trade-off between the cost of the motion models and their efficiency to reduce the prediction error.

In this paper, the problem of exploiting the redundancies existing in the motion information is addressed. Due to the high similarities existing between the motions in successive frames, their corresponding motion fields are highly correlated. Under the assumption of smooth motions, any motion vector can be roughly approximated by the one estimated for the previous frame. Moreover, the quadtree splitting process induces redundancy at the level of the motion representation. It can be reduced by adopting a multi-level motion estimation where the motion estimation is performed through successive refinements of an initial motion vector.

This paper proposes a new block-based motion estimation algorithm which reduces the redundancies existing in the motion information. Based on the motion estimation introduced in [3], a new multi-level motion estimation combined with a tracking stage is developed. Furthermore, the motion parameters are validated according to the cost criterion at each level of the multi-level estimation process. The trade-off between the motion information and the DFD information is thus taken into account.
2. Multi-level motion representation and estimation

The aim of the proposed motion representation and estimation method is the exploitation of the redundancy existing in the motion vector fields. In the proposed method, the redundancy induced by the splitting process is exploiting by a multi-level motion representation and the similarity of successive motion vector fields is taken into account by a new tracking method.

The next paragraphs will describe these two new techniques and the used motion estimation technique.

Motion estimation The motion parameters are computed thanks to the adaptive motion model representation proposed in [3] and using an initial square block segmentation. A hierarchy of motion models is used where the simplest model of this hierarchy is the null model (no motion parameter) and the more complex one, i.e., the linear model, uses six motion parameters (two translational, one rotation, one zooming and two hyperbolic motion parameters [4, 5]). The parameters are estimated through the gradient method and a deterministic relaxation [6]. Taking into account the trade-off existing between motion information and DFD information, the motion parameters are validated by an entropy criterion which models the total bit-rate [7]. The regions where the DFD energy is higher than the predefined threshold $S$ are split according to a quadtree segmentation. The motion parameters and the model labels are entropy coded and transmitted with the quadtree segmentation to the decoder.

Multi-level motion representation Figure 1 illustrates the proposed multi-level representation. Different levels $i$ of segmentation are defined, each of these corresponding to a given size block segmentation. More specifically, the first level (i.e. $i = 1$) is the initial block-based segmentation. The other following levels (i.e. $i > 1$) are obtained by a quadtree segmentation of the block whose DFD energy is above the threshold $S$. The number of levels depends on the preset initial and the minimum region sizes.

If $\overrightarrow{d}_{i}(t)$ represents the corrective displacement estimated at the level $i$ and at time $t$, the total displacement $\overrightarrow{d}_{T}(t)$ at a level $i > 1$ is

$$\overrightarrow{d}_{T}(t) = \overrightarrow{d}_{i-1}(t) + \overrightarrow{d}_{i}(t). \quad (1)$$

The displacement vectors $\overrightarrow{d}(t)$ used to calculate the prediction error are calculated as follows:

$$\overrightarrow{d}(t) = \overrightarrow{d}_{init} + \sum_{i=1}^{i_{max}} \overrightarrow{d}_{i}(t). \quad (2)$$

Figure 1. Illustration of the multi-level estimation (three levels).

where $i_{max}$ corresponds to the maximum level number and $\overrightarrow{d}_{init}$ represents the initial value for the level one. This initial value is calculated using the previous estimated vector field as explained in the next paragraph.

The multi-level motion estimation can be viewed as successive approximation of the motion vector field, where, at each level, only correction motion parameters are estimated. The first consequence is that the error energy is reduced from one level to the next one. From a coding point of view, this representation has two advantages in comparison with the mono-level approach:

1. Simpler models are validated by the motion model adaptation criterion, thus reducing the number of transmitted motion parameters.

2. Thanks to the fact that the successive correction term are generally centered around zero, the transmission cost of the motion parameters decreases from one level to the next one.

Tracking of the motion vector field. The tracking is introduced in order to exploit the redundancy between two successive motion vector fields. The basic idea is to transmit only the difference between these two motion fields in order to reduce the transmission cost. In the proposed method, the displacements $\overrightarrow{d}_{T}(t-1)$ estimated at the first level of the previous motion compensated frame (time $t-1$) are taken as initial values for the first level of the current frame $t$. mathematically, the initial
displacement $d_{\text{init}}$ is:
\[
d_{\text{init}} = d_0^{(1)}(t - 1). \tag{3}
\]
The total motion vector field $\mathbf{d}_0(t)$ for the level $i = 1$ is then given by (see also Fig. 1):
\[
\mathbf{d}_0^{(1)}(t) = d_0^{(1)}(t - 1) + \mathbf{a}(t). \tag{4}
\]
Finally, the motion vector field used for the motion compensation is calculated using the corrective displacements estimated at each level and using the displacement estimate at the first level of the previous frame. The final displacement vector $\mathbf{d}(t)$ used for the motion compensation is calculated as follows:
\[
\mathbf{d}(t) = d_0^{(1)}(t - 1) + \sum_{i=1}^{i_{\text{max}}} \mathbf{a}^{(i)}(t). \tag{5}
\]

3. Experimental results

To evaluate the new motion estimation method proposed in this paper, simulations within a coding scheme are presented. Its performances are compared to the motion estimation described in [9]. The latter can be seen as a mono-level representation. Experiments have been carried out on the luminance component of the odd fields of the CCEIR 601 sequence "Table Tennis".

In the considered coding scheme, the motion information is computed through the evaluation of its 0-th order entropy. The splitting procedure permits furthermore to localize the DFD regions where the quality of the motion compensation is not acceptable. In order to reduce the transmission cost, the DFD is entropy coded and transmitted only in the smallest regions. With regard to quantization, the DFD, the translational parameters and the linear parameters are uniformly quantized with steps of respectively 30, $\frac{1}{4}$ and $\frac{1}{8^{\text{th}}}$ pixels. The first motion estimation uses the first original frame, while the decoded images are used afterwards. In all the simulations, the maximum and minimum block sizes are respectively 32x32 pixels and 4x4 pixels. The splitting threshold $S$ is equal to 150.

The experimental results compare the performances of the proposed motion representation (multi-level representation and tracking) with the mono-level approach (without tracking). Figures 1 and 2 show respectively the PSNR and the motion information as a function of the frame number and the Table 1 gives the average motion information and PSNR. It can be seen that a reduction of around 20% of the motion information is obtained by using the proposed motion representation (0.16 bit/pixel) in comparison with the mono-level approach (0.2 bit/pixel) for the same PSNR (around 30.3 dB, the visual quality remains approximatively the same in the two cases) of the decoded sequences.

<table>
<thead>
<tr>
<th></th>
<th>motion information (bits/pixel)</th>
<th>PSNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono-level representation</td>
<td>0.20</td>
<td>30.3</td>
</tr>
<tr>
<td>Multi-level representation and tracking</td>
<td>0.16</td>
<td>30.4</td>
</tr>
</tbody>
</table>

Table 1. Average bit-rate and PSNR for the mono and multi-level representations on the sequence "Table Tennis".

4. Conclusion

In this paper, a new multi-level motion estimation for image sequence coding application is presented. The method combines an adaptive motion representation which uses a hierarchy of motion models, a multi-level motion representation and a tracking of the motion vector field. The efficiency of the proposed method is demonstrated by the experiments which show a reduction of 20% of the motion information in comparison of the mono-level approach without tracking.

Our current investigation involves the development of a multi-level region-based segmentation to avoid the blocking artifacts generally induced by the block-based segmentation.

References

Figure 1: motion information as a function of frame (sequence "Table Tennis").

Figure 2: PSNR (dB) as a function of frame (sequence "Table Tennis").
A new motion field enhancement technique for video coding

Fabrice MOSCHENI, Frédéric DUFAX, Iole MOCCAGATTA and Markus SCHÜTZ

Signal Processing Laboratory
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
Tel: +41 21 693 36 32, Fax: +41 21 593 76 00
E-mail: moscheni@itssun2.epfl.ch

Abstract. Video coding schemes rely on motion estimation and compensation methods to exploit temporal redundancies. Due to their simplicity and low overhead information, block matching techniques are widely used. They however induce annoying artifacts due the hypothesis of uniform motion within each block. In this paper, a VQ-based segmentation entailing a refinement of the motion field is proposed. The segmentation is controlled by an entropy criterion leading to an optimal bit allocation between motion, segmentation and DFD information. Simulation results show an improvement of the PSNR as well as a removal of the block artifacts. A significant enhancement of the visual quality is reached.

1. Introduction

In the framework of video coding, block matching motion estimation and compensation techniques [1] are widely used due to their simplicity and ease of hardware implementation. These techniques combine the advantage of a low transmission cost of the motion vectors with a good suitability toward block-based coding schemes (DCT, etc.). However, the motion is assumed to be uniform inside each block. This hypothesis fails when a block is positioned on the boundary of objects moving in different directions. The resulting wrong motion vector assigned to the block does not exploit the temporal correlation while introducing block artifacts in the Displaced Frame Difference (DFD). The efficiency of the coding scheme is thus reduced.

In order to improve the motion estimation, different techniques have been studied. They include variable-size block matching [2, 3] and region-based motion estimation [4, 5]. In all these different segmentation approaches, accurate motion field and low cost overhead information are two contradictory requirements. The problem of the final bit rate being a trade-off between the amount of motion information, the segmentation information and the DFD has not yet been addressed. An optimization of the global bit allocation is therefore not reached.

In this paper, a new segmentation technique of the motion field based on Vector Quantization (VQ) is proposed. The purpose is to improve block-based motion estimation techniques as for example block matching motion estimation. The motion field is refined through segmentation. Blocks where the motion estimation fails are segmented and different motion vectors are assigned to every segmented region. The segmentation pattern is coded and transmitted by a VQ technique. The refinement strategy is moreover controlled by an entropy criterion. It aims at minimizing the global bit rate and reaching an optimal bit allocation between motion, segmentation and DFD information. This is achieved by taking into account the trade-off existing between the motion and segmentation information, on one side, and DFD information on the other side.

2. Segmentation and coding of the motion field by VQ

The proposed technique enhances the motion field estimated by a block-based motion estimation (e.g. block matching motion estimation). Every block is a candidate for segmentation in N regions, each of whom is assigned a motion vector coming from the set S of allowable motion vectors. Most naturally, the set S may be said to consist of the motion vectors obtained for the considered block itself and its neighborhood. Assuming spatial consistency of the motion field, the latter neighborhood is defined as the blocks connected to the block of interest.

The pattern defined by the association of a motion vector d to each pixel location f
\[
\tilde{d}(\bar{r}) \left\{ \begin{array}{ll}
\tilde{d}_1 & \text{if } \bar{r} \in \text{region 1} \\
\tilde{d}_2 & \text{if } \bar{r} \in \text{region 2} \\
\vdots & \\
\tilde{d}_N & \text{if } \bar{r} \in \text{region } N,
\end{array} \right.
\]

(1)

determines the \( N \) segmented regions, where
\[
\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_N \in \mathcal{D}.
\]

(2)
The pattern represents the segmentation information and has to be transmitted to the decoder. In the proposed method, a VQ technique is used to code efficiently this information [6]. The chosen pattern is transmitted under the form of its corresponding codewords \( \tilde{c} \).

Different values \( N \) ranging from 2 to \( N_{\text{max}} \) are considered and for each choice of \( N \), a different codebook \( C_N \) is defined. Taking \( t \) as the time index, \( u(\bar{r}, t) \) the image intensity, \( \mathcal{W} \) the matching window and \( || \cdot || \) a distance measure, the choice of the best segmentation pattern for a given \( N \) is achieved by minimizing

\[
\min \sum_{\bar{r} \in \mathcal{W}} || u(\bar{r}, t + 1) - u(\bar{r} - \tilde{d}(\tilde{c}, \bar{r}), t) ||
\]

(3)
through an exhaustive search of all the codewords \( \tilde{c} \) of the codebooks \( C_N \) and all the allowed combinations of motion vectors. Regarding the codebook design rule, a synthetic codebook has been preferred. The difficulties of the training procedure are thus avoided.

3. Analytical entropy criterion on the segmentation strategy

To optimize the bit allocation between motion, segmentation and DFD information, the criterion proposed in [7] has been applied. It allows to control the segmentation procedure described in Sec. 2, so as to minimize the global bit rate. For each block, the amount of information without segmentation \( R^{(1)} \) and with segmentation in \( N \) regions \( R^{(N)} \), \( N = 2, 3, \ldots, N_{\text{max}} \) are compared. Hence the segmentation control is written

\[
N \in \{2, 3, \ldots, N_{\text{max}}\} \Rightarrow \min \frac{R^{(N)}}{R^{(1)}} < 1
\]

(4)
where
\[
R^{(N)} = R^{(N)}_{\text{seg}} + R^{(N)}_{\text{DFD}}.
\]

(5)

with \( R^{(N)}_{\text{seg}} \) and \( R^{(N)}_{\text{DFD}} \) the bit rate required to transmit respectively the motion information, the segmentation information and the prediction error for the block. Unlike \( R^{(N)}_{\text{seg}} \) and \( R^{(N)}_{\text{DFD}} \) which are easily estimated, the case of \( R^{(N)}_{\text{DFD}} \) is more difficult. Assuming an entropy coder, \( R^{(N)}_{\text{DFD}} \) is given by

\[
R^{(N)}_{\text{DFD}} = n \cdot H^{(N)}_{\text{DFD}},
\]

(6)
with \( n \) the number of pixels within the block and \( H^{(N)}_{\text{DFD}} \) their entropy after segmentation in \( N \) regions. Relying on a memoryless Laplacian model of the DFD, an analytical formula for the DFD 0-th order entropy is derived [7] and introduced in Eq. (6). The 0-th order hypothesis is justified by the marginal correlation remaining in the DFD [8, 9].

4. Implementation and application to video coding

To validate the proposed motion field enhancement technique, simulation results obtained within a video coding scheme are presented and compared to the classical full-search block matching algorithm. Experiments have been performed on the luminance component of the test sequences “Flower Garden” and “Table Tennis” in common intermediate format (CIF, frame size of 288 x 352 pixel).

In the considered coding scheme, the first frame of a group of pictures is coded in intraframe mode using a wavelet transform [10], while the following frames are predicted by motion compensation. Moreover the DFDs are uniformly quantized and entropy coded using an adaptive arithmetic coder [11]. This scheme is consistent with the model of the DFD adopted in the segmentation criterion for it supposes no correlation between pixel of the DFD. Furthermore, the motion vectors are coded by using the adaptive arithmetic coder.

The initial block-based motion vectors are estimated by full-search block matching, with a block size of 16 x 16 pixel, a maximum displacement of \pm 15 pixel, and one pixel accuracy. As far as the proposed technique is concerned, the experiments have been carried out for two cases. In a first stage, only one codebook \( C_2 \) has been considered where the block segmentation has been limited to two regions of connected pixels. The codebook includes only the patterns whose boundaries are straight lines. The resulting number of codewords is thus 1275. In a second stage, two codebooks \( C_2 \) and \( C_3 \) are taken into account. The first being equal to the previously defined codebook, the second one contains segmentation patterns of 3 regions whose boundaries are straight lines. For the latter, the total number of codewords si
8171. For all codebooks, the 4-connectivity choice has been made to define the neighborhood of the segmented block. These segmentation constraints are necessary to balance computational complexity, amount of side information and quality enhancement of the motion field.

Figure 1 compares the performances obtained by the proposed technique in the two above stated cases and the classical block matching algorithm. The performances are evaluated in terms of bit rate versus Peak-Signal-Noise-Ratio (PSNR). Although the PSNR is a poor measure of the visual quality, it is commonly used due to the lack of perceptually reliable visual quality measures.

The gain over the classical full-search technique varies from 1 to 1.5 dB. This clearly demonstrates the improvement obtained thanks to the VQ-based segmentation of the motion field under the entropy criterion. Furthermore, the gain due to the use of the 3-regions codebook is negligible. A segmentation limited to 2 regions is thus quasi-optimal in terms of improvement of the PSNR. This results from the small size of the block compared to the size of the moving objects. The number of objects present in a block badly motion compensated is most of the time equal to 2.

Furthermore, the visual quality is strikingly enhanced by the proposed technique. The latter is indeed very efficient in improving motion compensation along moving edges. Figure 2 shows results obtained with the sequence “Table Tennis”. It compares close-ups of the original image with the motion compensated image in both cases of full-search block matching alone and after applying the proposed technique. Only the 2-regions codebook is used in the latter. The block artifacts present in the full-search block matching case disappear when the proposed technique of motion field enhancement is employed. The motion compensated image is thus visually significantly better.

5. Conclusion

A new technique for enhancing the block-based motion field in video coding schemes is proposed. In particular, it allows to improve block matching techniques. Based on a segmentation of motion field, it relies on VQ to transmit the segmentation information. The segmentation strategy is controlled by an entropy criterion and a minimization of the global bit rate is reached. Simulations have shown a gain of 1.5 dB in terms of PSNR when applying the proposed method to the full search block matching motion estimation. Moreover, a characteristic of the proposed technique is to considerably improve the motion compensation along moving edges. It entails that the visual quality is greatly enhanced. Finally, it is shown that the possibility of segmentation in 3 regions brings a negligible improvement compared to the case when only a 2 regions segmentation is allowed.

Figure 1. Comparison of the full-search block matching and the proposed motion estimation technique with respectively 1 and 2 codebooks, a) “Flower Garden” and b) “Table Tennis”.

References


Figure 2. Comparison between a) the original image, b) the full-search block matching motion compensated image and c) the motion compensated image obtained with the 2-regions codebook.
A New Block-Matching Algorithm for Estimating Multiple Image Motion Vectors

Takahiro SAITO and Takashi KOMATSU

Dept. of Elect. Eng., Kanagawa Univ., 3-27-1 Rokkakubashi, Yokohama, 221, Japan
Tel/Fax: +81[45]481 5661/491 791, Email: kurikuri@cc.kanagawa-u.ac.jp

Abstract. To estimate local motion from moving images, many algorithms have been proposed. Although the concept of the existing block-matching algorithm is very simple, it works well even under real noisy conditions, and hence it is often used in various practical application fields such as video processing. However, the existing block-matching method does not work well if a given analysis window covers multiple regions moving in different directions. To solve this problem, in this paper we extend the basic concept of the existing block-matching algorithm, and thus propose a new block-matching algorithm for the analysis of multiple distinct motions. Through computer simulations, we demonstrate the potentiality of the new block-matching algorithm.

1. Introduction

The motion estimation at a point in an image must be based on pattern information in a neighborhood of that point. The neighborhood is referred to as the motion analysis window. To estimate local motion from moving images and/or to estimate local disparity from stereo images, many algorithms have been proposed[1]. These algorithms can be classified into two approaches: matching methods[2] and gradient-based methods[3,4]. The gradient-based method works well if the magnitude of motion is not too great and the abrupt discontinuous changes of image intensity do not appear within a given analysis window. However, the gradient-based method is particularly sensitive to noise. On the other hand, although the concept of the existing block-matching algorithm is very simple, it works well even under real noisy conditions, and hence it is often used in various practical application fields such as video processing. However, the existing block-matching method does not work well if a given analysis window, viz. an analysis image block, contains multiple regions moving in different directions, and/or if block boundaries and object boundaries do not coincide. In these cases, not only the existing block-matching method but also the existing gradient-based method does not work satisfactorily.

To solve the problem, we extend the basic concept of the existing block-matching method, and propose a new block-matching algorithm for the analysis of multiple motions. The new extended algorithm is based on the assumption that there may be multiple distinct regions undergoing coherent motion within a given analysis image block. Recently one time with another, the study group of Bergen and others[5] and the study group of Shizawa and others[6] have independently proposed new gradient-based algorithms for the analysis of two motions by processing three consecutive image frames. However, their algorithms are based on the brightness change constraint equation about spatiotemporal second derivatives of image intensity, and hence their algorithms seem to be sensitive to noise like the existing gradient-based method. On the other hand, our new algorithm can estimate more than or equal to two motions by processing two or three consecutive image frames, and moreover our algorithm seems to be robust even under real noisy conditions like the existing block-matching method.

2. Local Motion Configurations and Motion Estimation

The size of the analysis image block is a critical factor in local motion estimation. The appropriate size depends on such factors as the size and velocity of objects in the scene, and hence it is extremely difficult to determine the optimal size of the analysis image block. Therefore, there may be a number of motion configurations occurring within an analysis image block when its size is arbitrary determined. The local motion configurations, basically, can be classified into three cases:

Case I: The block contains a single pattern undergoing coherent motion, and hence the block has only one motion.

Case II: The block consists of multiple distinct regions undergoing coherent motion, but each pixel within the block has only one motion.

Case III: The block contains multiple moving image patterns that appear superimposed, and hence each pixel within the block has multiple motions. Examples include not only actual moving transparent objects, but also moving shadows, spotlights, reflections, etc.

The existing block-matching algorithm can deal with only the first case (Case I). In this paper, we present an extended algorithm that can deal with the second case (Case II). Some time or other, we will report an extended algorithm that can handle the third case (Case III).

3. A New Block-Matching Algorithm for Estimating Multiple Motion Vectors

Figure 1 illustrates the concept of our new block-matching algorithm for the analysis of multiple motions in the second motion configuration case (Case II) in contrast with that of
the existing block-matching algorithm. For the analysis of multiple motion vectors, firstly the new algorithm segments a given analysis image block into multiple distinct regions undergoing coherent motion, simultaneously applies its matching mechanism to each region separately, and then provides its multiple matching results as multiple motion vector estimates; the new algorithm finally provides a single motion estimate for each region. A segmentation process is usually based on image intensity, but the segmentation process introduced here is based on the assumption that each region undergoes coherent motion. We form the segmentation process satisfying the following two requirements:

(i) For given image segmentation, the algorithm should provide multiple motion vector estimates giving the minimum value of a certain cost function that is defined on the basis of a displaced frame difference compensated with multiple motion vector estimates.

(ii) For given multiple motion vectors, the algorithm should provide proper image segmentation giving the minimum value of the cost function identical to that mentioned in the requirement (i).

We organize the new algorithm as follows:

[Algorithm]

(0) Set the number of regions $p$ in the segmentation. Define an analysis image block with $M \times M$ pixels in the present image frame $X$, and define a search area in the next image frame $Y$ as a $L \times L$ block centered at the location corresponding to the center of each analysis image block.

(1) Compute displaced squared frame differences between a given analysis image block defined in the present image frame $X$ and any block with $M \times M$ pixels whose center is located in the search area defined in the next image frame $Y$, and store the displaced squared frame difference for each pixel within the analysis image block, in matrix form.

(2) Choose $p$ blocks from among all the $M \times M$ blocks whose centers are located in the search area. For each pixel within the analysis image block, choose a proper block giving the minimum displaced squared frame difference from among the $p$ blocks, and compute the sum of the minimum displaced squared frame differences within the analysis image block. We refer to the sum as S-DFD (Segmental-DFD) for short.

(3) Apply the procedure of the step

\[ M_{in} = \sum_{i,j \in BL} (X_{ij} - Y_{i+m_j,j+n_j})^2 \]

(1)

Figure 1. The concept of the new block-matching algorithm and that of the existing block-matching algorithm.

Figure 2. The test moving image sequence used in the simulation and the four analysis image blocks of A, B, C, and D.

Figure 3. The measured relationship between the minimum S-DFD value and the number of regions $p$ in the segmentation.
One of the rest important questions is how to determine the proper number of regions \( p \). Through computer simulations conducted on real moving image sequences, we derive a good way of determining the proper number of regions \( p \). In all the experimental simulations presented here, the size of an analysis image block and that of a search area are fixed at 25 x 25 and 13 x 13 respectively. Figure 3 shows a measured relationship between the minimum S-DFD value and the number of regions \( p \) for the analysis image blocks of A, B, C, and D defined on the moving image sequence as shown in figure 2, where the doll located near the center of the frame is moving horizontally in the right direction with the velocity of approximately 4 pixels per frame. For the analysis image blocks of A and B, which are defined in a single image region undergoing coherent motion, the increase of the number of regions \( p \) from 1 to 2 does not much decrease the minimum S-DFD value. On the other hand, for the analysis image blocks of C and D, which contain two distinct regions undergoing coherent motion, the increase of the number of regions \( p \) from 1 to 2 decreases the minimum S-DFD value to less than 10% of the preceding minimum S-DFD value at \( p=1 \), but the increase of \( p \) from 2 to 3 does not much decrease the minimum S-DFD value. The decreasing rate of the minimum S-DFD value depends on both image contents and noise properties, but here we experimentally derive a good way of determining the proper value of \( p \). The good way is as follows: we start with the initial value of 1 for \( p \), and then we increase \( p \) by one if and only if the increase of \( p \) by one decreases the minimum S-DFD value to less than 20% of the preceding minimum S-DFD value before the increase of \( p \).

4. Experimental Simulations

Figure 4 compares performance of the new block-matching algorithm with that of the existing block-matching algorithm for the analysis image block C defined in figure 2. Figure 4(b) shows performance of the existing algorithm. In figure 4(b), for various hypothetical motion vectors \((m, n)\), the sum of the displaced square frame differences, called DFD for short, is calculated within the analysis image block C and is shown, but as for the DFD value for ease of observation the logarithmic value of the inverse of DFD is shown. Hence, the graph shows its peak at the location where the DFD value is minimized. In figure 4(b), the graph shows its gently-sloping ridge along the lines of \( m=0 \) and \( m=1 \), and shows its vague peak near the location of \((0, 3)\). In this case, the existing algorithm provides \((0, 3)\) as a motion vector estimate, and hence the motion vector is not accurately estimated. On the other hand, figure 4(a) shows performance of the new algorithm. Figure 4(a) shows the S-DFD value for various hypothetical horizontal components \( n_1, n_2 \) of two motion vectors \((0, n_1), (0, n_2)\). The graph is symmetric about the line of \( n_1 = n_2 \), because the order of \( n_1 \) and \( n_2 \) is completely meaningless in the new algorithm. In figure 4(a), the graph clearly shows its sharp peak near the location of \((n_1, n_2) = (0, 4)\), which means that the new algorithm works well and provides two accurate motion vector estimates of \((0, 0)\) and \((0, 4)\) stably.

Figure 5 shows the locations of analysis image blocks where the new block-matching algorithm fixes the number of regions \( p \) at 2, as an square image block painted two colors of white and gray, and the regions colored white or gray denote two segmented image regions within each analysis image block. Figure 6 compares motion vector estimates provided by the whole process of the new algorithm for the real moving image of figure 2 with those provided by the existing algorithm. In figure 6, each motion vector estimate is displayed at the location of the center pixel of each analysis image block. As shown in figure 6(b) and figure 6(d), the existing algorithm does not provide correct motion vector estimates.
along the contour of the moving doll. On the other hand, as shown in figure 6(a) and figure 6(c), the new algorithm provides correct motion vector estimates along the left-hand contour of the moving doll, but does not provide correct motion vector estimates along the right-hand contour of the moving doll. A given analysis image block defined in the area along the right-hand contour of the moving doll contains part of the still background covered by the doll moving from the left to the right, and the covered background does not appear in the next image frame. Hence, correct motion vector estimates are not provided here. Figure 7 illustrates this problem. In figure 7, for the analysis image blocks of A and B the new algorithm provides correct motion vector estimates, but for analysis image blocks such as the block C the new algorithm cannot provide correct motion vector estimates.

To solve this problem, we should introduce the concept of a three-frame algorithm which estimates multiple motion vectors by applying the new extended block-matching algorithm to three consecutive image frames. The three-frame algorithm, applying the new extended block-matching algorithm not only to a combination of the present image frame and the next image frame but also to a combination of the present image frame and the previous image frame, produces two separate estimates of multiple motion vectors, and finally chooses the better estimate giving the smaller minimum S-DFD value between the two estimates.

5. Conclusions
We extend the basic concept of the block-matching motion estimation method, and propose a new algorithm for estimating multiple motion vectors by processing two or three consecutive image frames. Computer simulations conducted on real moving image sequences demonstrate that the new algorithm presented here provides a correct estimate of multiple motion vectors even in the area along the boundaries between multiple differently moving regions in a scene, and that the new algorithm is a very potential practical means as a motion estimation technique.

Further investigation are as follows: development of an extended block-matching algorithm that can analyze multiple motions in the third motion configuration case (Case III), reduction of the computation time by introducing various fast search techniques such as the tree-search method into the new extended block-matching algorithm, construction of a larger motion analysis system that can detect what kind of local motion configurations occurs within an analysis image block and can assemble local motion estimates into an overall interpretation of scene motion, application to various real problems of moving image processing, etc.

References
Multiple Motion Estimation by Robust Parameter Estimation over Multiple Frames

Serge AYER and Philippe SCHROETER

Signal Processing Laboratory, Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland, Tel: +41 21 695 26 01, E-mail: ayer@de.its.epfl.ch

Abstract. This paper presents a method for time-varying motion analysis which uses both a large spatial region and a large temporal support. The motion parameters are modeled spatially and temporally, by defining them as linear combinations of some orthogonal time functions. The main advantage of the proposed algorithm is its ability to constrain the computation of the flow field spatially and temporally at the same time. The number of time basis varies according to the complexity of the movement and is determined automatically by means of a multiresolution approach and of statistical tests. The motion estimation problem is thus formulated as one of time-varying parameter estimation over multiple frames, and the estimation of these parameters is accomplished through a robust regression technique. As these robust regression methods are resistant to a high percentage of outliers in the data, they allow us to overcome the problem of multiple moving objects inside the region of analysis and to recover the motion parameters of the different moving objects.

1. Introduction

Motion estimation has been given a very large attention in the research community. Of particular interest in the different methods proposed for solving this problem is the size of both spatial and temporal domains. The smallest possible spatial and temporal regions are used by the optical flow approach. Such a flow field assigns to each pixel of one image a translational vector containing local motion information computed between image pairs. However, as small spatial and temporal regions do not always carry sufficient motion information, the optical flow computation can be very inaccurate, and the need for increasing both the spatial and temporal supports exists. An attempt to solve this problem is the use of parametric motion estimators, which describe the motion over a larger spatial region in terms of a parametric model. In this case, a model is used in order to constrain the flow field computation. However, when the region of analysis contains multiple moving objects, one is compelled to use an estimation method that can recover both the model parameters and the motion discontinuities at the same time. To achieve this goal, several ideas have been proposed, among which we can discriminate those based on line process (discontinuities detection) [1] and outlier detection [2]. In this paper, we propose a method for time-varying motion analysis which uses both a large spatial region and a large temporal support and which uses robust regression methods to overcome the problem of multiple moving objects inside the region of analysis.

In Section 2.1, we present an overview of the method. More details about the time-varying motion estimation are given in Section 2.2. The region-based goodness of fit process, which allows multiple estimations, is described in Section 2.3. Finally, we show how orthogonal time functions may be used for obtaining an optimal representation of the motion parameters (Section 3.).

2. Model-Based Time-Varying Motion Estimation

2.1 Overview of the Method

The motion estimation problem is formulated as one of time-varying parameter estimation over multiple frames, and the estimation of these parameters is accomplished through a robust regression technique (i.e. LMedS or LTS [3]). As these robust regression methods are resistant to a high percentage of outliers in the data, they allow us to overcome the problem of multiple moving objects inside the region of analysis. In Fig. 1, we give an overview of the different basic operations which are involved in the implementation of the algorithm. The estimation is performed in a hierarchical and iterative way. In our implementation, we use the Laplacian image pyramid as multisresolution representation. The estimation process begins at the coarsest resolution level, using zero initial estimates. Several estimation steps are performed at this resolution level, and the parameters are then projected down to the next pyramid level, where the same operations take place (starting now from the projected estimates). The whole process repeats until the resolution level 0 is reached. The motion
estimation algorithm is first applied to the whole image and then successively to regions in which the estimated parameters are not valid (see Section 2.3).

2.2 From Parametric to Time-Varying Parametric Estimation

If we consider only two frames in the sequence and assume intensity constancy (i.e., the brightness of a small surface patch is not changed by motion), the problem of motion model fitting can be posed as the minimization over the region of analysis of an objective function of

\[
I(x,t) - I(x - u(x,p),t-1)
\]

where \( p \) denotes the model parameters, \( u(x,p) \) the flow field in that region, and \( I(x - u(x,p),t-1) \) the image at \( t-1 \) warped towards \( t \) \([4, 5, 6]\).

With this formulation, the use of several pairs of frames for motion estimation would lead to several distinct pairwise estimations. In this case, the parameter evolutions could be described as a time series. However, for robustness and efficiency purposes, the integration of more measurements into a single estimation process is desirable. Since the motion to be modeled is time-varying, we think that the model parameters to be estimated should also be time-varying. In our method, each coefficient in the model is allowed to change in time by defining it as a linear combination of some known time functions. Thus, each motion parameter is itself specified by a set of parameters, namely the coefficients in the linear combination. This approach has been inspired by several methods of time-varying parametric modeling in speech processing \([7]\).

By limiting our attention to such a time-varying model, we are clearly constraining the possible types of time variations that can be allowed. However, constraints on the nature of the time variations are essential in order to limit the degrees of freedom in the time-varying parameters, so that incoherent and noisy estimations can be avoided. A judicious choice of the basis functions \( f_i(t) \) can provide an adequate representation of the time variations of the motion parameters. The use of orthogonal time functions for these purposes will be explained in Section 3.

With a model of this form, the coefficients in the linear combination have to be estimated from the image sequence, and the problem of motion estimation is posed as the minimization of an objective function of

\[
I(x,t_j) - I(x - u(x,p(t_{j-1})),t_{j-1})
\]

over the region of analysis and over \( M+1 \) frames \((j = 1, \ldots, M)\). In this formulation, \( p(t_j) \) denotes the motion parameter vector at time \( t_j \), whose components are modeled as linear combinations of some known time functions \( f_i(t) \):

\[
p(t) = (p_0(t), \ldots, p_{P-1}(t))
\]

with

\[
p(t) = \sum_{k=0}^{F-1} a_{i+k} f_i(t)
\]

where \( P \) is the number of motion parameters (e.g., translational model \( P = 2 \), affine model \( P = 6 \)) and \( F \) the number of time functions, \( 1 \leq F \leq M \).

The problem of estimating the parameter vector \( a \) from equation (2) leads to a system of non-linear algebraic equations. This system has no closed form solution and, consequently, one is compelled to find an estimate of the parameter vector \( a \) numerically. Suppose that an initial estimate \( \hat{a} \) is available (in our case, it will be the parameter vector \( 0 \) at the coarsest resolution level and the projected estimates at the other levels). The problem may be linearized using \( \hat{a} \) and the first order Taylor expansion of equation (2):

\[
\Delta I(u(x,p))|_{a=\hat{a}} = (\nabla I)^T \frac{\partial u}{\partial p} \frac{\Delta a}{\Delta \hat{a}} |_{a=\hat{a}}
\]

where

\[
\Delta I(u(x,p)) = I(x,t_j) - I(x - u(x,p(t_{j-1})),t_{j-1})
\]

and

\[
\nabla I = \nabla I(x - u(x,p),t_{j-1}).
\]

This shows that after linearization of the model, the problem of time-varying model-based motion estimation can be reformulated to one of parameter estimation. The same robust linear regression methods as in \([5, 4]\) may be used to estimate the parameters \( a \) from the resulting system of equations.
2.3 Region-based goodness of fit measures

The motion estimation algorithm is first applied to an entire region of the image (here the whole image). We then determine the statically segmented regions \([8]\) in which the estimated parameters are not valid, and apply further motion estimations to these regions. Thus, by successive applications of the robust motion estimation algorithm, the number of regions for which the previously estimated motion parameters could not explain the motion well is reduced to zero. The result of the multiple motion estimation algorithm is a set of time-varying motion parameter vectors \(\mathbf{p}_j(t)\). These vectors may be used along with gray level information and time integration in order to obtain an accurate segmentation of the moving objects (see results in \([9, 10]\)).

Experimental results indicate that the proposed scheme is robust to the presence of different moving objects and is also general enough to deal with scenes with moving or static cameras, with objects close to or far from the camera, and stationary or non-stationary environment.

3. Sequential Fitting with Orthogonal Functions

Equation (5) shows that, after linearization of the model, the problem of time-varying model-based motion estimation can be reformulated to one of parameter estimation. However, the degree of freedom of the parameter time variation i.e. the number \(F\) of time functions, \(1 \leq F \leq M\), is not known a priori and needs to be determined. We want here to study the use of orthogonal time functions to fit the motion model. Without lack of generality, we concentrate on the case of translational model.

From equation (5), we see that we have to solve a set of linear equations of the form \(\Delta f(u(x, p))) = \mathbf{a} - \mathbf{a}\), where \(\mathbf{y}\) is the vector \(\Delta f(u(x, p)))\) and \(\mathbf{b}\) is the parameter vector \((\mathbf{a} - \mathbf{a})\) to be estimated.

In general, we cannot speak of finding the least squares estimates and the regression sum of squares due to a given time function (i.e. due to \(b_{ik+p_k}, 0 \leq i \leq P-1\)) without accounting for their dependency on the other regressors (i.e. the other time functions). However, the model can be shown to have an orthogonal structure, we can determine the least squares estimates of \(b_{ik+p_k}, 0 \leq i \leq P-1\), that are free of any dependency on the other regressors \(b_{ik+p_j}, j \neq k, 0 \leq i \leq P-1\). In this case the least-squares estimate of any parameter set is independent of the number \(F\) of time functions, which is a very desirable property. In other words, if the order of the model is changed from \(F\) to \(F+1\), only the coefficients related to the functions \(f_j, F+1, 0 \leq i \leq P-1\), have to be computed. The coefficients related to the functions \(f_j, 0 \leq k \leq F-1, 0 \leq i \leq P-1\), do not change due to the orthogonality property of the polynomials.

By computing the product \(X^TX\), we may express orthogonality conditions, i.e. conditions that make the matrix \((X^TX)^{-1}\) block-diagonal. These orthogonality conditions are:

\[
\begin{align*}
\sum_{i=0}^{M-1} w_{0,i} f_{0,r}(t_i) f_{0,s}(t_i) &= 0, \quad \forall r, s \text{ with } r \neq s \\
\sum_{i=0}^{M-1} w_{1,i} f_{1,r}(t_i) f_{1,s}(t_i) &= 0, \quad \forall r, s \text{ with } r \neq s \\
\sum_{i=0}^{M-1} w_{2,i} f_{2,r}(t_i) f_{2,s}(t_i) &= 0, \quad \forall r, s \text{ with } r \neq s,
\end{align*}
\]

where the weights \(w_{0,i}, w_{1,i}\) and \(w_{2,i}\) are

\[
\begin{align*}
w_{0,i} &= \sum_{j=0}^{N-1} l_i^2(x_j, t_i), \\
w_{1,i} &= \sum_{j=0}^{N-1} l_i^2(x_j, t_i), \\
and \quad w_{2,i} &= \sum_{j=0}^{N-1} l_i^2(x_j, t_i) l_j^2(x_j, t_i).
\end{align*}
\]

These conditions of orthogonality may be obtained in a number of ways. By assuming for \(k = 0, 1, 2\) that

\[
w_{k,i} \sim w_{k,i}, \quad \forall i, j
\]

equations (8) to (8) may simply be reformulated as

\[
\sum_{i=0}^{M-1} f_r(t_i) f_s(t_i) = 0, \quad \forall r, s \text{ with } r \neq s
\]

Setting \(t_i = i, i = 0, \ldots, M-1\), the following system of trigonometric orthogonal functions may be used:

\[
\begin{align*}
f_0(t) &= 1, \\
f_r(t) &= \cos(2\pi rt/M), \quad r \text{ odd, } r \neq 0, \\
f_r(t) &= \sin(2\pi rt/M), \quad r \text{ even, } r \neq 0.
\end{align*}
\]

Note that the use of other orthogonal polynomials such as Chebyshev polynomials is also possible. However, we may also find a set of orthogonal polynomials without assuming that equation (9) is respected. As suggested in \([11]\), we can use the three-term recurrence relationship (in the following equations \(k\) takes the value 0 or 1)

\[
f_{k,r+1}(t) = 2(t - a_{k,r+1})f_{k,r}(t) - b_{k,r}f_{k,r-1}(t),
\]

beginning with the initial polynomials

\[
f_{k,0}(t) = 1 \text{ and } f_{k,1}(t) = 2(t - a_1).
\]

Here \(t\) is normalized so that \(-1 \leq t \leq 1\) and the \(a_{k,r}\) and \(b_{k,r}\) are chosen to make the orthogonality conditions (6) and (7) hold, namely,

\[
\begin{align*}
a_{k,r+1} &= \frac{\sum_{i=0}^{M-1} w_{k,i} l_i^2(t_i)}{\sum_{i=0}^{M-1} w_{k,i} l_i^2(t_i)}, \\
b_{k,r+1} &= \frac{\sum_{i=0}^{M-1} w_{k,i} l_i^2(t_i)}{\sum_{i=0}^{M-1} w_{k,i} l_i^2(t_i)}.
\end{align*}
\]

702
However, also with this formulation for the orthogonal functions, condition (8) is only fulfilled approximately. Nevertheless, experimental results show that these polynomials yield very good orthogonality approximations. Using the orthogonal set of functions defined in equation (11) or (12), the covariance matrix \( (X^T X)^{-1} \) becomes a block-diagonal matrix, where each block corresponds to a given time function. Thus, the estimates and regression sum of squares for any model parameter set \( b_{i+k}, i = 0, \ldots, P-1 \), are free of any dependency on the other regressors set in the model. Sequential fitting of the model is therefore computationally easy, and we can use either the forward selection or backward elimination procedures for determining the appropriate number of time functions. If we wish to assess the significance of a given parameter set, \( b_{i+k}, i = 0, \ldots, P-1 \), then the following statistical test may be used:

\[
H_0: b_{P-k} = b_{1+k} = \ldots = b_{P-1+k} = 0
\]

\[
H_1: b_{i+k} \neq 0 \text{ (for at least one i)}
\]

(14)

This test can be performed using the \( F \)-statistic [3, 11]. Alternatively, we may also test hypotheses on the individual regression coefficients. These tests are helpful in determining the significance of each of the regressors in each parameter set.

<table>
<thead>
<tr>
<th>variable</th>
<th>p-value (individual)</th>
<th>p-value (set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b_0, b_1 )</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>( b_2, b_3 )</td>
<td>0.92</td>
<td>0.99</td>
</tr>
<tr>
<td>( b_4, b_5 )</td>
<td>0.71</td>
<td>0.00</td>
</tr>
<tr>
<td>( b_6, b_7 )</td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td>( b_8, b_9 )</td>
<td>0.58</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Figure 2. Probability values using sequential fitting and orthogonal time functions \((M = 5, P = 2)\)

Table 2 shows the probability values estimated for a sequence in which the motion in the \( x \)-direction is constant and the motion in the \( y \)-direction is oscillating at the maximum frequency. The \( p \)-value is the probability that the corresponding random variable becomes larger than the value that was actually obtained. This table shows that sequential fitting allows us to detect the time functions which are relevant for describing the parameters time variation and to delete the regressors which are not significant, yielding an optimal representation of the motion parameters.

4. Conclusion

Critical points for obtaining an accurate flow field are the size of the region of analysis and the way to constrain the flow field computation. In this paper, we proposed to compute the flow field both on large spatial and temporal support by means of time-varying motion models.

The parameters of the model are expressed as a linear combination of orthogonal time functions, whose number is automatically obtained by using multi-resolution and statistical tests. The main advantage of the proposed algorithm is its ability to constrain the computation of the flow field spatially and temporally at the same time. Thus, more information is used in the estimation scheme, increasing the stability and robustness of the motion parameter estimates. Furthermore, the motion constancy is not required. An application of the method to the problem of motion segmentation is shown in [10].

References


Motion estimation with violation of the constancy brightness constraint

Marco Mattavelli and André Nicoulin

Signal Processing Laboratory, Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland, Tel: +41 21 693 26 01, E-mail: marco@lts.de.epfl.ch

Abstract. In the field of motion estimation, the constancy luminance principle has always been used as hypothesis to constrain the problem of optical flow extraction. When real conditions do not satisfy this assumption, classical motion estimation techniques fail. On the basis of models of image representation and image acquisition, a general model that considers brightness changes by means of a multiplicative and an additive function, has been proposed. In this paper we discuss the problems that the new model poses for the extraction of the optical flow. We show that, in some conditions, the parameters of the model can alias the motion parameters. One consequence is that reliable estimates of the parameters have to be obtained by avoiding aliasing conditions, another consequence is that some classical techniques which use multiresolution approaches to solve the problem of local minima for motion estimation, cannot be used anymore. Finally an algorithm that estimates separately brightness change parameters and motion parameters avoiding aliased conditions is presented with simulation results.

1. Introduction

Motion estimation based on the extraction of information from gray levels of image sequences is usually based on the assumption that in time and with movements, object points maintain the same luminance values. Although this hypothesis helps in constraining an ill posed mathematical problem, in many cases it fails and the obtained motion estimates are affected by major errors. Some examples of inconsistency of the constant luminance assumption are: variation of lighting conditions produced by intensity variation, movements of the lighting sources, motion of objects that produce shadows or reflections on other parts of the scene. In the field of optical flow estimation by differential, recursive or not recursive techniques, the constancy luminance constraint has always been kept as a basic principle to solve the finite difference equations. This constraint is always joined to some other regularization or smoothness conditions on the structure of the optical flow and/or on the residual errors.

Recently it has been proposed to extend the basic gradient constraint equation to consider brightness changes [1]. A least mean square error (LMSE) solution of this gradient equation has been proposed in order to solve the problem. In this paper we discuss other possible solutions and we show that LMSE solutions based on the differential equation or on LMSE correlation methods are not satisfactory. We reformulate the motion estimation problem separating explicitly objects and illumination conditions, and we propose an iterative algorithm that separately estimates motion parameters and illumination variation functions.

In the following section the conditions that induce variations of illuminations in image sequences are briefly presented and a general model that deals with variations of illumination is derived. Sec. 3. discusses the main problems that arise when solving the motion estimation problem without the brightness constraint. A criticism of the LMSE solutions based on the differential equation and on a correlation solution is given. Sec. 4. reformulates the problem separating illumination conditions and objects. Sec. 5. presents the proposed new motion estimation algorithm while sec. 6. reports simulation results and performances. Finally concluding remarks are given in sec. 7.

2. A brightness change model

Sampled images can be modelled as the product of two positive contributes: an illumination component \( L(x, y, t) \) and a reflectance component \( R(x, y, t) \) [2]. Various phenomena may imply complex evolution of these two components. We could list the following ones: the modulation in space of lighting that can induce time modulation of illumination in moving objects, the movement of spatially non uniform lighting sources on stationary or moving objects and many other combination of object motion and lighting conditions variations [3, 4]. There is no way to obtain separately \( L(x, y, t) \) and \( R(x, y, t) \) since only the luminance \( I(x, y, t) \) is available. Therefore a relation between different instant of times can be derived considering a global multiplicative time

---

1 This work was supported by the Swiss National Funds for Scientific Research under CERS contract 2215.2. and contract 20-30376.92
varying 2-D function that models the variation of illumination on the measured luminance. This way the constraint equation that describes the motion in the sequence becomes:

\[ I(x,y,t+1) = G(x,y,t+1) \cdot I(x-d_x,y-d_y,t) + O(x,y,t) + \Gamma. \] (1)

The modeling of the acquisition process leads to consider two other factors: additive noise and non-linearity of the acquisition devices due to noise of the sensor and to saturation effects. These factors can be modeled considering an additional additive function \( O(x,y,t) \) and Gaussian noise \( \Gamma \). A new constraint equation can be written as:

\[ I(x,y,t+1) = G(x,y,t+1) \cdot I(x-d_x,y-d_y,t) + O(x,y,t) + \Gamma. \] (2)

which for small variations is equivalent to:

\[ I(x+\delta x, y+\delta y, t + \delta t) = G(x,y,t) \cdot I(x,y,t) + O(x,y,t) + \Gamma. \] (3)

After Taylor expansion limited to first order terms it can be expressed in differential form [1]:

\[ I_t + \nabla \cdot (I \cdot \nabla g) - I \cdot \nabla c = 0, \] (4)

where \( G = 1 + \delta g, O = n + \delta c, g_t = \lim_{\delta t \to 0} \delta g / \delta t \) and \( c_t = \lim_{\delta c \to 0} \delta c / \delta t \).

3. The brightness change problem

The optical flow estimation with the constant brightness constraint is at a pixel by pixel level an unconstrained problem because only one equation for two unknowns is available. Therefore it is possible to obtain only the normal component of the displacement vector. This condition is well known as the “aperture problem”. The solutions to this problem are usually obtained in different ways, usually by adding various consistency or smoothness constraints on the optical flow structure.

The brightness change model with a multiplier and an additive function is even more unconstrained: the unknowns in this case are four and even the normal component of the motion is undetermined. The solutions of the problem are completely arbitrary unless other constraints are added. The most natural solution consists in considering a local neighborhood of the considered pixel and looking for a LMSE solution of a over-constrained equation system. A first consequence is that we must assume constant or slowly varying brightness change functions and motion parameters inside the local neighborhood. Implicitly the size of the neighborhood defines the class of functions for which we can expect meaningful solutions. The formulation of the problem in the LMSE form is straightforward if the constraint equation is in the differential form of Eq. 4, and has been proposed in [1]. It consists in the LMSE solution of the overconstrained equation system on a local neighborhood of size \( n \cdot n \) of the considered pixel. Some results of the solutions of this equation for different natural images are reported in [1]. The average gradient magnitude in the square region has been used to define the value of \( n \). The results can be considered satisfactory when no motion but only illumination variations are present but are not reliable in case of motion larger than one pixel.

The problem can also be formulated in a non-differential form in terms of the minimization of a local multimodal correlation function over a four-dimensional search space. The iteration of motion and joint estimation of the brightness change parameters converges to a joint parameter estimation.

In our opinion, these formulations are unsatisfactory for reasons that, in the case of the differential approach, are connected. First, the estimation of motion vectors and brightness change parameters is based on the minimization of a unique function that depends locally on spatial and temporal derivatives or on a correlation error. These minimizations lead to what we define the “aliasing” problem, condition that is more serious than the classical “aperture problem”. In fact for smooth luminance surfaces the constraint Eq. 2 allows transformations brightness change parameters and motion parameters that attain the minimum for many values of the parameters. In other words the brightness change parameters and the motion parameters become completely interchangeable and the obtained solutions can be “aliased” estimations of the parameters. A complete analysis of aliasing conditions is beyond the scope of this paper and is omitted for brevity. A second reason that makes the differential approach unsatisfactory is that the differential formulation is not able to solve the problem of local minima. This problem is intrinsic to differential methods even in the case of the constant luminance principle. In this case it is solved with pyramidal or multisresolution approaches iterating the differential solutions on low resolution images and using the obtained motion estimates as initial values for the following iteration at a higher resolution. So, in order to overcome the problem of local minima, we have to solve the problem on “smoothed” versions of the images and that returns to the minimization of a global function of spatio-temporal derivatives that leads to the “aliasing” problem discussed above. Our observations and criticisms are in accordance with the results of [1]. In fact when there is no motion a LMSE constraint on spatio-temporal derivatives is sufficient to achieve the correct solution but, it fails in the most interesting cases of actual motion.
4. Statement of the problem

The projection $R(x, y, t)$ of a real scene on the image plane evolves in time with projected apparent motion given by the equation:

$$R(x, y, t + 1) = R(x - V_x(x, y), y - V_y(x, y), t), \quad (5)$$

where $V_x(x, y)$ and $V_y(x, y)$ can represent any motion. The problem of motion estimation consists in finding the best estimate of $V_x(x, y)$ and $V_y(x, y)$ from the sampled image observations

$$I(x, y, t) = R(x, y, t) \cdot L(x, y, t) + N(x, y, t) + \Gamma,$$

$$I(x, y, t + 1) = R(x, y, t + 1) \cdot L(x, y, t + 1) + N(x, y, t + 1) + \Gamma, \quad (7)$$

where $I(x, y, t)$ is the intensity of sampled image, $L(x, y, t)$ is the illumination component of the image, $N(x, y, t)$ is the additive functions of the brightness change model, and $\Gamma$ is zero mean Gaussian noise. With this definition the estimated motion is related to the reflectance component of the objects that in fact is the information we are interested in. From Equations 6 and 7 we obtain the constraint equation:

$$I(x, y, t + 1) = I(x - V_x(x, y), y - V_y(x, y), t) - G(x, y, t + 1) + O(x, y, t + 1) + \Gamma, \quad (8)$$

expressed in function of the observed image intensity, where $G(x, y, t + 1) = L(x, y, t + 1)/L(x, y, t)$ and $O(x, y, t + 1) = N(x, y, t + 1) - N(x, y, t)$.

5. The proposed algorithm

As we have discussed in previous paragraphs, the main problem in the estimation of the parameters for the brightness change model is the “aliasing” among motion and additive functions for smooth surfaces even when the gradient is relatively high. The main idea in the proposed algorithm is indeed to estimate the parameters in a way to avoid these aliased conditions. This can be reached by separating the information present in the images into two sets according to the its relevance for a particular parameter subset and separately estimate each parameter subset. In order to separate the estimations we consider the images as the sum of their low-pass and high-pass components:

$$I(x, y, t) = I_L(x, y, t) + I_H(x, y, t). \quad (9)$$

From Equations 6 and 7 we can derive:

$$I(x, y, t) = R_L(x, y, t) \cdot L(x, y, t) + N(x, y, t) + R_H(x, y, t) \cdot L(x, y, t) + \Gamma_H + \Gamma_L. \quad (10)$$

Assuming that $L(x, y, t)$ and $N(x, y, t)$ are smooth functions, the application of a linear smoothing operator $S[\cdot]$ such as a low-pass filter to Eq: 10 does not affect significantly the terms $R_L(x, y, t) \cdot L(x, y, t)$, $N(x, y, t)$ and $\Gamma_L$. Therefore we have:

$$I_H(x, y, t) = I(x, y, t) - S[I(x, y, t)] \quad (11)$$

$$I_L(x, y, t) = S[I(x, y, t)] \quad (12)$$

$$I_H(x, y, t) \simeq R_H(x, y, t) \cdot L(x, y, t) + \Gamma_H \quad (13)$$

$$I_L(x, y, t) \simeq R_L(x, y, t) \cdot L(x, y, t) + \Gamma_L \quad (14)$$

The constraint equations for $I_H$ and $I_L$ can then be written as:

$$I_H(x, y, t + 1) \simeq I(x - V_x(x, y), y - V_y(x, y), t) - G(x, y, t + 1) + O(x, y, t + 1) + \Gamma_H \quad (15)$$

$$I_L(x, y, t + 1) \simeq I(x - V_x(x, y), y - V_y(x, y), t) - G(x, y, t + 1) + O(x, y, t + 1) + \Gamma_L. \quad (16)$$

In the previous paragraph we have discussed that the constraint for $N(x, y, t)$ and $I(x, y, t)$ of being slowly varying functions in the considered local neighborhood is the condition for admitting meaningful solutions to the problem. Therefore the hypothesis of smooth functions is justified and does not constitute a strong limitation for the solutions. The spatial cut-off frequency of the smoothing operator has to be related to the dimension of the neighborhood. It should be larger or equal to the inverse of the dimension of the local neighborhood for each dimension. In this way it is possible to eliminate the additive function term $O(x, y, t + 1)$ in the high pass image and to estimate the multiplicative function $G(x, y, t)$ alone as shown in Eq. 15. The problem of aliased estimations among the two brightness change parameters is therefore eliminated for the high pass image. The estimation of motion vectors based on the high-pass image instead might be not reliable for two reasons. It is underestimated in the case of smooth surfaces, and the SNR of the estimation calculated in relation to noise is lower than if it were calculated on the original image, in absence of aliasing given by the brightness change parameters. Therefore the estimation of motion parameters is iterated using the information extracted from the estimation of the brightness parameters in order to solve this problem. These vectors can then be used to obtain a new estimate of $G(x, y, t)$ using Eq. 15 and the procedure can continue as described above until a convergence criterion stops the iterations. The possibility of aliased estimation among motion parameters and brightness change parameters is largely reduced. The cases of smooth surfaces are not critical anymore because they are transformed into horizontal quasi-planar surfaces of different gray levels. Not aliased estimation of the function $G(x, y, t)$ are therefore possible. On the basis of this considerations, the following algorithm is proposed:
1. extraction of $I_H$ and $I_L$ from $I(x, y, t)$ and $I(x, y, t + 1)$,

2. preliminary first estimation of motion vectors $\hat{V}_x(x, y)_n$, $\hat{V}_y(x, y)_n$, $n = 1$ for each local neighborhood $i$ on $I_H(x, y, t)$ and $I_H(x, y, t + 1)$, set $n = 2$,

3. estimation of $\hat{\hat{G}(x, y, t + 1)}$ on $I_H(x, y, t)$ and $I_H(x, y, t + 1)$ for a displacement corresponding to the motion vectors $\hat{V}_x(x, y)$, $\hat{V}_y(x, y)$

4. estimation of new vectors $\hat{V}_x(x, y)_n$, $\hat{V}_y(x, y)_n$ on $I_L(x, y, t) \cdot \hat{g}(x, y)$ and $I_L(x, y, t + 1)$

5. estimation of $\hat{\hat{\hat{O}(x, y, t + 1)}$ on $I_L(x, y, t) \cdot \hat{\hat{G}(x, y, t + 1)}$ and $I_L(x, y, t + 1)$ using Eq. 16

6. estimation of new motion vectors $\hat{V}_x(x, y)_n$, $\hat{V}_y(x, y)_n$ from $I(x, y, t) \cdot \hat{\hat{G}(x, y, t + 1)} + \hat{\hat{O}(x, y, t + 1)}$ and $I(x, y, t + 1)$ using Eq. 8

7. comparison of the motion vector $\hat{V}_x(x, y)_n$, $\hat{V}_y(x, y)_n$ with the vectors obtained with the previous iteration $\hat{V}_x(x, y)_{n-1}$, $\hat{V}_y(x, y)_{n-1}$. Convergence of the algorithm if estimated motion vectors are equal to previous iteration vectors, else the iterative algorithm return to step 3 incrementing $n$.

The estimation of a parameter, noted $\hat{\beta}$, consists in the minimization of the following expression by gradient search methods for the brightness change parameters and by exhaustive search for the motion vectors:

$$\hat{\beta} = \arg \min_{\beta} \sum_{x, y \in \Delta} (F(I(x, y, t), \hat{\beta} - I(x - \hat{V}_x(x, y), y - \hat{V}_y(x, y), t + 1))^2$$ (17)

where $\Delta$ is the local neighborhood, $\hat{\beta}$ are the previously estimated parameters and $F$ is the brightness change model function.

6. Simulation results

Simulations of the proposed algorithm have been performed on natural images artificially processed by translational motion, brightness change according to the model and additive white Gaussian noise. The brightness change functions are combinations of constant, constant gradient and sinusoidal variations. Since the aim of this paper is to propose an algorithm that solve the specific motion estimation problem with violation of the constancy brightness constraint given by the model, the choice of using artificially processed images allows a accurate evaluation of the algorithm performances with known references. Other estimation errors that could be caused by other effects such as uncovered background and motion discontinuities do not affect the results. The proposed algorithm applied to the images Mobile Calendar and Flower Garden artificially perturbed, even by very large brightness changes, converges after three iterations, to 97% and 100% of correct motion estimates. Instead for the same brightness changes instead the joint LMSE estimation algorithm described in section 3, showed systematically much poorer performances. It is important to notice that the algorithm reaches the results using only the information available in the local neighborhood extended to the motion vector range without any information coming from parameters or motion vectors estimated in side areas.

7. Conclusion

In this paper we have discussed the new problems posed by a model that includes multiplicative and additive functions as brightness changes. We have discussed the "aliasing" among motion and brightness parameters when LMSE solutions of all parameters are considered. A new motion estimation algorithm that avoids "aliasing" by a multiresolution iterative parameter and motion estimation has been proposed. Simulation results have shown that it yields correct motion estimates even for large brightness changes while LMSE based solutions fails for not considering the aliasing problem and being trapped in local minima. Further work has to be done to improve the algorithm in order to cope with brightness change function that include discontinuities and to apply regularization algorithms in the case "aliasing" does not allow a reliable local estimation.

References


Moving Segment Detection
in Monocular Image Sequences under Egomotion

Włodzimierz KASPRZAK † and Heinrich NIEMANN ‡‡

†Bavarian Research Center for Knowledge-Based Systems (FORWISS),
Am Weichselgarten 7, D-91058 Erlangen, E-Mail: kasprzak@forwiss.uni-erlangen.de
‡University of Erlangen-Nuremberg, Institute of Computer Science (IMMD 5),
Martensstr. 3, D-91058 Erlangen, E-Mail: niemann@informatik.uni-erlangen.de

Abstract. Two moving segment detection problems in image sequences of road scenes under egomotion are discussed: the classification of image contours into moving or stationary and the 3-D object motion estimation. In both cases two solution classes are compared. The approaches of the first class (bottom-up, application-independend) are searching for image feature correspondence in subsequent images. The second class approaches perform a model-based classification and object motion estimation. Hence the first problem is solved either by image motions of several contour points or by geometric vanishing point detection. After a road detection step the contours are grouped and their depth positions are hypothesized. The two approaches applied to the problem of object motion estimation use either the 2-D length change rate of a contour group or 3-D positions of corresponding groups in subsequent images.

1. Introduction
During early image sequence processing one primarily wants to redetect the previously detected features in the next image in order to stabilize the image description and to obtain the visual motion field [1]. In case of true camera motion in the 3-D space (also called egomotion), not limited to the lateral camera motion, the visual motion consists of two unknown components: the first one corresponds to the object motion and the second one to the egomotion. The visual motion of stationary objects is not unique – it depends on the object position relative to the camera. For a separation of true moving segments from the stationary background the visual motion part caused by the egomotion should be estimated. On the other hand for the calculation of this unknown motion part the stationary segments should be used only.

In this paper two moving segment detection problems in monocular image sequences of road scenes under egomotion are discussed and two classes of methods for solving them are compared: bottom-up methods based on image motion and model-driven methods, based on geometry detection in 2-D or 3-D space. Related works are described for example in [2] (road scene analysis), [3] (focus of expansion detection), [4] (adaptive estimation) and [5] (3-D egomotion estimation).

2. System outline
Let us consider a system for dynamic road scene analysis with a moving observer as depicted in Figure 1. It consists of an application-independent module for image contour detection and image motion estimation (2-D) (left bottom scheme part), that is integrated with a 2.5-D module for contour classification and road detection (top part), and they are both interacting with two model-based modules for object initialization (3-D) and tracking (4-D) (right part).

The 2-D module has been described in [6]. Its task is to detect closed image contours and to estimate the image motion vectors of the contours.

After the contour classification step has been finished (as described in section 3) application-specific knowledge about the scene is used next for the detection of the road class, road width and the observer position relative to the middle road axis.

Because the measured image motion as well as the image locations are very sensitive to the discretization error, two stabilization schemes are applied and redundant measurements are provided. A weighted averaging of individual measurements in a short sequence of up to 5 images is performed and an adaptive scheme is applied (by means of a recursive linear filter) for the stabilization of the image measurements in a long image sequence.

The contour grouping step starts with the backprojection of the contours into the 3-D space over the road plane. Then it tries to aggregate neighbour contours from the same class together. For example the search for an "obstacle" group starts with a non-stationary "road" contour on the bottom of the image. It looks for image neighbours that are located near the first contour, when projected back to the 3-D space over the road. The group should satisfy the geometry restrictions given by the object model (i.e. width-to-height ratio)

All initialized objects are supplied to the tracking module. The relative motion of "stationary" objects is caused by the egomotion of the camera vehicle. The current egomotion measurement is the result of weighted averaging of individual stationary road stripe velocities. The weights correspond to the measurement variances of every object. Both stationary and true moving objects are tracked by the use of a model-to-image match (point- or edge-based) and their states are recursively updated by means of adaptive filtering methods. The tracking module is scope of the paper [7].

In the current paper the design of the contour classification and the object initialization steps is discussed in the context of the following question: how can the 3-D object recognition task in image sequences from a low-cost camera
be supported by the application-independent image motion of the contours. During the contour classification step the image motion in several contour points may allow the separation of moving contours from the stationary ones. In the second step the translational velocity of an object hypothesis may be initialized by using the image contour change rate instead of detecting the depth position difference of corresponding hypotheses.

3. Contour classification

3.1 VP-based contour classification

An example of the contour data results from the 2-D module is given in Figure 2(a). The vanishing point (VP) in the image is detected as the center point of an area with the highest density of hypothetic road line crossings (Figure 2(b)). On the basis of the VP location the image pixels are classified into three classes: "road", "surrounding area" and "heaven". The contours containing some number of "road" pixels are classified as "road" contours, the contours without such pixels but containing enough "surrounding" area pixels are classified as surrounding contours. The remaining contours constitute the "heaven" (Figure 2(c)).

After this basic classification the "road" contours with a large amount of VP-edges (i.e. pointing towards the VP point) are classified as "road stripes" (stationary segments) and the remaining "road" contours are "obstacles" (moving segments) (Figure 3).

3.2 FOE-based contour classification

In the general case the existence of a vanishing point is not guaranteed. In such situation an image motion-based method for moving segment detection could be used instead of the VP-based method.

In Figure 4(a) a closed discrete contour, its features and discrete disparity vectors are shown. The unknown motion \((v_x, v_y)\) of a continuous contour feature is approximated by a weighted averaging of N-1 disparity vectors for a discrete feature point \(- (v_x, v_y)\). The weights are directly related to the additional component of contour motion \(v_z\) – the relative contour length change rate. For the motion vector set of a true moving contour a dynamic focus of expansion point (C-FOE) is estimated. This is the point in an image (in general a region), where the motion lines induced by the feature motion vectors vanish.

The general method of moving segment detection is based on the distinction of individual focus of expansion points for each contour. The idea is to classify contours being the projections of assumed road stripes and the "surrounding" contours into stationary contours. The motion vectors of stationary contours induce then the current dynamic focus of expansion point (FOE). This is the center of an area in the image plane, with the highest density of C-FOE points for stationary contours. A moving contour should differ from the stationary background by its visual motion (its C-FOE does not match the stationary FOE point) (Figure 4(b,c)).

4. Object initialization

The initialization of an object (or generation of an object hypothesis) is equivalent to the initialization of a parametric state vector on the basis of one group features and model-dependent restrictions, with the knowledge about the current camera-to-road transformation. A state vector consists of the trajectory and shape parts. The trajectory subvector \(x(k)\) at time point \(t_k\) is a five-dimensional vector

\[
x(k) = [(p_x(k), p_z(k), \theta(k)), (V(k), \omega(k))]^T,
\]

that consists of the position \((p_x(k), p_z(k))\) relative to the camera vehicle, the orientation \(\Theta(k)\) of the translational motion and the magnitudes \(V(k)\) and \(\omega(k)\) of translational and angular velocities.

The localization parameters (position and orientation) are model-based estimated by projecting the image group features back onto the road coordinates assuming the on-road position. Then the model-based restrictions about the length-to-width and height-to-width ratios allow the orientation estimation.

The angular velocity is assumed to be equal to the current egomotion state parameter \(\omega(k)\). The translational velocity along the depth axis can be hypothesized on two ways. The geometry-based method performs a short-time tracking of the object depth, whereas the application-independent
Table 1: The average errors and the error variances of the VP and FOE estimation in 5 image sequences

<table>
<thead>
<tr>
<th>Image</th>
<th>$\delta_x$</th>
<th>$\sigma_x^2$</th>
<th>$\delta_x^*$</th>
<th>$\sigma_x^2*$</th>
<th>$\delta y$</th>
<th>$\sigma_y^2$</th>
<th>$\delta y^*$</th>
<th>$\sigma_y^2*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 5</td>
<td>1.5 - 6.2</td>
<td>17 - 156</td>
<td>1.1 - 5.2</td>
<td>1 - 28</td>
<td>0.8 - 6.4</td>
<td>6 - 64</td>
<td>0.9 - 6.8</td>
<td>1 - 7</td>
</tr>
<tr>
<td></td>
<td>$x_{VP} = x_{VP} - 20$</td>
<td>$x_{VP} = x_{VP} - 20$</td>
<td>$y_{VP} = y_{VP} - y_0$</td>
<td>$y_{VP} = y_{VP} - y_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - 5</td>
<td>9.5 - 30.7</td>
<td>1148 - 1978</td>
<td>7.5 - 26.9</td>
<td>31 - 340</td>
<td>23.6 - 35.3</td>
<td>107 - 215</td>
<td>21.9 - 32.7</td>
<td>28 - 57</td>
</tr>
</tbody>
</table>

Method applies the contour length change rate $v_c$ as follows:

$$V_c(k) = p_x(k) \left( \frac{\gamma(k)}{1 + v_c(k)} \right) - 1$$

(2)

where the coefficient $\gamma(k)$ depends on the object type and the rate $v_c(k)$.

5. Experiments

Five test sequences have been used with complexity of 100-120 contours in one image and with average contour length of 100-110 pixels.

5.1 FOE vs. VP-estimation

In Table 1 the detected errors $E$ and the variances $\sigma^2$ of the detection $(X, Y)$ and estimation $(X^*, Y^*)$ of the FOE and VP points in five image sequences are given. The original VP locations $(x_0, y_0)$ have been manually measured in the images. As the camera vehicle is moving approximately towards the VP-line, the reference value for the FOE detection error was the same as for the VP error. While comparing the errors and variances an immediate conclusion is that the quality of VP detection is about ten times better than the quality of FOE detection.

Thus the FOE-based approach has failed to reach its goals in practical tests. Unfortunately the differences in measured motion of different contour points have been too small for a robust and stable determination of a C-FOE point for a great number of contours.

5.2 Object motion initialization

There are two road hypotheses tracked in parallel that correspond to a 2- or 3-lane road class. The contours are grouped into 20 - 30 groups with a tracking success of 90 - 95%. Some 10 groups correspond to a moving object hypothesis (there were 3-4 vehicles in the scene) and 10-12 groups induce a generation of up to 6 road stripes for each road hypothesis (Figure 5).

The Table 2 summarizes the results of repeated depth $p_z$ and object motion $V$ initializations for a middle road stripe hypothesis in 20 images. The original values of depth $p_z$ and velocity $v_0$ of the moving object have been measured manually in the image sequence.

The errors of depth estimation were up to ±25% but the errors of translational velocity along the depth axis, calculated from the 3-D location differences, were much higher between −57% and 67%. The quality of the same velocity, but computed by the $v_c$-based method, was better — errors of ±17.5% have been observed. There is a big contour detection instability in the image interval 7-12 as the stripe is passing a highlighted area in the road. During the estimation of the $v_c$ change rate, these errors are partly filtered out by the measurement stabilization procedures and by the redundancy of measurements (the border length and diagonal length change rates are combined).

These errors should be related to the discretization errors of object initialisation in a synthesized image sequence. For a synthetic road stripe of similar size (256x256x8 bit images, contour length from 34 to 106 pixel) the detected measurement error of $v_c$+1 or of the contour motion ($v_c$, $v_{p}$) was up to 0.005/δl or up to 7%. At the same time the errors of individual border point motions were several times larger than this error. With known vehicle speed (ca. 111m/r) the error of translational velocity was below 2m or 7.4%.

6. Conclusion

Two classes of approaches for moving segment detection in image sequences under true camera egomotion have been compared: general image motion-based methods and model-based methods. For the first detection problem the processing results of a model-based method (VP-detection) have been of much better quality than the results of a general solution (dynamic FOE-detection). For the estimation of translational velocity along the depth camera axis a contour-length-change-based method was proposed, which is of better quality than the 3-D location difference measurement of corresponding object hypotheses.

Acknowledgements

The support from the 'Deutsche Forschungsgemeinschaft', Bona, F.R.G., is gratefully acknowledged. The real images are by courtesy of the BMW AG, Munich, F.R.G.

References

Figure 2: Basic contour classification: (a) contours, (b) vanishing point, (c) "road" and "surrounding" contours

Figure 3: VP-based contour classification: (a) VP-edge detection, (b) "obstacle" contours, (c) "road stripe" contours

Figure 4: FOE-based contour classification: (a) the image motion of one contour and the C.FOE-point, (b) the motion vectors of the contours, (c) the detected FOE point

<table>
<thead>
<tr>
<th>Space</th>
<th>Image feature or object state</th>
<th>Image k =</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2 5 8 11 14 17 20</td>
</tr>
<tr>
<td>2-D</td>
<td>Contour length [pel]</td>
<td>34 40 48 74 78 92 106</td>
</tr>
<tr>
<td></td>
<td>Stabilized $v_0$ [pel/$\tau$]</td>
<td>0.064 0.068 0.058 0.063 0.072 0.095 0.106</td>
</tr>
<tr>
<td>3-D:</td>
<td>$VP$-based depth $p_x$ [m]</td>
<td>-43.70 -39.40 -46.15 -38.95 -27.00 -25.95 -18.10</td>
</tr>
<tr>
<td>repeated</td>
<td>$\delta p_x$-based $V$ [m/$\tau$]</td>
<td>1.42 1.03 0.49 1.06 1.04 1.04 1.92</td>
</tr>
<tr>
<td>init</td>
<td>$v_x$-based $V$ [m/$\tau$]</td>
<td>1.46 1.45 1.38 1.33 1.16 1.39 1.15</td>
</tr>
<tr>
<td>3-D:</td>
<td>$p_x$ [m]</td>
<td>-43.27 -39.79 -36.31 -32.84 -29.36 -25.88 -22.41</td>
</tr>
<tr>
<td>original</td>
<td>$V_0$ [m/$\tau$]</td>
<td>1.16 1.16 1.16 1.16 1.16 1.16 1.16</td>
</tr>
</tbody>
</table>

Table 2: Repeated object initialization (depth and translational velocity of "stationary" objects (the first middle road stripe ($p_x, \Theta_S$ are constant and $\omega_S = 0$)) ($\tau = 0.04$sec, pel-pixel side).

Figure 5: Object detection: (a) road lane hypothesis, (b) the car object hypotheses, (c) the road stripe objects
Subspace-based Adaptive Algorithms for the Blind Equalization of Multichannel FIR Filters

David GESBERT, Pierre DUHAMEL† and Sylvie MAYRARGUE

France Telecom-CNET/PAB/RGP/RCN
38-40 rue du Général Leclerc, 92151 Issy les Moulineaux, France
†Télécom Paris, Département Signal, 48 rue Barrault, 75004 Paris Cédex 13, France
E-mail: duhame@eng.ucl.ac.uk, Tel: ++33 1 4039 2825, Fax: ++33 1 4034 2625

Abstract. This paper proposes an algorithm for equalizing channels (possibly non-minimum phase and time-varying) using second order statistics only. This problem occurs in the context of digital radio-communications, and is based on the use of a multichannel formalism. We introduce the concept of mutually referenced equalizers (MRE) for blind spatio-temporal equalization. In the noiseless case, the obtained filters coincide with delayed versions of Zero-Forcing equalizers (ZFE). In the presence of additive noise, the simulations indicate a good behavior and show performances that outperform those of the ZF equalizer at low SNRs, being close to the MMSE results. Furthermore, the main improvement is the possibility of obtaining an adaptive version of the method which, due to the use of second order statistics only, has intrinsically fast convergence properties.

1. Introduction

Transmission through radio channels is impaired by multipath propagation. The introduced intersymbol interference (ISI) makes it necessary the use of an equalization device for the transmitted symbols to be recovered. Equalization traditionally relies on the use of training sequences which are periodically emitted. Hence, some part of the communication time slot is used for acquiring the equalizer taps values rather than for transmitting useful information. The increasing need for higher bit rates makes attractive the methods which do not use such training sequences.

In the recent years, so-called 'blind' techniques have mainly been based on the use of higher-order statistics (even the well known Constant Modulus Algorithm implicitly relies on the use of 4th order statistics). The resulting algorithms often show poor convergence abilities, and are subject to possible local minima. The recent papers by Tong et al. [4, 5], Moulines et al. [2], showed that blind identification of possibly non-minimum phase systems using the second-order information only of the observed signals was feasible, provided that several measurements could be acquired per emitted sample, either by multiple sensors reception or by temporal oversampling. Once the system has been identified, a ZF equalizer is easily obtained. In contrast, we hereby address directly the problem of spatio-temporal equalization without pre-identification.

The new criterion is first presented in a block algorithm, showing the intrinsic possibilities of this approach. The equalizer main characteristics are investigated theoretically in the noiseless case. Then, an adaptive algorithm minimizing the same second-order criterion is presented. Simulations comparing the robustness of the ZF, MR and MMSE equalizers with respect to additive noise are provided.

2. Problem formulation

Our approach builds on the same formulation as the previous work of Tong et al [4, 5]. Notations are as follows:

- $N$: Time extent of the measurements
- $M$: ISI length, assumed to be finite
- $s_n$: Transmitted sequence, assumed to be temporally white with power $\sigma_s^2$
- $b_n$: White noise samples, with power $\sigma_b^2$

Once sampled at the symbol rate, the received sequence $x_n$ is a noisy version of the convolution between the transmitted symbols and the baseband equivalent of the channel response $H = (h_0, h_1, \ldots, h_M)$, which includes all the equipment filtering, modulation and demodulation effects:

$$x_n = \sum_{m=0}^{M} s_{n-m} h_m + b_n$$

In the case of oversampling and/or multi-antenna reception, there are $L = KN$ such impulse responses, corresponding to various virtual channels. Let $x_n^{(i)}$ be the signal received on the $i$th channel with response $H^{(i)} = [h^{(i)}_0, h^{(i)}_1, \ldots, h^{(i)}_M]$. Let $X^{(i)}(n)$ denote the vector obtained by stacking the $N$ last received signals $x_n^{(i)}$. By construction, the measurements between time $n$ and $n - N + 1$ depend on a finite number of transmitted symbols, equal to $P \stackrel{def}{=} M + N$. This results in the following matrix equation:

$$\begin{pmatrix}
  x^{(i)}_{n} \\
  \vdots \\
  x^{(i)}_{n-N+1}
\end{pmatrix} =
\begin{pmatrix}
  H^{(i)}_{0,0} & 0 & \ldots & 0 \\
  \vdots & \ddots & \vdots & \vdots \\
  0 & \ldots & H^{(i)}_{M,P-1} & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & \ldots & 0 & H^{(i)}_{M,N-1}
\end{pmatrix}
\begin{pmatrix}
  s_n \\
  \vdots \\
  s_{n-P+1} \\
  \vdots \\
  \vdots \\
  s_{n-N+1}
\end{pmatrix}
$$

Or, with straightforward notations:

$$X^{(i)}(n) = \mathbf{H}^{(i)} S(n) + B^{(i)}(n) \quad (1)$$

Then, form the snapshot defined by $X_{LN}(n) = (X^{(0)}(n), \ldots, X^{(N)}(n))$ collecting the whole set of available data. Stacking equations (2) for $i = 1 \ldots L$ yields:

712
\[ X_{LN}(n) = H_{LN} S_P(n) + B_{LN}(n) \quad (3) \]

The filtering matrix \( H_{LN} \) (which contains \( L \) vertically stacked blocks \( H_{LN}^1 \)) thus has \( P \) columns and \( LN \) rows.

3. The Mutually Referenced Equalizers

3.1. The subspace approach

Assume that the filtering matrix \( H_{LN} \) is full column rank. Hence, its left-inverse \( H_{LN}^1 \) is such that: \( H_{LN}^1 H_{LN} = I_P \), where \( I_P \) denotes the \( P \times P \) identity matrix. Thus, the transmitted symbol vector \( S_P(n) \) in eqn.(3) can be approximately recovered from the snapshot \( X_{LN}(n) \) using the columns of \( H_{LN}^1 \) as ZF equalizers:

\[ H_{LN}^1 X_{LN}(n) = S_P(n) + H_{LN}^1 B_{LN}(n) \quad (4) \]

In the paper by Tong et al., the SVD of \( H_{LN} \) is first estimated from a subspace-based decomposition of two data correlation matrices with various time-shifts. A simplification of the method was proposed in [2], inspired by the MUSIC algorithm. Once the channel has been identified, the pseudo-inverse \( H_{LN}^1 \) is easily obtained.

In the proposed method, we rather exploit a structural property of the left-inverse of \( H_{LN} \), which characterizes some possible equalizers. More explicitly, in the noiseless case, the outputs of the \( P \) filters represented by the columns of \( H_{LN}^+ \) should be equal, up to delays ranging from one to \( P-1 \) symbol durations (see eqn.(4)). The corresponding criterion is introduced in section 3.2. However, we first give some conditions for the approach to be valid.

Existence of \( H_{LN}^1 \). All subspace based methods referenced above, as well as this one, rely on the left-invertibility of the filtering matrix \( H_{LN} \). This property is ensured by the following assumptions:

- \( H_{LN} \) is vertical, i.e.: \( LN \geq P \)
- \( H_{LN} \) is full column rank \( P \).

The latter point corresponds to the absence of any common zeros between the \( Z \)-transform polynomials associated to channels \( H_{LN} \).

3.2. MR Criterion

Consider a set of \( P \) FIR equalizers with \( LN \) taps: \( V_1, \ldots, V_P \) with \( V_p = [v_{1p}, \ldots, v_{Lp}, n]_1 \) for each \( p = 1 \ldots P \). The successive snapshots \( X_{LN}(n) \) form a common input sequence to these equalizers. At time \( n \), \( P \) scalar outputs, denoted \( y_p(n) \), are measured: \( y_p(n) = V_p^+ X_{LN}(n), p = 1 \ldots P, \) where + denotes the trans-conjugation. The equalization criterion, outlined in section 3.1, corresponds to minimizing the averaged distance, up to one symbol duration delay, between the \( P \) successive outputs. A scalar error vector is correspondingly defined as \( e_p(n) = y_p(n) - y_{p+1}(n+1) \), for \( p = 1 \ldots P \). Let \( E_{P-1}(n) \) denote the vector \( [e_1(n), \ldots, e_{P-1}(n)] \) and \( V \) the \( LN \times P \) matrix with \( V_p \) as the \( p^\text{th} \) column. The error vector writes:

\[ E_{P-1}(n) = [I_{P-1} 0] V^+ X_{LN}(n) - [0 I_{P-1}] V^+ X_{LN}(n+1) \quad (5) \]

We seek a \( LN \times P \) complex-valued matrix \( V_{opt} \) that globally minimizes the following cost function: \( C(V) = \mathcal{E} \parallel E_{P-1}(n) \parallel^2 \), where \( \mathcal{E} \) denotes the mathematical expectation.

3.3. Equalizer characterization

Noisless case. In the case \( \sigma_s^2 = 0 \), there exists a subspace of \( LN \times P \) matrices \( V \) cancelling the cost function \( C(V) \). Each column of \( V \) thus corresponds to a delayed ZF equalizer.

Proof. Consider a \( LN \times P \) matrix \( V \) that cancels \( C \). According to eqn (5), we have, almost surely, for all \( n \):

\[ [I_{P-1} 0] V^+ X_{LN}(n) = [0 I_{P-1}] V^+ X_{LN}(n+1) \quad (6) \]

then, from eqn (3):

\[ [I_{P-1} 0] V^+ H_{LN} S_P(n) = [0 I_{P-1}] V^+ H_{LN} S_P(n+1) \]

Since \( s_m \) is a white sequence, we get:

\[ [I_{P-1} 0] V^+ H_{LN} [0 I_P] = [0 I_{P-1}] V^+ H_{LN} [0 I_P] \]

The latter equation leads to

\[ V^+ H_{LN} = \alpha I_P \text{ for some complex } \alpha \quad (7) \]

\[ V^+ X_{LN}(n) = \alpha S_P(n) \quad \text{for all } n \quad (8) \]

Hence, in the absence of noise in the model, the minimization of \( C(V) \) provides a set of \( P \) ZF equalizers \( V_p, p = 1 \ldots P \) (the columns of \( V \)) such that \( V_p^+ X_{LN}(n) = \alpha s_{m+p-1} \). Equalization is thus performed up to a to multiplicative constant.

Presence of noise. In the case where \( \sigma_s^2 > 0 \), eqn (6) does not hold anymore, and the obtained equalizers (MR-E) differ from ZF-Es. As expected, the difference between these two type of equalizers increases as the SNR lowers. A theoretical study of the MR-E performances with respect to noise is still under investigation, only simulation results are provided in this paper. However, since the MR criterion makes use of many informations (a whole set of ZF equalizers), one can expect that it will perform better than a single ZF equalizer. This is checked by simulation in section 4.

3.4. Minimization algorithm

This subsection provides the general solution for the blockwise global minimization of \( C(V) \), subject to the non-triviality constraint \( Tr(V^+ V) = 1 \).

Kronecker product formulation. For any \( q \times r \) matrix \( M \), let \( M(i) \) denote all the elements of \( M \) strung out in a long column vector: \( M(i) = [m_{11}, \ldots, m_{qr}] \).

Then define for any equation \( AV = W \) (i) with \( A, V, B \) properly sized matrices, the matrix \( A \otimes B \) such that (i) is equivalent to \( [A \otimes B] V(i) = W(i) \) (Kronecker product).

The criterion \( C(V) \) may accordingly be expressed in terms of \( V = Y(i) \) as follows:

\[ E_{P-1}^+(n) = X_{LN}^+(n) V \begin{bmatrix} I_{P-1} & 0 \\ 0 & I_{P-1} \end{bmatrix} - X_{LN}^+(n+1) V \begin{bmatrix} 0 & I_{P-1} \\ I_{P-1} & 0 \end{bmatrix} \quad (9) \]

Let \( Y(n) = X_{LN}(n) \otimes I_{P-1} \) then:

\[ E_{P-1}(n) = Y(n) V \quad (10) \]

where * denotes the complex conjugation. Let \( \mathcal{E}(Y(n)^+ Y(n)) \). The global minimization of \( C(V) \) under constraint \( Tr(V^+ V) = 1 \) corresponds to the computation of the minimal eigenvector of \( \mathcal{E}Y \). Straightforward calculations lead to:

\[ \mathcal{E}Y = A \otimes R_X - J_P \otimes R_L X - J_P^X \otimes R_L^X \]
where \( R_X = E(X_L(n)X_L(n)^T) \), \( R_iX = E(X_L(n + 1)X_L(n)^T) \), \( \Delta \) is a \( P \times P \) diagonal matrix defined as \( \text{diag}(1,2,2,...,21) \), and \( J_P \) is a \( P \times P \) matrix with ones on the diagonal above the main one, and zeroes elsewhere.

Algorithm. The procedure is as follows:
1. Compute covariance matrices estimates \( \hat{R}_X \) and \( \hat{R}_iX \).
2. Build the corresponding \( \hat{R}_Y \) (eqn.10). Find an eigenvector \( V_{opt} \) associated to the minimum eigenvalue of \( \hat{R}_Y \).
3. Reshape \( V_{opt} \) into a \( LN \times P \) matrix \( V_{opt} \).
4. Use any column of \( V_{opt} \) as an equalizer (Choosing the column with minimal norm will however reduce the noise effects).

3.5. Adaptive implementation

The adaptive implementation of the described equalization method mainly relies on the availability of an efficient algorithm for tracking the 'smallest' eigenvector of the time-varying \( LNP \times LNP \) matrix \( R_Y \). Among the classical solutions, based on the minimization of Rayleigh quotients, the gradient-based techniques are most appealing for their simplicity. A first order solution (LMS) was tested. The algorithm was derived from the method by Thompson, see [6] or [3] for practical implementation. The resulting equalizer showed poor convergence ability (more than 10000 snapshots were required to reach convergence in the mean). Note that rank deficiency of the correlation matrix \( \hat{R}_X \) is inherent to this approach (in the noiseless case, the matrix is built to be of reduced rank). Hence, the LMS algorithm cannot provide a satisfactory solution, at least in a straightforward approach.

A second-order solution (conjugate gradient) was implemented and showed an excellent behavior. The algorithm by Chen et al.[1] was chosen and extended to the case of complex signals. The algorithm was modified to reduce the problems induced by the time variation of \( R_Y(n) \) (estimate of \( R_Y \) at \( n \)). A block updating of \( R_Y(n) \) was implemented instead of instantaneous updating, which brought about noticeable gain in the convergence rate, without increasing the complexity. This improvement is due to a more accurate computation of the successive conjugated search directions when the matrix does not change.

The initial complexity is in \( O(LNP^2) \), while the estimate \( \hat{R}_Y(n) \) has some structural properties (hermitian, sparse) that allow a reduction of the global cost. Though it requires more calculations than the LMS algorithm - \( O(LNP) \) - the conjugate gradient method naturally overcomes the ill-conditioning effects. Therefore, it seemed to fit particularly well to our problem.

4. Simulations

This section first compares the performances of the block subspace-based algorithm by Moulines et al.[2] (ZF), the MMSE and the MR equalizers. Four randomly chosen FIR channels were used to simulate the propagation paths (\( L = 4 \)). We had \( N = 3 \) and an ISI of 4 symbols (\( M = 4 \)), the signals being QPSK-modulated. Fig. 1-4 show the constellations of the equalized signals after 5000 snapshots and illustrate the respective robustness of ZF and MR algorithms with respect to noise at 8 and 25 dB SNR. Fig.5 shows for reference the constellations obtained by MMSE equalization at 8 dB SNR. The MRE clearly perform better than the ZF, being close to the MMSE results, as the SNR lowers. This is probably a consequence of the way the memory is treated in both algorithms: since the ZF tries to inverse the channel filters, it ignores the noise. In contrast, the MRE minimizes an MSE-like criterion that implicitly takes noise into account. As expected the difference vanishes as the SNR improves.

Adaptive implementation of the MRE via the conjugate gradient method is illustrated on fig.6. Variation of the bit error rate versus the number of iterations is provided. Rates are computed under the assumption of perfect synchronization, at 15 and 18dB SNR. In both cases, a unique realization of the stochastic algorithm is considered. The adaptive MR equalizers show very fast convergence properties at good SNRs. Of course, performances rapidly degrade for lower SNRs (See fig.6).

5. Discussion

A new multichannel equalization criterion (MRE) has been introduced that only uses the second-order statistics of the received signals. The corresponding equalizers generally perform better than the ZFE-estimate by Moulines et al., especially at low SNRs. Note however that the accuracy of the equalization depends on the choice of the parameters (\( N \) and \( L \)). In contrast to previous similar works, no diagonalization of the spatio-temporal correlation matrix is required. The MRE derivation is based on the minimization of a MSE criterion which is shown to be equivalent to a familiar problem: the computation of a minimal eigenvector. Consequently, numerous known algorithms might be used to perform the minimization adaptively. Among them, the conjugate gradient is most appealing for its robustness with respect to ill-conditioned correlation matrices. Yet, it implies heavy computational cost. The utilization of LMS-based algorithms which are normalized in some sense to offer a trade-off between convergence rate and complexity will be investigated.

References


714
Blind Channel Identification
Using 2nd-Order Cyclostationary Statistics

Stephan V. Schell and Dale L. Smith
Dept. of Electrical Engineering
Penn State University
University Park, PA 16802
(814) 863-8102
Fax: (814) 865-7065
Email: schell@gauss.ece.psu.edu

William A. Gardner
Dept. of Electrical Engineering
University of California
Davis, CA 95616
(916) 752-1951
Fax: (916) 752-8428

Abstract
Digital communication systems for personal and portable wireless communication, mobile and indoor communication, and manufacturing-shop communication are growing in importance. An important performance feature of many of these systems is that they must operate reliably in the presence of time-varying multipath propagation conditions. To address this problem, blind adaptive equalization has received much recent research attention. In this paper, we present a method which exploits the second-order cyclostationarity present in the data stream. A novel extension uses prior knowledge of the transmitter pulse to substantially reduce the amount of data and/or input SNR necessary for the channel identification while simultaneously reducing the computational complexity.

1 Introduction
Adaptive equalization is a standard technique for mitigating system performance degradation that results from time-varying multipath channel distortion. Many conventional equalization methods are available, although each exhibits one of the following drawbacks when considered for use with a rapidly time-varying channel: training symbols must be sent occasionally (sometimes frequently), which diminishes the resources that might otherwise be used to send messages; medium-to-high-quality estimates of the channel characteristics (e.g., its impulse response) are needed, and these are usually impossible to know a priori in the time-varying environment and can be difficult to estimate; and, the channel distortion must be mild enough that a low bit-error rate (BER) is obtained prior to equalization, which is typically not the case in the presence of moderate to severe multipath.

Alternatively, much work has been done on blind adaptive equalization methods that are intended to mitigate the effects of the time-varying channel distortion without suffering from the drawbacks of more conventional methods. Most of these blind methods operate on the received data, sampled at the symbol rate, and thus ignore the second-order cyclostationarity (SOCs) that almost all digital communication signals exhibit. Instead, they exploit (implicitly or explicitly) their higher-order statistics (HOS). Furthermore, some of them require the computationally intensive solution of multidimensional nonlinear optimization problems, which can preclude practical implementation where cost and latency must be low. Even in those operating environments or those operating scenarios where cost and processing time are not issues, some of these methods are known to exhibit misconvergence, wherein the iterative solution method converges to a useless or unreliable solution. Also, most of these methods require very high signal-to-noise ratios (SNRs) at the input and/or the collection of received data over a large number (hundreds or thousands) of symbol periods in order to operate properly.

In an attempt to bypass some of the pitfalls inherent in the HOS methods, various researchers have exploited the cyclostationarity present in almost all communication signals. Recent work by Gardner [1], Liu et al [2], Baccala and Roy [3], and Schell et al [4] (and many other researchers) shows that structured channels (e.g., MA, AR, ARMA) driven by cyclostationary inputs can be identified using only second-order statistics of cyclostationarity measured at the receiver.

In this paper, we present a second order blind channel identification procedure. Computer simulations in the paper show the bit error rate (BER) performance of this method when coupled with a fractionally spaced equalizer (see chapter 7 in [5]). This is compared to the BER performance when there is
a priori knowledge of either the channel (transmitter pulse plus propagation channel) or of just the transmitter pulse.

2 Data Model and Notation

Consider a QAM signal, corrupted by channel distortion and stationary additive white noise, sampled \( T \) times per symbol period:

\[
x(kT + n) = \sum_{i=0}^{L-1} g(iT + n)s(k - i) + i(kT + n),
\]

where \( g(t) \) is the unknown channel response having length \( LT \), \( T \) is the symbol period, \( s(k) \) is the \( k \)th symbol, and \( i(t) \) is the noise. The symbol sequence is modeled as being stationary and white. If we perform a time-division demultiplexing operation with period \( T \) we obtain \( T \) signals,

\[
y_n(k) = \sum_{i=0}^{L-1} h_n(i)s(k - i) + j_n(k)
\]

\[= h_n(k) * s(k) + j_n(k),\]

for \( n = 1, \ldots, T \) where

\[
y_n(k) = x(kT + n - 1),
\]

\[
h_n(k) = g(kT + n - 1),
\]

\[
j_n(k) = i(kT + n - 1).
\]

That is, the \( T \) signals \( y_1, \ldots, y_T \) are the outputs of \( T \) subchannels \( h_1, \ldots, h_T \) excited by the same input, \( s \).

We note here that this model also characterizes the output of a \( T \)-sensor antenna array receiving a signal \( s \), where the signal received at each sensor has propagated through a different channel.

To define other notation, denote

\[
y_n(k) = \begin{bmatrix} y_n(k) \\ \vdots \\ y_n(k - L + 1) \end{bmatrix},
\]

\[
h_n = \begin{bmatrix} \hat{h}_n(0) \\ \vdots \\ \hat{h}_n(L - 1) \end{bmatrix}, \quad z_{mn}(k) = \hat{h}_{mn}y_n(k),
\]

with \( \hat{h} = [\hat{h}_1 \cdots \hat{h}_T]^t \), and \( R_{mn} = \langle y_m(k)y_n^H(k) \rangle \). Superscripts \((\cdot)^*\), \((\cdot)^t\), and \((\cdot)^H\) denote conjugation, transposition, and conjugate transposition, respectively. Thus, the vector \( \hat{h} \) is a shuffled estimate of the overall channel response,

\[
\hat{g}(kT + n - 1) = \hat{h}_n(k)
\]

where \( k = 0, \ldots, L - 1 \) and \( n = 1, \ldots, T. \)

3 Subchannel Response Matching (SRM) Algorithm

3.1 Algorithm Motivation

We propose to identify the unknown channel \( g(n) \) by identifying its subchannels \( h_1, \ldots, h_T \) using subchannel response matching. That is, considering for the moment only two subchannels indexed by \( m \) and \( n \), we propose to find \( \hat{h}_m \) and \( \hat{h}_n \) such that the two signals defined by

\[
\hat{z}_{mn}(k) = \hat{h}_m(k) * y_n(k)
\]

\[
= \hat{h}_m(k) * h_n(k) * s(k) + \hat{h}_m(k) * j_n(k)
\]

and

\[
\hat{z}_{mn}(k) = \hat{h}_n(k) * y_m(k)
\]

\[
= \hat{h}_n(k) * h_m(k) * s(k) + \hat{h}_n(k) * j_m(k)
\]

are optimally close in the mean-square sense. The following minimization problem follows directly:

\[
\min \left\{ \langle \hat{z}_{mn}(k) - z_{mn}(k) \rangle^2 \right\},
\]

subject to the constraints that neither \( \hat{h}_m \) nor \( \hat{h}_n \) is identically zero.

Clearly, in the absence of noise, the choices \( \hat{h}_m(k) = h_m(k) \) and \( \hat{h}_n(k) = h_n(k) \) yield an optimal match: \( z_{mn}(k) = \hat{z}_{mn}(k) \). Since the criteria for identifiability for a mathematically equivalent approach are given in [3, 2], they are not repeated here.

This subchannel response matching problem was shown in [4] to be equivalent to the following problem:

\[
\min \hat{h}^t S \hat{h},
\]

subject to the constraint \( \| \hat{h} \|^2 = 1 \), where

\[
S = \left[ I \otimes \sum_{n=1}^T R_{nn} \right] - R,
\]

\[
R = \begin{bmatrix} R_{11} & R_{21} & \cdots & R_{T1} \\
R_{12} & R_{22} & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
R_{1T} & \cdots & \cdots & R_{TT} \end{bmatrix},
\]

and the identity matrix has dimension \( T \times T \), and \( \otimes \) denotes the Kronecker product. Since it can be shown that the matrix \( S \) is positive semidefinite, then (5) is a well-known problem whose solution is simply the conjugate of the least dominant eigenvector of \( S \).
3.2 Algorithm Summary

The algorithm can be summarized as follows:

1. For each pair \((m, n)\), where \(m = 1, \ldots, T\) and \(n = 1, \ldots, T\), estimate \(\mathbf{R}_{mn}\) from the subchannel outputs \(y_m(k)\) and \(y_n(k)\).

2. Find and conjugate the least dominant eigenvector of the matrix \(\mathbf{S}\) formed according to \((6)\). This yields the collection \(\mathbf{h}\) of subchannel estimates.

3. This vector can then be shuffled according to \((3)\) to obtain a complete impulse response of the overall channel.

4 Exploiting Prior Knowledge of the Transmitter Pulse

In this section we generalize the least-squares algorithm of the previous section to exploit knowledge of the transmitter pulse.

4.1 New Algorithm Derivation

Here we assume that the impulse response \(p(n)\) of the transmitter filter is known and has length \(L_p\). Denote by \(c(n)\) the impulse response of the propagation channel with length \(L_c = LT - L_p + 1\). Then we can reexpress the overall channel response \(h(n)\) as

\[
g(t) = \sum_{m=0}^{L_c-1} p(t-m)c(m). \tag{8}\]

Thus, the \(k\)th sample of the \(n\)th subchannel can be expressed as

\[
h_n(k) = g(kT + n - 1) = q_n(k)c, \tag{9}\]

where

\[
q_n(k) = [p(kT + n - 1) \cdots p(kT + n - L_c)]^T \tag{10}\]

and \(c = [c(0) \cdots c(L_c - 1)]^T\). Consequently,

\[
h = [h_0^\top \cdots h_T^\top]^\top = Qc \tag{11}\]

\[
h = [q_0^\top \cdots q_{LT}^\top] = Qc \tag{13}\]

where \(Q\) has dimension \(L_c \times LT\) and is defined by

\[
Q = [Q_1 \cdots Q_T]. \tag{13}\]

The algorithm of the previous section is then easily generalized by adding the constraint \(\mathbf{h} = \mathbf{Q}\hat{\mathbf{c}}\). The resulting algorithm thus expends its resources to estimate only the unknown propagation channel response by the vector \(\hat{\mathbf{c}}\). Since \(L_c < LT\) (and in some cases, such as those in which a bandwidth-efficient signal propagates through a channel having a short memory, \(L_c \ll LT\)), the computational complexity of the eigendecomposition is reduced. In fact, the problem reduces from an eigendecomposition of an \(LT \times LT\) matrix to an \(L_c \times L_c\) calculation.

After estimating \(\hat{\mathbf{c}}\) using the extended algorithm, the overall channel response can be estimated by forming

\[
\hat{\mathbf{h}} = \mathbf{Q}\hat{\mathbf{c}}. \tag{14}\]

4.2 New Algorithm Summary

To exploit prior knowledge of the transmitter pulse, the following algorithm is proposed:

1. For each pair \((m, n)\), where \(m = 1, \ldots, T\) and \(n = 1, \ldots, T\), estimate \(\mathbf{R}_{mn}\) from the subchannel outputs \(y_m(k)\) and \(y_n(k)\).

2. Find and conjugate the least dominant eigenvector of the matrix \(\mathbf{Q}\mathbf{S}\mathbf{Q}^H\), where \(\mathbf{Q}\) and \(\mathbf{S}\) are formed according to \((10),(11),(13)\) and \((6)\), respectively. This yields the estimate \(\hat{\mathbf{c}}\) of the propagation channel.

3. The vector \(\hat{\mathbf{h}} = \mathbf{Q}\hat{\mathbf{c}}\) can then be formed and shuffled according to \((3)\) to obtain a complete impulse response of the overall channel.

5 Simulations

In this section, several simulations compare the bit error rate (BER) performance of the SRM method(s) with the performance obtainable with apriori knowledge of the entire channel response (transmitter filter convolved with the propagation channel) or with apriori knowledge of just the transmitter filter. For the simulations, a channel estimate is generated using the appropriate method, and then a fractionally spaced equalizer is generated using the method given in [5]. Four different cases are compared: a known channel, a known transmitter pulse, the channel estimated with the SRM method, and the channel estimate from the SRM method with prior knowledge of the transmitter pulse. The simulations use 100 trials at each SNR \((0,3,6,\ldots \text{dB})\) to calculate the BER.

5.1 Simulation 1 : Two-ray Multipath Channel

This simulation considers a two ray multipath model. The transmission pulse combined with the channel has impulse response

\[
p(n) + \left( -0.9 + 0.1j \right)p(n-1)
\]

where \(p(n)\) is a square-root Nyquist-shaped pulse with 35% excess bandwidth.

The signal is QPSK with a symbol rate of 0.25 \((T = 4)\). In each trial, 162 symbols were used for equalization. For the SRM method(s), the channel length was known \((LT = 48)\) as was the length of the
the propagation channel ($L_c = 2$). For equalization, a 23 tap T/4 equalizer was used. The results of this simulation are shown in Figure 1. In this case, the SRM with prior knowledge blind equalizer has performance almost exactly the same as when the channel is known ahead of time. The regular SRM method does not perform well at these data lengths and SNR's.

5.2 Simulation 2: Four-ray Multipath Channel

This simulation is similar to the previous one, except with four multipath components. The transmission pulse combined with the channel has impulse response

$$4p(n) - 4p(n - 1) + 4p(n - 2) + (-.2 + .1j)p(n - 3)$$

where $p(n)$ is a square-root Nyquist-shaped pulse with 35% excess bandwidth.

The signal is QPSK with a symbol rate of 0.5 ($T = 2$). In each trial, 162 symbols were used for equalization. For the SRM method(s), the parameters $LT = 28$ and $L_c = 4$ were used. For equalization, a 19 tap T/2 equalizer was used.

The results are shown in Figure 2. In this case, the BER of the SRM method with prior knowledge is considerably better than merely knowing the transmitter pulse, and is approaching the performance possible with a known channel at higher SNR's.

6 Conclusions

This paper presents a second order blind equalization method which takes advantage of the cyclostationary nature of the data. The simulations show that performance using the SRM method in some cases can approach that of a system which has prior knowledge of the complete channel impulse response. Since it is a second order method, just a small amount of data is necessary, which makes blind equalization a viable option in the presence of rapidly time varying multipath. In addition, the extension to the case of a known transmitter pulse gives much better performance, as well as reduced computational complexity compared to the original algorithm.

References


Adaptive Blind Equalisation and Demodulation
Without Channel and Signal Parameter Extraction†

Josep SALA-ALVAREZ, Gregori VÁZQUEZ-GRAU

Dpt. of Signal Theory and Communications
Universitat Politècnica de Catalunya
Apt 30002, 08071 Barcelona, Spain
Tel: +34-3-4016440
Fax: +34-3-4016447
E-mail: alvarez@tscc.upc.es / gregorit@tscc.upc.es

Abstract. The problem of filtering a source of known statistics in noise and/or interference of unknown statistics is addressed in this paper. In a general framework, the optimum estimation (detection) of the desired signal requires perfect knowledge of the probability density functions of the interfering signal and usually leads to complex non-linear algorithms. Herein we have restricted ourselves to linear estimators and optimize performance within this subset according to a criterion derived from information-theoretic considerations. For validation purposes, the particular application of blind data demodulation has been considered. Synchronisation, carrier phase rotation and equalization are performed with the same algorithm to force a source of a known pdf at the output of the adaptive system. Performance at low SNR is derived through simulations.

1. Introduction
The motivation of this paper stems from an ample variety of communications problems where a receiving station has to retrieve the transmitted data from a channel-distorted versions of the original signal. The purpose has been to derive an adaptive algorithm that can ensure blind demodulation, without a training signal. All available statistical knowledge of the clean modulated signal may be employed. In this sense, acquisition is performed on availability of a statistical reference rather than on a time reference (training sequence). The algorithm herein developed is especially suited for those applications where at least partial knowledge of the signal statistics is available and very little is known about the characteristics of the noise or interfering signals. Provision against effects such as coloured noise, unknown channel impulse response, unbalanced I-Q channels, can be achieved by the same algorithm and a suitable architecture [1]. We have striven for its robustness and simplicity. As a difference to cumulant-based algorithms [2], that estimate the channel from statistics of the incoming signal, this algorithm can be classified as of the Bussgang type. A suitable cost function is derived whose minimization ensures correct demodulation.

2. Signal Model
The signal model of a linearly modulated signal and related notation are presented in this section. Let us assume that a sequence of symbols $a_k$ is modulated on a Nyquist pulse $p(t)$ and transmitted through a channel of impulse response $h_c(t)$. The received signal is expressed by the following equality,

$$ r(t) = \sum_{k=-\infty}^{k=\infty} a_k \left( p^\ast h_c \right)(t-kT) + n_c(t) $$

where $n_c(t)$ stands for an additive Gaussian noise process. Synchronisation with the symbol sequence as well as carrier phase offset are implicit in the channel response $h_c(t)$. The demodulation architecture is defined in terms of a linear filter $h_0(t)$ working at the symbol rate on the incoming signal, such that the estimates of the symbol sequence are given by,

$$ \hat{a}_k = (r \ast h_0)(t-kT) $$

Note that a linear operation on the incoming data is sufficient to resolve stationary effects as symbol time offset or phase rotation (provided that the architecture of the linear filter $h_0(t)$ is suitably defined). Frequency Doppler can also be compensated for without inclusion of an NCO (Number Controlled Oscillator) if the Doppler uncertainty range is much smaller than the signal bandwidth. Synchronisation can be corrected for with a simple time shift of $h_0(t)$. In a digital implementation we will assume that samples are taken at the rate $1/T_s$, such that the available signal is discrete with $n_c=r(kT_s)$. The filter $h_0$ will be considered as a FIR of $N_c$ coefficients. Hence, if successive snapshots of the signal are taken at the rate of once per symbol, $r_k = [r(k), r(k-1), \ldots, r(k-N_c+1)]$, the received signal can be expressed as the convolution of the symbol sequence $a_k$ with a vector sequence $h_r$ of length $N_c$ that contains the discrete snapshots of the impulse response $r^\ast h_c$ in the format already defined for $r$.

†This work has been supported by the National Research Plan of Spain CICYT, Grant Number TIC92-0800-C05-05

720
\[ r_z = a^* h_r + n_r(k) = \sum_{i=0}^{\infty} a_i h_r(k-i) + n_r(k) \]  

The symbol estimates are then obtained via multiplication with the coefficient vector \( w \) of \( h_k \),

\[ \hat{a}_k = w^T r_z = a_k w^T h_r(0) + \sum_{i=0}^{\infty} a_i w^T h_r(k-i) + n_r(k) \]  
The objective of the filter \( w \) is to minimise the power of the interfering terms, Gaussian and convolutive noise, to minimise the bit error rate on \( a_k \). It would be desired that \( w \) were orthogonal to all vectors \( \{ h_r(i), i \neq 0 \} \) although this seldom happens as \( h_r \) is usually infinite response. The coefficients of \( w \) are updated to minimise a suitable cost function to be derived in the next section.

3. Derivation of the Cost Function

Assuming perfect knowledge of all relevant model parameters, we seek to derive the optimal estimator of \( a_k \) within the subset of all possible linear estimators. For the model presented in the introduction, the estimator should (approximately) converge into the subspace orthogonal to the span of \( \{ h_r(i), i \neq 0 \} \). As the optimization criterion, we have chosen the minimization of a cost function akin to the discrimination between the pdf of \( a_k \) and that of its estimate \( \hat{a}_k \). Let the Kullback-Leibler information measure (or minus cross-entropy) [3] between two probability density functions, \( p_X(x) \) and \( p_Y(x) \), be expressed as,

\[ D_{KL} = \int p_X(x) \ln \frac{p_X(x)}{p_Y(x)} dx \geq 0 \]  

We wish that the statistics of the estimate be as close as possible to the prior pdf of \( a_k \). Let \( A \) denote the random variable at the output of the estimator for the optimal case when all parameters of interest (\( a \), pdf of \( n_z \)) are available. Also let \( \hat{A} \) denote the output of the estimating filter \( w \). Therefore, minimization of the discrimination between the real pdf of \( \hat{A} \) and the desired pdf of \( A \) is enforced.

\[ w = \arg \min \left\{ E_a \ln p_A(\hat{a}) - E_a \ln p_A(\hat{a}) \right\}; \hat{a} = w^T r \]  

The stumbling block of this approach is that the pdf of \( \hat{A} \) appears explicitly in the first term of the discrimination, which equals the minus entropy of the symbol estimates. Its estimation is not straightforward (some procedures are referenced in [2] based on high order statistics of the process and hence difficult to obtain). The minimum attainable output noise variance is also unknown. Therefore we propose instead minimisation of the following cost function, obtained as the second summand in eq.6,

\[ w, \sigma_a = \arg \min \left\{ -E_a \ln p_{A_0+N(\sigma)}(\hat{a}) \right\} \]  

where now minimisation has been also extended to the parameters characterising unknown additive noise affecting the output symbols. The adaptive algorithm implementing this criterion steers the output statistics through \( w \) and controls the shape of the non-linearity through \( \sigma \). The output pdf is modelled as that of \( A_0+N(\sigma) \), with \( A_0 \) a random variable with the same pdf of the symbols and \( N(\sigma) \) additive noise parametrised by its variance \( \sigma \) (or maybe some other parameter for non-Gaussian distributions, see Section 4.1). Intuitively, when the distribution of \( A \) is such that its most likely values coincide with the maxima of the pdf of \( A_0+N(\sigma) \), the expectation of the non-linearity in eq.7 reaches its minimum value. Although the parametrisation of the objective pdf may not be exact (due to interference or convolutional noise), this effect renders a robust cost function for an adaptive algorithm. Application of this cost function also minimises noise variance at the output of the adaptive filter. Additive noise on a random process \( X(n) \) manifests itself as a broadening of the peaks of its pdf. If the maxima of \( p_X(x) \) coincide with the (convex) peaks of the non-linearity, this broadening should be minimised so as to gather most samples of \( X(n) \) around the peaks of the objective pdf. Thus, the expectation in eq.7 is minimised. In general, the proposed cost function will be multimodal depending on the involved statistics. The criterion outlined in eq.7 should be understood as one of absolute minimization. When the joint minimisation in eq.7 is carried out, it does not necessarily imply that the distribution of \( \hat{A} \) equals that of \( A \) although it is similar. In practice we have observed this distributional bias in high noise situations as constellation shrinkage. A rigorous proof falls out of the scope of this article. Nevertheless, if the pdf of \( \hat{A} \) is kept fixed, partial minimisation with respect to \( \sigma \) can cause that \( p_{\hat{A}} \rightarrow p_{A} \) (provided that the parametrisation of \( A \) is exact enough), as endorsed by the cross-entropy (eq.5).

3.1 Relationship with DFE

We will show in this section that the proposed cost function is equivalent to DFE algorithms in the medium to high range of signal-to-noise ratios and in the tracking regime. To that purpose let us consider the distribution of a discrete constant envelope modulation as given in eq.12 with the constraint that \( \sigma^2 = 1 \) for all symbols. Let us also assume that SNR is high: \( |a|^2 < 1 \), with \( n_z \) additive noise and that we are already in a tracking regime. Then, when symbol \( a_k \) is present, all Gaussian exponentials in eq.12 except that corresponding to the actual symbol are vanishingly small (depending on the reliability assigned by the tentative variance \( \sigma \)). The criterion can then be approximated with most likelihood to a least squares of the type,

\[ -E_a \ln p_{A_0+N(\sigma)}(\hat{a}) = E_a \frac{1}{\sigma^2} [\hat{a} - a_k(\hat{a})]^2 \]  

save additive constants, with \( a_k(\hat{a}) \) the closest symbol to the estimate \( \hat{a} \). It has been thus proven that the proposed criterion is equivalent to a Decision Directed (DD) algorithm for high SNR. For low SNR, other symbols than the closest also influence the cost function, for the reliability of the decision is not so high. After some
algebra, the criterion in eq.7 can also be shown to equal the CMA [4] algorithm for constant amplitude modulations.

4. Adaptive Algorithm

The coefficients of the adaptive algorithm will be updated in terms of the gradient of the cost function defined in Section 3. Note here that the gradient will also have to be calculated with respect to the parameterisation of the noise (see eq.7). Nevertheless, as the expectation operator cannot be realised in practice, we will take a finite sample estimate, such that the non-linearity evaluated at the output of the matched filters is given by the following expression,

\[ J = -N_m^2 \sum_{k=1}^{N_m} \ln p_{\hat{a}_k+N(\sigma)}(w|x_k) = N_m^2 \sum_{k} J_k \]

When \( N_m > 1 \), \( J \) turns out to be a memory non-linearity. When the gradient with respect to the coefficients is calculated we have that,

\[ \nabla_{w_k} J = -N_m^2 \sum_{k=1}^{N_m} e^{i\theta_k} \hat{p}_{\hat{a}_k+N(\sigma)}(\hat{a}_k) r_k \]

\[ = -N_m^2 \sum_{k=1}^{N_m} e^{i\theta_k} r_k \]

Where a generalised error \( e(\hat{a}_k) \) is introduced in terms of the target distribution. Note though that this is not the actual error with respect to the transmitted symbols. From this expression, we can see that we can evaluate the covariance of the terms in the summation weight by the exponential of \( J_k \). Therefore, when the output value \( \hat{a}_k \) has a low likelihood with respect to \( \hat{a}_k \), the exponential term is very high. In this way, when we are far from convergence, the gradient tends to accelerate. As we approach convergence such that the cost function decreases, the weighting terms become less and less important, stabilizing the algorithm. The gradient with respect to the noise parameter is given by,

\[ \nabla_{\sigma} J = -N_m^2 \sum_{k=1}^{N_m} \frac{d}{d\sigma} \hat{p}_{\hat{a}_k+N(\sigma)}(\hat{a}_k) \]

For the particular case of linearly modulated signals, the objective pdf is given by the convolution of the discrete constellation pdf and the model of the underlying noise distribution. Assuming a Gaussian distribution we are led to,

\[ p_{\hat{a}_k+N(\sigma)}(\hat{a}_k) = (\frac{2}{\pi \sigma^2} M_0)^{-1} \sum_{i=1}^{M_0} e^{-|\hat{a}_k - a_i|^2/\sigma^2} \]

with \( M_0 \) the number of modulation levels. The error \( e(\hat{a}_k) \) with respect to the coefficients will then be expressed in terms of the following non-linearities,

\[ q_i(\hat{a}_k) = e^{i\theta_k} a_i^2/\sigma^2 \sum_{i=1}^{M_0} e^{-|\hat{a}_k - a_i|^2/\sigma^2} \]

as,

\[ e(\hat{a}_k) = \sum_{i=1}^{M_0} q_i(\hat{a}_k)(\hat{a}_k - a_i) \]

Note that the set of non-linearities \( \{ q_i, 1 \leq i \leq M_0 \} \) play the role of a measure of the likelihood that \( \hat{a}_k \) contains the symbol \( a_i \). They are estimates of the conditional probability \( p(a_i|\hat{a}_k) \) parametrised by the tentative variance \( \sigma \). If \( \sigma \) is taken to be the actual noise variance, we obtain by application of the Bayes conditional probability rule that \( q_i(\hat{a}_k) = p(a_i|\hat{a}_k) \). Moreover, it holds that \( \sum_{i=1}^{M_0} q_i(\hat{a}_k) = 1 \) as was to be expected. Thus, the error defined in [4.6] appears as a weighted average of the error between the output sample \( \hat{a}_k \) and all symbols in the constellation. Symbols farther apart than the closest neighbours will have a lesser influence on \( e(\hat{a}_k) \) as parametrised by \( \sigma \).

In its turn, the derivative with respect to the tentative variance is expressed as,

\[ \frac{d}{d\sigma^2} J_k = -\frac{1}{\sigma^2} \left( \frac{1}{\sigma^2} \sum_{i=1}^{M_0} q_i(\hat{a}_k)(\hat{a}_k - a_i)^2 \right) \]

\[ = -\frac{1}{\sigma^2} \left( \frac{\hat{a}_k^2}{\sigma^2} \right) \]

This gradient can also be associated to a measurement of the error between the tentative variance \( \sigma \) and the actual noise variance \( \sigma_n \). Just note that a weighted average is performed of the squared errors of the symbol estimate \( \hat{a}_k \) with respect to the constellation symbols.

A common pitfall when adaptively changing the cost function itself (as we are doing through adaptation of the tentative variance) is encountered when the filter coefficients are not updated with an equivalent step-size normalized to the curvature of the new cost function. The consequence of this is higher misadjustment noise and possibly divergence as a higher curvature must be offset with a smaller step-size. Then the gradient with respect to the coefficients should be normalised by \( 1/\sigma^2 \), which is approximately the curvature of \( J \) evaluated at the peaks. The tentative variance has been bounded from below as \( \sigma^2 = \sigma_0^2 + \Delta^2 \), in order to preclude that the cost function migrates to an approximately Gaussian distribution during adaptation. This would recover noise instead of data at the equalizer output. \( \Delta \) becomes then the adaptive parameter. The coefficient update equation results,

\[ w_{k+1} = w_k - \mu (\Delta^2 \nabla J) \]

\[ \sigma_1^2 = \sigma_0^2 + \Delta^2 \]

4.1 Inclusion of Other Distributions

Up till now a Gaussian model has been assumed for the distribution of the additive noise \( N(\sigma) \). Nevertheless, other models can also be utilised. In fact, the algorithm is quite robust as concerns the choice of the noise pdf provided that the model pdf and the actual one display similar shapes. We put forward here utilisation of a Laplacian noise distribution to take into account possible interference not accurately modeled by a Gaussian (longer tails). The real Laplacian distribution for the noise \( N(\lambda) \) is given by,

\[ p_{N(\lambda)}(n) = (2\lambda)^{-1} e^{-|n|/\lambda} \]
In this way, the criterion in high noise is found equivalent

to minimisation of the absolute error between the output
sample $\hat{a}_k$ and the closest symbol $a_i(\delta)$.

$-E_b \ln p_{\hat{a}_k,X(\delta)}(\hat{a}) = E_b \frac{1}{\lambda} \left[ \text{Re} \ \delta + |\text{Im} \ \delta| \right]$, $\delta = \hat{a} - a_i(\delta)$

save additive constants. The expressions for the gradient are
calculated as in eq.10 and eq.11.

5. Simulations

Blind demodulation of a QPSK signal filtered by a mixed-
phase channel has been considered for the simulations.
Performance of the algorithm averaged over a set of
realizations in a wide range of SNR is depicted in figure 1.
(There term $N_{av}$ in eq.9 has been chosen as 1). It displays
the inverse of the squared tentative variance parameter (a
Gaussian model has been assumed for the noise) versus
time which can also be interpreted as the evolution of the
EbNo before the decision device. It appears that the range
of acquisition falls in the order of one to two thousand
symbols depending on the input SNR. The step size for the
adaptation of the coefficients has been chosen as 0.005
while that of the tentative variance is 0.05. In general
the tentative variance accepts quite high adaptation rates
without danger of divergence. The acquisition curve of the
tentative variance is usually very steep during convergence.
The smoothed behaviour observed in figure 1 is due rather
to statistical averaging of several realizations (the elbow
does not always occur at the same instant) than to the actual
behaviour of a single realization. A realization of the
coefficients error with respect to the optimum equalizer is
depicted in figure 3.

6. Conclusions

An algorithm capable of forcing a signal of an a priori
known pdf at the output of an adaptive system has been
presented. It has also been shown that it is able to
minimise (unknown) noise variance at the output. The rate
of blind acquisition is rather fast as compared with other
methods. Some research is currently being done to solve the
capture problem of the algorithm to distributions that are
similar to the a priori template.

Figure 1. Evolution of the inverse of the squared tentative
noise variance before detection or estimated EbNo for
several input SNR's (dB): 16.90, 13.37, 10.88, 8.94, 7.35,
6.02, 4.86, 3.83, 2.92.

Figure 2.: Evolution of the in-phase channel for an input
SNR of 4.86 dB (see figure 1 for output EbNo).

Figure 3. Evolution of the equalizer coefficients error
with respect to the optimum filter at an input SNR of 4.86
dB (see figure 1 for output EbNo).

References

[1] Josep Sala, Gregori Vázquez, "Global Beamforming and
Demodulation for Linearly Modulated Signals", Seventh SP
Workshop on Statistical Signal and Array Processing.
Quebec, 1994.
d'ordres supérieurs pour le traitement du signal". Fascicule
III, Les Houches, collège pré-doctoral de physique. August
[3] Richard E. Blahut, "Principles and Practice of
International Editions. 1991
Stability and convergence analysis of the Constant Modulus Algorithm. Comments on finite equalization schemes and the stability of the Normalized CMA

Katia HILAL* and Pierre DUHAMEL**

* Université de Versailles St Quentin en Yvelines, France
** Telecom-Paris 16, rue Barrault 75013 Paris France
E-mail: duhamel@sig.ens.fr, Fax: +331 45887885

Abstract. This paper puts together several results concerning the convergence properties of the Constant Modulus Algorithm (CMA). First, the steady states of the criterion are recalled. Then, we concentrate on the steady states of the algorithm itself. A powerful method (Ordinary Differential Equation - ODE-) for studying adaptive algorithms is compared to a time-domain characterization. Comparison of these results provides several useful informations on the steady states of the algorithm. The stability of the recently proposed normalized version of the CMA is also studied.

1. Introduction

The Constant Modulus Algorithm (CMA), introduced by Godard in 1980 [2] is one of the most popular adaptive blind equalization algorithms. It involves a non-linear adaptation rule, which makes its convergence analysis difficult. Treichler et al studied its performance [1] under various approximations. Later, the global convergence of the algorithm was proven for an infinite length equalizer [4], while Ding et al [5] showed the existence of undesirable equilibria in the finite equalizer case.

This paper first recalls the classical characterization of the CMA steady states. Section 3 gives an explicit solution of the Ordinary Differential Equation (ODE) of the CMA in the real valued case, showing that its convergence rate depends only on the initialization and on the input signal statistics. Then, section 4 provides a time-domain characterization of the CMA steady states which enlightens the consequences of the hypothesis made over the step-size or the equalizer length in the classical studies. Section 5 applies these results to the Normalized-CMA (NCMA) [6].

Throughout the paper, a slanted capital letter denotes a vector $X$, a bold face letter denotes a matrix $R$, $X^*$ denotes the conjugate of $X$, $X^H$ and $X^t$ denote its transpose and transpose conjugate.

2. The steady states of the Constant Modulus criterion

The CMA is a stochastic gradient descent algorithm derived from the following cost function:

$$ J = E \{ (|y_n|^2 - r)^2 \} $$

where $y_n = X_n^H H_n$ is the filter output at time $n$, $X$ is the input signal vector, $H$ is the equalization vector and $r$ is a positive constant which, without loss of generality, is set to "1". The corresponding adaptation equation is usually written in one of the two forms:

$$ H_{n+1} = H_n - \mu X_n^H (|y_n|^2 - 1) $$

where $g[y] = y |y|^2$. Equation (1) has the form of a "Bussgang" algorithm [9]. Classically, define the stationary states as those filters $H$ for which the gradient of the criterion is zero. For the CMA, this reads:

$$ E \{ X^* H_s \} = R H_s $$

or, more explicitly

$$ E \{ X^* X^t H_s (X^t H_s X^t) \} = R H_s $$

where $R$ is the input signal autocorrelation matrix, statistical expectation being taken over different realizations of vector $X$. Of course, these stationary states do not characterize stable states (which form a subset of the stationary ones), nor the states which remain unchanged when running the adaptive algorithm (due to the stochastic gradient procedure).

Note also that the above characterization depends only on the criterion, and not on the specific algorithm. Thus this characterization of the stationary states does not reflect the precise behavior of the algorithm. Equation (2) is similar to the Wiener solution of an MMSE adaptive filter [9]. In what follows, we concentrate on the algorithm itself rather than on the criterion.

3. Ordinary Differential Equation (ODE) of the CMA

The ODE method was first introduced by Ljung [11] and developed later by Kushner [13] and Benveniste [12]. It relates the motion of the coefficient estimates of the adaptive algorithm

$$ H_{n+1} = H_n + \mu F(H_n, X_n), $$

to the behavior of an unforced deterministic ODE (when it exists):

$$ \frac{\partial H}{\partial t} = f(H) \quad \text{where} $$
\[ f(H) = E\{F(H,X)\} \text{ for fixed } H \]

This ODE provides a continuous-time evolution of the filter coefficients, and the correspondence between the sampled and the continuous one is given by \( t = n \times \mu \), which holds for small \( \mu \) and under stationarity conditions. Hence, stationary states of the equalizer are those for which \( f(H_0) = 0 \). Stable states are the stable solutions of the ODE, i.e., those for which the real part of the eigenvalues of \( G_s = \text{grad}_H(f)H_0 \) is negative.

### 3.1. Derivation of the ODE

Our study is undertaken for real-valued variables, since the complex case is much more intricate. Despite this restriction, we enlighten and explain the origin of several problems that were encountered in the literature concerning the CMA. The solution of the ODE is first derived for \( L = 1 \), the general result for a length \( L \) filter being given afterwards. Following the above definitions, the ODE of the real-valued CMA reads:

\[ \frac{\partial H}{\partial t} = RH - E\{XX'HX'XH'X\} \]  

For \( L = 1 \), the above equation is easily solved by setting \( z = \frac{1}{H_0} \), yielding the solution:

\[ \frac{1}{H_0^2} \frac{E\{(z^2)\}}{E\{(z)\}} = \left( \frac{1}{H_0^2} - \frac{E\{(z^2)\}}{E\{(z)\}} \right) \exp(-2E\{(z^2)\}t) \]

The case \( L > 1 \) is not that easily obtained. The solution is provided below without demonstration, and checked afterwards.

\[ I = R^{-1}E\{XX'HX'H\} + \exp(-2Rt)AE\{XX'HX'H\} \]

Matrix \( A \) is obtained by imposing some initial condition \( H_0 \) at time \( t = 0 \) in the above equation:

\[ A = E\{XX'H_0X'H_0\}^{-1} - R^{-1} \]

An expression which clearly involves only \( H_0 \) and the 4th and 2nd order statistics of the signal.

Although difficult to derive, the above solution is easily checked by multiplying each side by \( H \) and then taking its derivative:

\[ \frac{\partial H}{\partial t} = 3R^{-1}E\{XX'HX'H\} \frac{\partial H}{\partial t} + 3E\{XX'HX'H\} \frac{\partial H}{\partial t} - 2E\{XX'HX'H\} \frac{\partial H}{\partial t} \]

The first line of the above equation can be transformed according to (5), yielding:

\[ \frac{\partial H}{\partial t} = RH - E\{XX'HX'H\} \]

which is actually the ODE of the CMA. Observation of (5) leads to the following comments:

### 3.2. Steady states of the CMA

When \( t \to \infty \), equation (5) tends to a subset of (2), characterized by:

\[ I = R^{-1}E\{XX'H_0X'H_0\} \]

Now check the stability of these stationary points, by considering the quantity \( G_s = \text{grad}_H(f)H_0 \):

\[ G_s = -R^{-1}E\{XX'HX'H\} \]

Since \( R \) is the input signal autocorrelation matrix, its eigenvalues are nonnegative. Hence the filters that satisfy (7) are stable solutions.

The ODE provides another result concerning the convergence rate towards these steady states: the convergence rate does not depend on the actual value of the equalizer, and is governed mainly by the input signal autocorrelation matrix \( R \) as for the LMS algorithm.

The effect of the intersymbol interference (ISI) can thus be explained: Convergence to (2) depends on the conditioning of matrix \( R \) which is only related to the ISI involved in the received sequence when the emitted symbols are uncorrelated (which is the case in many practical situations).

This section showed that, under classical assumptions (small \( \mu \) and stationarity conditions), the CMA converges, with a given convergence rate to any stable solution of coefficient satisfying (7). However, the practical use of the algorithm showed cases of misbehavior and of local minima, even in cases where the above conditions were met. Hence, as a conclusion, these problems seem to be due not to the actual algorithm, but to the specific criterion which is used. More specifically, these problems are not due to the truncation of the equalizer. However, the presence of noise, the finite length of the equalizer as well as the use of a large step may result in a deviation of the stable states of the actual algorithm from the ones of the criterion.

The next section develops a deterministic characterization of the steady states (which does not rely on an infinite length assumption) providing further explanation of this result.

### 4. The time domain equation

In [14], various adaptive algorithms were reformulated as a recursive structure. The CMA recursive reformulation is easily obtained by rewriting the algorithm equations at time \( n \) as:

\[ H_n = H_0 - \sum_{i=0}^{n-1} \mu_i X_i e_i \]

\[ y_n = X_n' H_n = X_n' H_0 - \sum_{i=0}^{n-1} \mu_i X_i X_i' e_i \]

where \( e_i = y_i - y_i \) is the algorithm “error” at time \( i \).

This structure allows a study of the actual average behavior of the algorithm around a steady state \( H_n \) to be performed. The steady states are now defined as the filters for which the time average of the algorithm adaptations \( \Delta H_n = H_{n+1} - H_n \) is null, a condition which is in fact related to the convergence of the mean of the equalizer. In other terms:

\[ \lim_{n \to \infty} \left\{ \frac{1}{n} \sum_{i=0}^{n-1} \Delta H_i \right\} = \lim_{n \to \infty} \left\{ \frac{1}{n} (H_n - H_s) \right\} = 0 \]

For a fixed step-size \( \mu \), a time-domain characterization of steady states is easily obtained by substituting (8) in the above equation:

\[ X_n' B_n X_n' H_n = X_n' B_n G_n \text{ when } n \to \infty \]
where

\[
X_n = [X_0, X_1, \ldots, X_n]
\]

\[
G_n = [g(y_0), g(y_1), \ldots, g(y_n)]
\]

\[
B = (I + \mu S_n)^{-1}
\]

\[
S_n = \begin{bmatrix}
0 & 0 & \cdots & 0 \\
X_1^* X_1 & 0 & \cdots & 0 \\
X_1^* X_1 & X_1^* X_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
X_1^* X_1 & X_1^* X_2 & \cdots & X_1^* X_{n-1}
\end{bmatrix}
\]

(11)

B is a triangular matrix with 1's on the diagonal, the other terms being related to estimates of the input signal correlation coefficients and to the step-size \( \mu \). Note that internal dimensions of the involved matrices tend to infinity with \( n \). Since \( X_1^* B_n X_1 \) is invertible and full rank, equation (10) gives a general characterization of the steady states of the CMA.

Eq. (10) is easily seen to reduce to a temporal estimation of the classical steady states characterization (2) in the following cases:

1. \( B_n \approx I \) which means that either \( \mu \) is very small or the input signal correlation estimates are negligible.

2. when the equalization operation reduces to the identification of a finite length filter that satisfies \( G_n = X_1^* H_n \) in which case (10) is obviously true. This corresponds to cases where \( y_n - g(y_n) \) is always zero, i.e. cases where perfect equalization with a filter of the same length as \( H_n \) is possible and the emitted sequence has indeed constant modulus. Since perfect equalization cannot be obtained for non minimum phase channels [16], this condition will never be met, except in academic situations.

Item number 1 deserves a few comments: it shows that the classical assumptions which are usually done for studying analytically the adaptive behavior of algorithms are somewhat linked: Preliminary studies relied on the assumption that the input data was uncorrelated. The ODE method did not require explicitly the same condition, but requires \( \mu \) be very small. Item 1 thus shows that these two assumptions are somewhat linked: Both result in \( B_n \approx I \). Hence, the prediction of the steady states must be expected to be somewhat unprecise for very correlated input data (which means strong ISI: the very cases where the equalizers are important) and/or large steps, whatever the prediction method. On the other hand, all classical results for large step-sizes currently available use the independence assumption, which specifies that \( \{X\} \) is an i.i.d. sequence [17].

Hence, since equation (10) reduces to (2) in these two cases, the time-domain characterization (10) seems to be an accurate method of predicting the steady states of the actual algorithm, since it is valid in the cases of truncation of the equalization filter, correlation of the input signal, or large step-size. The results of section (3) are consistent with the new characterization, since the ODE method holds only for small \( \mu \), in which case both results are equivalent.

We now concentrate on the stability conditions of a Normalized version of the CMA (NCMA) [6] which is the CMA counterpart of the Normalized-LMS.

5. Stability analysis of the NCMA

The NCMA [6] was obtained with reference to the step-size that allows the a posteriori error of the CMA to be zero at each time \( n \):

\[
\mu_{n,n}^{opt} = \frac{||y_n||^2 - ||y_n||}{\mathcal{E}_n[||y_n||^2 (||y_n||^2 - 1)]} = \frac{||y_n||}{\mathcal{E}_n[||y_n||^2 (||y_n|| + 1)]}
\]

(12)

where \( \mathcal{E}_n = X_1^* X_1^* \) is the energy of the input signal seen by the filter at time \( n \). Thus, the actual NCMA is:

\[
H_{n+1} = H_n - \alpha \mu_{n,n}^{opt} X_1^* y_n (||y_n||^2 - 1)
\]

\[
= H_n - \frac{\alpha}{\mathcal{E}_n} X_1^* y_n (1 - \frac{||y_n||}{||y_n||})
\]

(13)

where \( \alpha \) is a "normalized" step-size (energy independent), introduced as in the NLMS to be able to attain various tradeoffs between convergence rate and residual error. The study of the stability of the NCMA by the above methods is somewhat difficult due to the precise form of the corresponding adaptation equation. The only result we found was that the step value of the NCMA always lies in the stability domain of the step of the classical CMA. However, this result was not sufficient to ensure stability of the NCMA, since its adaptation step is time-varying.

Hence, we used the technique developed by Macchi et al for the LMS and the Decision-Directed algorithms [7] directly on the NCMA. This method allows to find the conditions on the step-size such that the algorithm remains stable. It consists in separating the contribution of the whole set of iterations to some filter value \( H_n \) in two parts: A first one corresponds to the transient, which should tend to zero, while the second one corresponds to the limit value, and should remain bounded for the adaptive filter to be stable.

The organization of the computation follows the LMS case: equation (13) can be developed as follows:

\[
H_{n+1} = \left(1 - \frac{\alpha}{\mathcal{E}_n} X_1^* X_1^*\right) H_n + \frac{\alpha}{\mathcal{E}_n} X_1^* X_1^* e^{\phi_{n+1}}
\]

\[
= \prod_{i=0}^{n} \left(1 - \frac{\alpha}{\mathcal{E}_i} X_1^* X_1^*\right) H_0 + \alpha \sum_{i=0}^{n} \left(\prod_{k=i}^{n-1} \left(1 - \frac{\alpha}{\mathcal{E}_k} X_1^* X_1^*\right)\right) X_1^* e^{\phi_i}
\]

(14)

where \( \phi_i \) denotes the phase of the output signal at time \( i \).

The algorithm is convergent if the first term of (14) tends to zero with \( n \) (extinction of the transients) and if its second term is bounded.

Just like in [7], we assume that the input signal has finite memory \( M \) i.e. that its correlation coefficients vanish for all time shifts larger than \( M \). In an equalization application, this assumption is correct for channels with finite length base-band models and uncorrelated transmitted symbols. The precise computations are very similar to the ones found in [7]. The first condition (extinction of the transients) reads for the NCMA:

\[
\alpha < \frac{L}{M}
\]

with \( M \) the memory of the signal and \( L \) the filter length. It is easily checked that the same condition ensures the second term of (14) to remain bounded (since we consider that the input signal also is). Note that, in practical cases, \( L \geq M \) (otherwise the equalization cannot be completed exactly) and that \( \alpha \leq 1 \) is usually sufficient to ensure stability.
6. Conclusion

This paper reviews several previous results on the stability of CMA and its steady states and then provides a characterization of its convergence rate by means of ODE techniques. Then, a time-domain characterization of the steady state values of the algorithm is provided, without resorting to the condition that the equalizer has infinite length, vanishing step-size or i.i.d. input sequence. The equivalence of these two results is then discussed. (It is shown that they are equivalent under conditions which are seldom met in actual situations). Finally, stability conditions of the recently proposed NCMA are provided.

Acknowledgments

This work was undertaken while both authors were with the Centre National d’Etudes des Télécommunications Issy les Moulineaux, France.

References


A Fast Adaptive Algorithm for Multichannel System Identification - Application to DFE

Sergios THEODORIDIS, George V. MOUSTAKIDES, Kostas BERBERIDIS

Dept. of Computer Engineering, Univ. of Patras, 26500 Patras, GREECE, and
Computer Technology Institute (CTI), P.O. Box 1122, 26110 Patras, GREECE.

Abstract. The major computational contribution in fast transversal adaptive algorithms comes from updating the associated forward and backward predictors, of the input time series, which are implicitly assumed to be of the same order as the unknown system. However it is quite common in practice to extract all the predictable information of the input series with predictors of much lower order. This paper presents a method that incorporates an apriori information about the predictors’ orders, assumed to be less or equal to that of the multichannel system to be identified. This results to a new class of algorithms trading off performance to computational complexity. The general multichannel problem with different number of taps per channel is treated. The applicability of the proposed algorithm to Decision Feedback Equalization is demonstrated.

1. Introduction

The design of efficient adaptive algorithms for multichannel system identification is of major importance in a wide range of disciplines such as seismic signal processing, wide band adaptive array design, digital communications, control, etc. [1]. The major computational contribution in fast transversal adaptive algorithms comes from updating the associated forward and backward predictors of the input time series, which are implicitly assumed to be of the same order as that of the unknown system [1]. However it is quite common in practice to extract all the predictable information of the input time series with predictors of much lower order. This idea was successfully exploited in the recently suggested class of Fast Newton Transversal Filters [4] where the prediction part was assumed to be of a lower order than that of the filtering part.

In general the number of taps in a multichannel FIR system need not be the same for the different channels involved (i.e. feedforward and feedback section). This is for instance the case in Decision Feedback Equalization which is the application of interest in this paper. Although such a case could be dealt by using the same number of taps for all channels, expecting the algorithm to zero the extra taps, this leads to an unnecessary computational increase and also may affect the accuracy of the obtained solution. In this paper the more general case of different number of taps per channel is treated. Simulation results verify that performance is traded off against complexity, by varying the predictor’s order.

2. Formulation of the Problem

Let us assume two input signals $x_1(n), x_2(n)$ which are combined by the linear system

$$\hat{d}(n) = -\sum_{i=1}^{m} c_{1i} x_1(n-i+1) - \sum_{j=1}^{l} c_{2j} x_2(n-j+1)$$

A recursive solution to the problem of estimating the unknown system’s parameters, based on input - desired output samples is given by the well-known Stochastic Newton method [1]

$$c_{ml}(n) = c_{ml}(n-1) + w_{ml}(n)[d(n) + c_{ml}^i(n-1)x_{ml}(n) - 1]$$

where $c_{ml} = [c_{11}, c_{12}, \ldots c_{1m}, c_{21}, c_{22}, \ldots c_{2l}]^T$ is the parameter vector and $x_{ml}(n) = [x_1(n), x_1(n-1), \ldots x_1(n-m+1), x_2(n), x_2(n-1), \ldots x_2(n-l+1)]^T$. The gain vector is given by $w_{ml}(n) = -\gamma n R_{ml}^{-1}(n)x_{ml}(n)$ where $R_{ml}(n)$ is an estimate at time $n$ of the input data correlation matrix and $\gamma(n)$ is a properly chosen positive gain sequence. In this paper, extending the idea introduced in [4] for the single channel case, this estimate of $R_{ml}(n)$ is produced by extrapolating the sample correlation matrix of a lower order $R_{rs}(n)$, where $r, s$ denote the prediction orders of the two input sequences respectively (with $r \leq m$ and $s \leq l$). Note that the multichannel case is not treated here as a straightforward generalization of the single channel case, having matrices in the place of scalars. The multichannel order evolution required by the algorithm, is achieved in steps involving each channel separately and leads to an algorithm involving scalar operations only.

Let us assume that the matrix $R_{rs}(n)$ is known and we seek to make an estimate of $R_{r+1,s+1}(n)$. If the latter matrix is partitioned as

$$R_{r+1,s+1}(n) \equiv \begin{pmatrix} K & L \\ L' & M \end{pmatrix}$$

then the unknown elements in the above matrix are, the upper right element of $K$ (denoted as $p_{12}^{rs}(n)$), the upper right element of $M$ (denoted as $p_{22}^{rs}(n)$), the upper right element of $L$ (denoted as $p_{rs}(n)$), and the lower left element of $L$ (denoted as $\tilde{p}_{rs}(n)$).

728
The above elements are computed from respective prediction problems following a saddle point approach. The involved prediction problems are defined as follows, 

a) Given the input samples \(x^1(n-1), \ldots, x^1(n-r), x^2(n-1), \ldots, x^2(n-s+1)\) predict \(x^1(n)\). The corresponding predictor, prediction error and error power are denoted as \(A^2_{s-1}(n), c^2_{s-1}(n)\) and \(\alpha^2_{s-1}(n)\) respectively.

b) Given \(x^1(n-1), \ldots, x^1(n-r), x^2(n), \ldots, x^2(n-s+1)\), predict \(x^1(n)\). The respective prediction quantities are denoted as \(A^2_{r-1}(n), c^2_{r-1}(n)\) and \(\alpha^2_{r-1}(n)\).

c) Given \(x^1(n), \ldots, x^1(n-r+1), x^2(n), \ldots, x^2(n-s+1)\), predict \(x^1(n-r)\). The respective prediction quantities are denoted as \(B^2_{r-1}(n), c^2_{r-1}(n)\) and \(\alpha^2_{r-1}(n)\).

d) Given \(x^1(n), \ldots, x^1(n-r), x^2(n), \ldots, x^2(n-s+1)\), predict \(x^2(n-s)\). The respective prediction quantities are denoted as \(B^2_{s-1}(n), c^2_{s-1}(n)\) and \(\alpha^2_{s-1}(n)\).

The minmax part of the saddle point approach is equivalent with making the minimum error power to be maximum with respect to the corresponding unknown element of matrix \(R_{r+s+1}(n)\). Thus, for example, the unknown element \(\tilde{p}_{-r}(n)\) will be estimated as the result of the predictor \(A^2_{s-1}(n)\) to have the worst possible prediction error power. Using the partitionings of Table 1 (where the definitions of the involved permutations and partitionings are given) and the well-known matrix inversion lemma for partitioned matrices [1], it can be shown after some algebra [5] that \(A^2_{s-1}(n)\) satisfies the following order update

\[
A^2_{s-1}(n) = S_{r-1} \begin{pmatrix} A^2_{s-1}(n-1) \\ 0 \end{pmatrix}
\]

(4)

and the estimated value of the unknown element is given by

\[
\tilde{p}_{-r}(n) = -B^2_{r-1}(n) E\{x_{r-1},s-1(n-1) x^2(n)\}
\]

where \(E\{\cdot\}\) is an estimate of \(E\{\cdot\}\). As we shall see the actual computed estimates of the autocorrelation elements are of no interest to us. It is the equivalent predictor's order update which will be exploited. Similar order updates can be obtained for the rest of the predictors, that is, \(A^2_{r-1}(n) = (\tilde{A}^2_{r-1}(n) \ 0)^T\), \(B^2_{r+1}(n) = (0 \ \tilde{B}^2_{r+1}(n))^T\), \(B^2_{s-1}(n) = (0 \ \tilde{B}^2_{s-1}(n))^T\), where \(\tilde{B}^2_{r-1}(n)\) and \(\tilde{B}^2_{s-1}(n)\) are intermediate auxiliary vectors computed via the relations, \(\tilde{B}^2_{r-1}(n) = T_{r-1} \ (0 \ \tilde{B}^2_{r+1}(n-1) n-1)^T\) and \(\tilde{B}^2_{s-1}(n) = T_{s-1} \ (0 \ \tilde{B}^2_{s-1}(n-1) n-1)^T\).

Continuing the above procedure, the matrix \(R_{m+n}(n)\) can be recursively estimated from \(R_{r+s}(n)\). Note that, with our method, this is possible only if \(m - r = l - s\), imposing a restriction on the order of the matrix \(R_{r+s}(n)\). It can also be shown that the matrix extrapolated in the above way remains positive definite [5]. Furthermore its inverse, if viewed as a 2 \times 2 block matrix, results in banded blocks. Specifically, in the places of the unknown elements of the extrapolated matrix, we have zeros in its inverse \(R_{m+n}(n)\).

3. The Multichannel FNTF Algorithm

Let us write the recursion of eq. (2) in the LS a-posteriori error formulation, i.e.

\[
c_{m+l}(n) = c_{m+l}(n-1) + w_{m+l}(n) e_{m+l}(n)
\]

(5)

where

\[
c_{m+l}(n) = d(n) + c_{m+l}(n) x_{m+l}(n)
\]

(6)

\[
w_{m+l}(n) = -\lambda^{-1} R^{-1}_{m+l}(n-1) x_{m+l}(n)
\]

(7)

We assume that \(R_{m+l}(n-1)\) is a scaled estimate of the correlation matrix extrapolated from a lower order covariance matrix \(R_{r+s+1}(n-1)\) which is computed as a least squares estimate. It is well known that the essence in deriving a fast algorithm is to achieve a fast computation of the Kalman gain vector (i.e. vector \(w_{m+l}(n)\) in our case). Notice that a constant weighting sequence \(\lambda\) has been adopted to allow for slow time variation tracking. As it has already been pointed out in the previous section the extrapolated matrices will not be explicitly involved. It is their related prediction (state space) parameters and their interrelation which will be accounted for, as it is always the case with fast algorithms.

Let us now define the partitioning \(A^2_{m+l}(n) = (\tilde{A}^2_{r+1}(n) \ \tilde{A}^2_{s+1}(n))^T\) for the predictor \(A^2_{m+l}(n)\) and similarly for the resting predictors \(B^2_{r+1}(n)\) and \(B^2_{s+1}(n)\). Applying successively the updating procedure as in eq. (4) and using the definitions of Table 1 we finally obtain

\[
T_{m+l+1} \begin{pmatrix} A^2_{m+l}(n+1) \\ 0_{m-r} \end{pmatrix} = \begin{pmatrix} a_{m+l}^2(n) \\ 0_{m-r} \end{pmatrix}
\]

(8)

For the rest predictors it can be shown, in a similar way, that

\[
S_{m+l} \begin{pmatrix} B^2_{m+l}(n) \\ 0_{l-s} \end{pmatrix} = \begin{pmatrix} b^2_{m+l}(n) \\ 0_{l-s} \end{pmatrix}
\]

(10)

\[
B^2_{m+l+1}(n+1) = \begin{pmatrix} b^2_{m+l+1}(n) \\ 0_{l-s} \end{pmatrix}
\]

(11)
Starting from the respective definitions and using the above updating formulae the following relations are obtained for the involved powers

\[
\begin{align*}
\varepsilon_{m}^{2}(n) &= \varepsilon_{r}^{2}(n) \\
\varepsilon_{m+1}^{2}(n) &= \varepsilon_{r+1}^{2}(n) \\
\varepsilon_{m}^{4}(n) &= \varepsilon_{r}^{4}(n - m + r) \\
\varepsilon_{m}^{6}(n) &= \varepsilon_{r}^{6}(n - m + r)
\end{align*}
\]

and similar relations can be obtained for the involved powers. Now using (8)-(12) and following similar steps as in the two-channel staircase algorithm of [3] the corresponding two-channel FNTF algorithm of Table 2 results. It is readily observed that the computations associated with the filtering part contribute to the complexity in proportion to the systems order \( m, l \). The contribution to the complexity of the prediction part is linearly dependent on the predictors orders \( r, s \). Specifically, the overall complexity of the algorithm of Table 2 is \( 10(r + s) + 2(m + l) \) MADS per time recursion, while the respective complexity of the algorithm of [3] is \( 12(m + l) \) MADS. Recently an exact block version of the algorithm has been developed reducing complexity to a portion of LMS per input sample [6].

4. Application to DFE

The above derived algorithm is directly applicable to Decision Feedback Equalization. To show its applicability and its performance we conducted the following experiment. A binary pseudorandom sequence was used as the bit information sequence sent to a channel which introduced intersymbol interference. The channel was a linear phase FIR filter with an impulse response spreading over 15 successive bits. A 20dB (SNR) white Gaussian noise was added at the output of the channel. The introduced distortion was rather severe due to the large dynamic range and the deep nulls which were present in the frequency response of the channel. Equalization of channels with deep nulls suggest the use of Decision Feedback Equalizers (DFE). A typical DFE consists of the feedforward anticausal part and the feedback causal part, and its output is given by

\[
\tilde{z}(t) = \sum_{i=1}^{N_1} c_{1}^{i} y(t + N - i) + \sum_{j=1}^{N_2} c_{2}^{j} \tilde{z}(t - j + 1) \tag{13}
\]

where \( \{y(t)\} \) is the received sequence and \( \{\tilde{z}(t)\} \) is a sequence consisted of the correct symbols in the training mode and the detected symbols in the decision directed mode respectively. A symbol rate decision feedback equalizer is typically a two channel system identification task. The inputs in the two channels are the sequences \( \{y(t)\} \) and \( \{\tilde{z}(t)\} \) respectively. The equalizer parameters \( c_{1}^{i} \) and \( c_{2}^{j} \) are estimated so that the error \( \tilde{z}(t) - z(t) \) is minimized. The equalizer used by this experiment consisted of 20 feedforward and 20 feedback taps. Five curves are shown in Figure 1. Curve 1 (the lower one) corresponds to the two-channel RLS algorithm. Curve 2 (dashed line) corresponds to the MFNTF algorithm with two-channel predictors of orders 15,15. As we can see an almost negligible degradation in performance results at a computational saving of the order of 25%. Curve 3 (dotted line) corresponds to the MFNTF algorithm with predictors of orders 10,10. Curve 4 (dashed-dotted line) corresponds to the MFNTF algorithm with predictors of orders 5,5. The latter has converged at about 2000 samples. In all the above cases the forgetting factor \( \lambda \) was taken equal to 0.99. The top curve corresponds to the Normalized LMS which at about 4000 samples (not shown in the figures) converges to the same misadjustment level as that of Curves 3 and 4. Thus we have demonstrated that the use of the multichannel Fast Newton algorithm provides the means of trading off performance with computational complexity having RLS at one end and NLMS at the other.

References


| \( x_{ij}(n) \) | \( = ( x_{i}(n) \tilde{z}_{i-j}(n) )' \) |
| \( x_{ij}(n) \) | \( = ( x_{i+j}(n) x_{i+j}(n) )' \) |
| \( \tilde{z}_{ij}(n) \) | \( = T_{ij}( x_{i+j}(n) x_{i+j}(n) )' \) |
| \( x_{ij}(n) \) | \( = S_{ij}( x_{i+j}(n) x_{i+j}(n) )' \) |

**Table 1:** Partitions and Permutations
**TABLE 2: The MFNTF Algorithm**

Define:

\[
\begin{bmatrix}
    p_{r+1}^{11}(n) \\
p_{r+1}^{22}(n)
\end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix}
    a_r^{11}(n) \\
a_r^{22}(n)
\end{bmatrix}
\begin{bmatrix}
    1 \\
1
\end{bmatrix},
\begin{bmatrix}
p_{r+1}^{21}(n) \\
p_{r+1}^{22}(n)
\end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix}
    a_r^{21}(n) \\
a_r^{22}(n)
\end{bmatrix}
\begin{bmatrix}
    1 \\
1
\end{bmatrix}
\]

\[
\begin{bmatrix}
    q_{r+1}^{11}(n) \\
q_{r+1}^{22}(n)
\end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix}
    a_r^{11}(n) \\
a_r^{22}(n)
\end{bmatrix}
\begin{bmatrix}
b_r^{11}(n) \\
b_r^{22}(n)
\end{bmatrix},
\begin{bmatrix}
q_{r+1}^{11}(n) \\
q_{r+1}^{22}(n)
\end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix}
a_r^{11}(n) \\
a_r^{22}(n)
\end{bmatrix}
\begin{bmatrix}
b_r^{11}(n) \\
b_r^{22}(n)
\end{bmatrix}
\]

\[k = n - m + r, \quad [p]_k \text{ denotes the } k\text{-th element of the vector } p\]

- Available from the previous recursion of the MFNTF: \(w_{ml}(n-1), \alpha_{ml}(n-1)\)
- Available from any LS multichannel algorithm: \(\theta_{r+1}^{11}(n), \theta_{r+1}^{22}(n), \theta_{r+1}^{22}(k), \theta_{r+1}^{22}(k)\)

### Prediction Part

\[
\hat{w}_{m+1}(n) = T_{m+1} \begin{bmatrix}
    0 \\
w_{ml}(n-1)
\end{bmatrix}^t - T_{m+1} \begin{bmatrix}
p_{r+1}^{11}(n) \\
p_{r+1}^{22}(n)
\end{bmatrix} \begin{bmatrix}
    0 \\
0
\end{bmatrix}^t
\]

\[
w_{m+1}(n) = \hat{w}_{m+1}(n) - \begin{bmatrix}
    0 \\
p_{r+1}^{11}(n)
\end{bmatrix} \begin{bmatrix}
    0 \\
0
\end{bmatrix}^t
\]

\[
\begin{bmatrix}
w_{m+1}(n) \\
0
\end{bmatrix}^t = \hat{w}_{m+1}(n) + \begin{bmatrix}
    0 \\
w_{ml}(n)
\end{bmatrix} \begin{bmatrix}
    0 \\
0
\end{bmatrix}^t
\]

\[
\begin{bmatrix}
w_{m+1}(n) \\
0
\end{bmatrix}^t = \theta_{m+1}^{22}(n) + \begin{bmatrix}
    0 \\
w_{ml}(n)
\end{bmatrix} \begin{bmatrix}
    0 \\
0
\end{bmatrix}^t
\]

\[
\alpha_{ml}(n) = \alpha_{ml}(n-1) + \begin{bmatrix}
    p_{r+1}^{11}(n) \\
p_{r+1}^{22}(n)
\end{bmatrix} \begin{bmatrix}
    \theta_{r+1}^{11}(n) \\
\theta_{r+1}^{22}(k)
\end{bmatrix}
\]

### Filtering Part

\[
\epsilon_{ml}(n) = d(n) + \begin{bmatrix}
    p_{r+1}^{11}(n) \\
p_{r+1}^{22}(n)
\end{bmatrix} \begin{bmatrix}
    \theta_{r+1}^{11}(n) \\
\theta_{r+1}^{22}(k)
\end{bmatrix}
\]

\[
\epsilon_{ml}(n) = \epsilon_{ml}(n)/\alpha_{ml}(n)
\]

\[
\epsilon_{ml}(n) = \epsilon_{ml}(n-1) + w_{ml}(n)\epsilon_{ml}(n)
\]

**FIGURE 1**

![Graph](image-url)

---

731
Genetic Algorithms in the Continuous Space for Recursive Adaptive Filter Design

Jean-Marc VESIN
Signal Processing Laboratory
Swiss Federal Institute of Technology
1015 Lausanne, Switzerland
Tel/Fax: +41 21 693 3996 / 693 7600, E-Mail: vesin@lissg3.epfl.ch

Abstract This paper presents the potential interest of a new type of genetic algorithm, called the genetic algorithm in the continuous space, for signal processing applications. The central idea is that the parameter vector can be used straightforwardly instead of its binarized concatenated representation as in classical genetic algorithms. This allows to have more control on the evolution of the parameters and to take profit of prior information on their range of admissible values. This approach is illustrated on a problem of all-pole system identification, for which significantly better performance is achieved.

1. Introduction

Genetic algorithms (GA) [1, 2] were introduced by J. Holland as a class of search procedures. They consist in an imitation of Darwinian evolution on a population of initially randomly selected feasible solutions. Starting from this initial population, successive generations are obtained through crossover (recombination) and mutation in keeping with principles such as "survival of the fittest", i.e., the best solutions are allowed to reproduce while the worst ones are eliminated. This process is repeated until acceptable solutions appear in the population. GA have been applied with success to problems such as the Traveling Salesman one [3], image enhancement [4] and system identification [5] to name but a few.

However, in all these applications, the "individuals" in a population are described by "chromosomes" consisting in bit strings, which makes the solution space discrete. These chromosomes are built by concatenating discretized values of the relevant parameters, and the operations of recombination and mutation correspond respectively to random shuffling of the bit strings of two individuals and modification of randomly chosen bits. Thus, apart from the discrete nature of the solution obtained, another disadvantage is the lack of control upon the operations of recombination and mutation with respect to possible inherent constraints on the solution.

The purpose of this paper is to advocate the use (at least for signal processing applications) of genetic algorithms in the continuous space (GACS). GACS consist in considering as chromosomes the classical parameter vectors and in applying mutation and recombination operations taking this structure into account (analysis of a simplified GACS incorporating only mutation can be found in [6]). The discretization of the solution is obviously removed, and proper choice of the mutation and recombination procedures can assure that the solution obtained will indeed satisfy possible constraints. In our view, this evolution of the concept of GA is similar to what happened to neural network (NN) techniques: Earlier development in this field was initially slowed by a strong adherence to the (supposed) structure of biological neurons. Once NN were recognized as new mathematical tools for multi-dimensional function approximation, new structures with no particular biological justification could emerge. In the case of GA, use of bit strings as chromosomes in association with the above described mutation and recombination schemes tends to be too faithful to the DNA-based biological model. GACS constitute a first step towards the acceptance of GA as ordinary optimization schemes. In order to illustrate the potential of GACS, an application to the same problem as the one described in [7], namely adaptive IIR filter design, is presented in this paper. First, conventional GA and GACS are briefly described. Then the specific application studied is presented, as well as the GACS developed for it. Simulation results are provided and extended to the IIR lattice filter case. Finally, possible improvements and potential applications of GACS are reviewed in the conclusion.

2. GA and GACS

2.1 Classical genetic algorithms

To solve an optimization problem with a genetic algorithm, one first builds a chromosomal representation of the parameters by quantizing them and appending their binary coded quantized values so as to obtain a string of 0's and 1's. An initial population of such chromosomes (individuals) is randomly created and the algorithm forms successive generations through the evaluation of each chromosome with respect to a given criterion and the action of three basic operators: reproduction, crossover, and mutation. Reproduction consists in copying (reproducing) the best chromosomes, i.e. the most highly rated ones. Crossover allows to mix the characteristics of two chromosomes a and b. In its simplest and most frequently used form, it consists in selecting at random a position on the string and build a new chromosome by taking from a the bits left of the position and concatenating to them the bits of b right of the position. For instance, if one has:

\[
a = 110010001 \\
b = 011001100
\]

and position 5 is chosen, then the resulting chromosome will be:

\[
c = 110011100
\]

Mutation consists in switching some randomly chosen bits of a chromosome. The mutation rate, that is, the probability
that a bit will be switched, is an important parameter in any GA. The GA differ in the more or less sophisticated rules used to select which chromosomes of a generation will be reproduced, subject to crossover or mutation [2].

2.2 Genetic algorithms in the continuous space
As already stated, we propose GACS as an alternative to conventional GA for some signal processing applications. The distinctive feature of GACS is that the parameter set is not coded into a bit string, which implies the use of modified crossover and mutation operators. The only reference known to us on GACS is [6], in which a convergence analysis is performed upon a simplified situation (mutation only).

With GACS, undesirable effects due to parameter quantification obviously disappear, and correspondence between chromosome components and the problem under study is not lost, since parameters conserve their individuality. In fact, in a lot of practical situations some parameters or parameter subsets are constrained to remain in a bounded and/or convex region. The central idea is that mutation and crossover operators can be tailored to these constraints, so that viable chromosomes "give birth" to viable chromosomes. A natural choice for mutation seems to add a small random number to all parameters values constituting a chromosome. One may check whether the new chromosome parameter values satisfy the constraints or not, and accept or reject this mutation. Crossover can easily be arranged to satisfy boundness constraints by using convexity. Quite naturally, the new chromosome has only to be located "somewhere in-between" its parents. In this way, natural selection plays its role since highly rated viable individuals will produce highly rated viable individuals as long as the evaluation function is smooth enough.

3. Description of the problem
For purposes of comparison with conventional GA, we consider exactly the same application as in [7], i.e., the recursive identification of a two-pole system. This framework is illustrated in figure 1.

![Figure 1. Recursive identification of a two-pole system.](image)

It is well known that the mean-squared error $E[e^2]$ is a quadratic function of the filter weights if an adaptive FIR model is used. This feature is at the heart of the ubiquitous LMS algorithm [8]. In the case an adaptive two-pole system is used as described in figure 1, it can be shown [8] that, under an assumption of whiteness for the input signal $x_k$, the mean-squared error surface is a non-quadratic single-minimum function in the adaptive filter coefficients $a_1$ and $a_2$. Moreover, stability of the model requires that the point of coordinates $(a_1, a_2)$ remains in the triangular region defined by:

$$
\begin{cases}
-1 < a_2 < 1 \\
1 - a_1 < a_1 < 1 - a_2
\end{cases}
$$

In fact, monitoring of the stability is one of salient problems in IIR adaptive filtering.

4. Description of the specific GACS
A population will be composed of $N$ chromosomes consisting of couples $(a_1(i), a_2(i))$, $i = 1, ..., N$. The mutation operator is defined by:

$$
\begin{cases}
a_1(d) = a_1(f) + v_1 \\
a_2(d) = a_2(f) + v_2
\end{cases}
$$

where $f$ designates the index of the initial chromosome, $d$ the index of the resulting one, and $v_1, v_2$, two Gaussian random variables with zero mean and variance $s^2$. The new chromosome can be accepted or not, depending whether it satisfies eq.(1) or not. The obvious convexity of the triangular stability region induces to consider a crossover defined by:

$$
\begin{cases}
a_1(d) = g(a(f) + (1 - g)a(m)) \\
a_2(d) = g(a(f) + (1 - g)a(m))
\end{cases}
$$

where $f$ and $m$ designates the indices of the "parent" chromosomes, $d$ the resulting chromosome one, and $g$ is a random variable uniformly distributed in the interval $[0, 1]$. Eq. (3) assures that crossover of chromosomes corresponding to stable filters yields a chromosome corresponding to a stable filter too, and that it attempts to combine the qualities of the two parents.

From an adaptive filtering viewpoint, successive generations correspond to successive time indices. The evaluation function used is the instantaneous squared error. In order to compute it for each chromosome, one needs to have previous outputs of the corresponding filters, which is not possible since the population changes at each time index. The solution consists in using the same past output values $\{x_k\}$ for all filters, and in updating the latest value at each time index with the output of the filter winner of the competition between individuals.

Finally, the population evolution scheme used is the one described in [9], which is well suited for small populations and very easy to implement. It requires the number $N$ of chromosomes to be odd, and includes the following steps:
1) Computation of the evaluation function for each chromosome.
2) Reproduction of the best chromosome (winner) for the next generation.
3) $(N-1)/2$ mutations of the winner for the next generation.
4) Local competitions between pairs of chromosomes for the $(N-1)$ remaining ones.
5) The $(N-1)/2$ winners of these local competitions are then subject to crossover in order to create $(N-1)/2$ new chromosomes which complete the next generation.

In step 4) local competitions take place in order to preserve the genetic diversity of the population. If the overall best chromosomes were selected for crossover, the population would soon tend to be too uniform.
Note that the only information used in this scheme is the instantaneous squared error. Since the GACS uses solely comparisons, the magnitude of the instantaneous error could be raised to any power. It means that this algorithm is not limited to a minimum mean-squared error criterion, and should behave equally well in non-Gaussian environments.

5. Simulation results

In all the experiments presented below, the coefficients of the system to identify were taken to be $c_1 = 1.2$ and $c_2 = -0.6$. The input signal $\{x_k\}$ was an i.i.d. Gaussian noise with zero mean and unit variance. The size of the population, as in [7], was chosen to be $N = 11$. The contour plots of the mean-squared error as well as the triangle of stability in the $(a_1, a_2)$ plane are shown in figure 2. The curve shown in this figure is the average trajectory over 200 runs of the GACS, this on 250 iterations. The tendency of the algorithm to follow a steepest-descent path is clearly visible. This is a bit surprising since the GACS does not use any gradient information. Figure 3 displays the evolution of the normalized mean-squared error for a mutation variance $\sigma^2 = 0.00025$. This curve was obtained by averaging 1000 runs whose initial conditions were chosen, as in [7], randomly in the left part of the triangle of stability (i.e., $a_1 < 0$). The mean values over these 1000 runs of the coefficients corresponding to the best chromosome (after 250 iterations) were computed to be 1.198 and -0.597, with standard deviations 0.032 and 0.038. This means that almost all the final results are very close to the optimum.

Out of these 1000 runs, 99.7% yielded a filter with a normalized mean-squared error of less than 0.1. This is to be compared to the result obtained in [7] with a conventional GA, which was 53% of the runs, meaning that this algorithm presented no improvement upon a random choice of the quantized filter parameters. Moreover, the mean-squared error after convergence, namely 0.0033, and the convergence time (around 200 iterations), compare favorably with the results shown in [8] with the LMS (800 iterations).

It is interesting to examine what the respective roles of the mutation and crossover operators are in the convergence process. Figure 4 shows the average over 1000 runs of a) the percentage of winners issued from mutation and b) the proportion of winners issued from crossover. One may remark that, after a very short transient period, mutation is predominant at the beginning of the convergence, while crossover is much less important. As convergence proceeds, the two operators trade roles, and crossover becomes prevalent. This may be explained as follows: In the beginning, mutation provides a thorough search of the $(a_1, a_2)$ plane, which allows to find the first "good" solutions. Later the population becomes more homogeneous, and all chromosomes tend to be closer in terms of evaluation. Crossover by its very nature is more suited to fine adjustments, and thus is more effective when the population is closer to the optimum. This scenario seems in fact to be the common rule in any type of GA, with the two operators having a somewhat different influence but an equal importance. One may note that, while traditional Darwinian theory has always put forward mutation, large-scale simulations have recently reminded geneticists that crossover constituted an important factor of evolution [10].

![Figure 2](image2.png)

Figure 2. Triangle of stability, contour plots and average trajectory from the origin in the $(a_1, a_2)$ plane.

![Figure 3](image3.png)

Figure 3. Average mean-squared error vs. iterations

![Figure 4](image4.png)

Figure 4. a) Percentage of winners issued from mutation. b) Percentage of winners issued from crossover.
6. Extension to the lattice structure

The lattice structure has been intensively used in adaptive filtering, in spite of its higher complexity, due to its good performance in finite word length implementations, its modular structure, and the fact that stability can be easily monitored [8]. It is well known that the condition for stability of the lattice implementation of an all-pole filter is that all the reflection coefficients satisfy:

\[ |k_i| \leq 1 \]  

Thus, here again one constrains the parameters to belong to a convex domain, namely the interval [-1, 1] on the real line. But now this constraint is expressed independently for each parameter, which seems reasonable enough since decoupling of the parameters motivated the derivation of the lattice structure in the first place.

From a practical viewpoint, if one uses lattice filters instead of transversal filters, the population is composed of \( N \) chromosomes consisting of couples \( (k_1(i), k_2(i)), i = 1, \ldots, N \), \( y_1 \) being the "pool" output used by all filters as described in section 4.1. one replaces the computation corresponding to a transversal filter:

\[ y_3 = a_1y_{2,1} + a_2y_{2,2} + x_4 \]

by:

\[ y_3 = k_1(1 + k_2)y_{2,1} \cdot k_2y_{2,2} + x_4 \]

The mutation operator is defined by:

\[ k_1(d) = k(f) + v_1 \]
\[ k_2(d) = k(f) + v_2 \]

with the same convention as in eq.(2) and \( v_1, v_2 \), two Gaussian random variables with zero mean and variance \( s^2 \). Mutation is not changed, except that viability of the chromosome resulting from the mutation is checked independently on the two parameters with eq.(4). The crossover operator is defined by:

\[ k_1(d) = gk(f) + (1 - g)k(m) \]
\[ k_2(d) = h(k(d) + (1 - h)k(m) \]

with \( g \) and \( h \) two random variables uniformly distributed in the interval [0, 1]. Eq.(8) assures that crossover of chromosomes corresponding to stable filters yields a chromosome corresponding to a stable filter too, but the operator acts now separately on the two parameters.

1000 runs were performed in the same conditions as in section 5. The same mutation variance \( s^2 = 0.00025 \) was also used, because it produced a mean-squared error curve very similar to that of the previous experiment in terms of convergence speed. On the whole the results were however a bit better: 100% of the runs yielded a filter with a normalized mean-squared error of less than 0.1. The mean-squared error after convergence was 0.0026 (0.0034 for the transversal filters). The optimal reflection coefficients values are \( k_1 = -0.75 \) and \( k_2 = 0.6 \). The mean values over these 1000 runs of the reflection coefficients corresponding to the best chromosome (after 250 iterations) were computed to be -0.750 and 0.598, with standard deviations 0.022 and 0.035.

7. Conclusion

We have presented in this paper an alternative approach, called genetic algorithms in the continuous space, to classical genetic algorithms in the context of signal processing applications. In the particular case of the identification of an all-pole model, this scheme has shown superior performance in simulations with respect to a classical genetic algorithm and to a standard LMS-based technique. Although no general conclusion can be drawn from this experiment, we think that this new approach can be applied to a variety of problems in adaptive filtering, filter design, or parameter estimation. Further work could include the use of improved strategies such as the one proposed in [11] in order to reduce the importance of the choice of the mutation parameter value. A criticism that can be made to GACS is that they have to be adapted to each application, while conventional genetic algorithms are more universal. It seems however that, for many techniques and in many practical situations, a characteristic of universality brings with it undesirable features such as slow convergence speed.

References

A New Adaptive Equalizer Based on the Canonical Piecewise Linear Model

Carlos J. PANTALEON-PRIETO, Aníbal R. FIGUEIRAS-VIDAL†

Dpto. Electrónica, ETSI Telecom-Univ. Cantabria, Av Los Castros s.n., 39005 Santander, Spain.
Tel/Fax: +34 [42] 201388/201402
†DSSR, ETSI Telecom-UPM, Ciudad Universitaria, 28040 Madrid, Spain.
Tel/Fax: +34 [1] 3367226/3367350 E-Mail: werner@gits.ssr.upm.es

Abstract: Optimal Bayesian symbol equalization demands the application of non-linear filters: in this paper a novel, computationally efficient, nonlinear equalizer is developed. It is based on a piecewise-linear (PWL) approximation of the optimal (non-linear) decision boundary. To model this decision border we use the canonical PWL structure, a model that can approximate almost any continuous PWL function while using the minimum number of parameters. An adaptive implementation of the proposed approach is applied to the equalization problem. Comparative simulations show how the canonical PWL equalizer offers good performance with much less parameters than other proposed approaches.

1. Introduction

Conventional methods to suppress the channel distortion in digital communications are based on (linear) FIR adaptive filters [1]: these equalizers are robust, easy to implement and have been deeply analyzed. It has been demonstrated, however, that even with a linear distortion, the optimal Bayesian equalizer has a non-linear structure [2], so it is necessary to incorporate some non-linearity if we seek for a better performance.

Many non-linear approaches have been presented to approximate the optimal decision function. In [2] and [3] several equalizer structures based on the Multilayer Perceptron (MLP) are developed. Although the MLP architecture improves the conventional linear equalizer, it implies many parameters, long training times, and the possibility of suboptimal solutions due to local minima. Other alternatives attempt the approximation of the decision function by a Radial Basis Functions (RBF) network [4]. This structure reduces the training time because of the linear dependence on some of the parameters, but it still needs very big architectures and high complexity. It is generally accepted that the much less computational cost of the RBF network over the MLP implies the need for more data patterns and more parameters in the network.

Piecewise linear equalizers have been also proposed, as in [5], where the input space is divided into a number of different regions and a linear filter is used for every region. This approach is not a careful approximation to the optimal solution, and can originate high degradation in some practical situations. As an alternative, in [6] a tree structured piecewise-linear equalizer is developed. In both cases, however, the number of linear parts is very high, so these structures need many parameters, with the associated memory and computational costs.

In this paper, we present a new adaptive piecewise linear equalizer architecture. It attempts to approximate the decision region with as few parameters as possible, i.e., as few linear functions as possible. It can be viewed as a natural extension of the linear equalizer, in the sense that we estimate our model as a refinement of the linear case. Furthermore, the increase in the number of parameters over the linear case is not high, and the convergence is quite fast. Finally, non-linear channels can be equalized with the same approach.

The paper is organized as follows. In Section 1 the equalization problem in data communication systems is presented. A brief review of the canonical PWL model is conducted in Section 2. Section 3 explores the canonical PWL filter and its application to channel equalization. Simulation results are shown in Section 4. Some conclusions and the presentation of further lines of research close this paper.

2. Nonlinear Channel Equalization

Many digital equalizers make a decision about which is the most probable transmitted symbol using a finite number of observations, N, of the channel output. If we model the communication channel as an FIR system with impulse response \( w \), and call \( A_k \) the symbol transmitted at instant \( k \), the channel effect is modeled as

\[
x_k = \sum_{i=0}^{L-1} w_i A_{k-i} + \epsilon_k
\]

where \( x_k \) is the output and \( \epsilon_k \) the additive Gaussian noise in
the channel. Because $A_k$ is chosen from a finite number of symbols, and we suppose an FIR channel response, there exist only a finite number of possible observation points $X_k = \{X_k, X_{k-1}, X_{k-N+1}\}$. In [2] it is shown how the optimal recovery of the transmitted symbol (in the sense of minimum bit-error-rate) is inherently a nonlinear classification task, even for linear channels.

![Figure 1](image_url)  
**Figure 1.** Channel output points and decision boundary of the optimal Bayesian equalizer with a noise variance of 0.1. Channel impulse response $\{1, 0.5\}$.

This result can be confirmed in Figure 1, that shows all the possible pairs of received data samples $[X_k, X_{k-1}]$ for a channel with impulse response $h=\{1, 0.5\}$ and a two level transmitted sequence $\{1, -1\}$. The optimal decision boundary for a noise variance of 0.1 is also plotted. As it can be seen, this decision boundary is not a straight line, and therefore the optimal solution is non-linear. In the same sense it is easy to see how this optimal solution can be very well approximated with a PWL model composed of three linear blocks, as the dotted lines in Figure 1 suggest.

Nonlinear equalization can be also be analyzed under a deconvolution framework [3]: with this approach, it is believed that the noise enhancement problem in linear equalization (the amplification of the channel noise due to the equalizer placing large gains around the spectral nulls of the channel spectral response) can be partially solved by using nonlinear equalization techniques.

3. The Canonical PWL Model

A conventional piecewise linear model, as that considered in [5], is composed of two steps: a subdivision of the domain space and the estimation of a linear model in each region. This approach needs many parameters to reach a reasonable performance and, so, it has long training times. As an alternative, the canonical piecewise-linear model offers a general modeling approach for piecewise-linear models that minimizes the number of parameters.

The canonical PWL representation was introduced by Chua et al [7] to perform the analysis, and modeling of nonlinear circuits. A canonical PWL function $f: \mathbb{R}^N \rightarrow \mathbb{R}^N$ takes the form

$$f(x) = a + Bx + \sum_{i=1}^{m} c_i \left( \langle \alpha_i, x \rangle - \beta_i \right)$$

with $B \in \mathbb{R}^{N \times N}$, $a, c_i, \alpha_i \in \mathbb{R}^N$, $\beta_i \in \mathbb{R}$ and $\langle \rangle$ denotes the inner product between two vectors. The canonical PWL model is restricted to approximate continuous PWL $n$-dimensional functions with $(n-1)$ dimensional linear boundaries. This is, the data space is divided in regions that are defined by hyperplanes of the form

$$\langle \alpha_i, x \rangle - \beta_i = 0$$

This model depends linearly on $a$, $B$, $c_i$ and non-linearly on $\alpha_i$ and $\beta_i$. In [7] an optimization technique is developed to obtain these parameters. It combines a least squares estimation of the linear dependent parameters with a grid search in the direction of the gradient to obtain the ones that exhibit nonlinear dependance.

The canonical PWL model can achieve an important reduction in number of parameters over conventional approaches, and very efficient computation methods have been developed to analyze networks made up of this kind of models [8].

4. Canonical PWL Adaptive Equalization

An equalizer, as presented in Section 2, can be seen as a nonlinear functional that maps input vectors of length $N$ in an output decision: for this application the model in eq.(2) must be restricted to functions from $\mathbb{R}^N \rightarrow \mathbb{R}$. If we also force a symmetric response and consider a three linear block model making $\sigma=2$, eq.(2) reduces to

$$f(x) = h^T x + c \left[ \alpha^T x - \beta \right]$$

with $h, \alpha \in \mathbb{R}^N$ and $c, \beta \in \mathbb{R}$. The linear equalizer is transformed in a non-linear one, using a piecewise linear approximation of the optimal decision function (note that, with $c=0$, we obtain the linear equalizer). The piecewise-linear model in eq.(4) is composed of three linear terms (we suppose a symmetric channel, so an equalizer with only two linear blocks is not possible)

$$f(x) = \begin{cases} h^T x + 2c \beta & \text{for } \alpha^T x \leq -\beta \\ (h - 2c \alpha)^T x & \text{for } -\beta < \alpha^T x < \beta \\ h^T x - 2c \beta & \text{for } \alpha^T x \geq \beta \end{cases}$$

Of course, in order to use this model as a non-linear equalizer we must be able to estimate its parameters adaptively. In [9] an adaptive implementation of the canonical PWL is presented: assuming that $d_k$ is a finite
memory continuous function of the input $x_n$, (as the boundary in Figure 1) and can be approximately represented by a PWL function, the MSE cost function in iteration $k$ is defined as

$$J_k = E[c_k^2] = E[(d_k - f(x_k))^2]$$

where $E(.)$ denotes expectation. An LMS adaptation algorithm is derived to obtain the parameters $h$, $c$, $\alpha$ and $\beta$ in eq.(4). The resulting expressions are

$$h_{k+1} = h_k + \mu_h \cdot c r r_k \cdot x_k$$

$$c_{k+1} = c_k + \mu_c \cdot c r r_k \cdot \left[\alpha_k^\top x_k - \beta_k\right]$$

$$\alpha_{k+1} = \alpha_k + \mu_\alpha \cdot c r r_k \cdot c_k \cdot \left[\alpha_k^\top x_k - \beta_k\right]$$

$$\beta_{k+1} = \beta_k - \mu_\beta \cdot c r r_k \cdot c_k \cdot \left[\alpha_k^\top x_k - \beta_k\right]$$

The equations shown in eq.(7) are not exactly the same as in [9] because in that application $\beta$ is given a constant value of 1. Although this simplification restricts the model in a general application, it is also useful in our case. In the two tap equalizer in Figure 1., the canonical expression in eq.(4) is overparametrized: we only need 4 parameters to define the piecewise-linear border and one more to define the gain (see eq.(5)). In our simulations we also give $\beta$ a fixed value. In [8] the convergence properties of this adaptation algorithm are reviewed and equations are given for the adaptation parameters.

The cost function (6) is quadratic in $h$ and $c$, so the algorithm converges to a global minimum if the rest of the parameters are fixed. In the general case, however, due to the unconvex MSE cost function, the adaptation algorithm may lead to a local minimum, giving a suboptimal solution. Nonetheless, if we start our estimation algorithm with the linear equalizer as the initial solution, at least we are going to maintain its performance. If a better result is pursued, many nonlinear optimization algorithms have been developed to overcome local minima and could be applied to this problem[10].

5. Simulation Results

In our simulations we analyze the performance of the proposed equalizer evaluating its probability of error for different SNR situations, as well as its speed of convergence.

The transmission channel is FIR with impulse response (1.0.5) as shown in Figure 1., and used in [5]. We adapt an equalizer as shown in eq.(4) with $\mu_h=0.4$, $\mu_c=0.01$, and $\mu_\alpha=0.1$. $\mu_h$ is decreased in every iteration according to

$$\mu_{h,k+1} = 0.99 \cdot \mu_{h,k}$$

(8)

to get a reduced residual error when convergence is achieved. The starting values for $h$ and $c$ are random. The $\alpha$ vector is initialized as an orthogonal vector to the estimate of $h$ in the first iterations. This initialization is made to avoid that $h$ and $\alpha$ take similar values, a situation that slows down the convergence of the algorithm. As we said before, we give to $\beta$ a constant value of 0.5.

![Figure 2. Probability of error for the linear '1-' and PWL 'o-o' equalizer for different Signal to Noise ratios. Channel $H(z)=1-0.5z^{-1}$.

Figure 2 shows the probability of error in equalizing the previously proposed channel for the linear and PWL equalizers of length two for different signal to noise ratios. It can be seen the superior performance of the piecewise linear equalizer, mostly at high signal to noise ratios. It has been obtained applying both equalizers to a sequence of 500,000 samples for every SNR value. These results are very similar to those obtained with the PWL model in [5], but the number of parameters is greatly reduced.

![Figure 3. Evolution of the Probability of error for the linear and PWL model. SNR=14 dB.](image-url)
In Figure 3 is presented the evolution of the probability of error with training. It can be seen how the proposed algorithm shows a speed of convergence comparable with the linear one. The parameters of both equalizers are the same as in the previous case. These results have been obtained by averaging 30000 times the first 2000 iterations of both algorithms and low pass filtering the results.

Finally, the channel \( H(z) = 0.3482 + 0.8704z^{-1} + 0.3482z^{-2} \) is equalized. The same parameters are maintained as in previous cases and the vector is also initialized as an orthogonal vector to the estimate of \( \mathbf{h} \). Note, however, that in this case many possible orthogonal vectors can be defined. We obtain ours by selecting the two components of higher modulus in \( \mathbf{h} \) and defining \( \mathbf{a} \) as orthogonal in that direction.

In Figure 4, we present the probability of error obtained when equalizing 500,000 sample sequences. We compare linear and PWL equalizers, both of them with a length of 4 and a delay of 1. It can be seen how the proposed model also outperforms the linear model in this case. With the same channel, MLP and RBF networks get better results [2,4], but with much bigger architectures.

![Figure 4. Probability of error for the linear --- and PWL -o-o- equalizer for different Signal to Noise ratios. Channel 0.3482+0.8704z^{-1}+0.3482z^{-2}.

6. Conclusions
Based on the fact that optimal Bayesian symbol by symbol equalization is an inherently non-linear problem, it has been shown how the canonical PWL equalizer constricts the optimal decision boundary off the linear one. It has been also shown that the proposed approach demands a reduced number of parameters as well as a low computational cost; becoming an intermediate possibility when the high complexity of other models cannot be accepted. Several simulations have confirmed the good behavior of the proposed model in comparison with linear approaches.

Nonetheless, some problems exist. The proposed estimation approach relies on a gradient optimization that can converge to local minima: some research should be conducted to avoid this situation. The optimal decision boundary is very well approximated by a PWL function in low noise applications: when the noise increases, however, the optimal border gets away from the PWL case, so it will be interesting if we could smooth out the PWL solution to get closer to the optimum border. Finally, when the proposed technique is applied to more realistic channels the optimal decision boundary becomes very complex. It demands many linear components and, in this case, gradient type approaches suffer high degradation so more adequate parameter estimation approaches should be investigated.

Acknowledgment
This work has been partially supported by CYCYT grant TIC92-0800-C05-01.

References
The Fast Subsampled-Updating Stabilized Fast Transversal Filter (FSU SFTF) RLS Algorithm

Dirk T.M. SLOCK† and Karim MAOUCHE‡

†, ‡ Institut EURECOM, 2229 route des Crêtes
B.P. 193, 06904, Sophia Antipolis Cedex, FRANCE
Tel/Fax: +33 93902606/ 93008627, E-Mail: slock@eurecom.fr
Tel/Fax: +33 93902632/ 93008627, E-Mail: maouche@eurecom.fr

Abstract. We present a fast new algorithm for Recursive Least-Squares (RLS) adaptive filtering that uses displacement structure and subsampled updating. The FSU SFTF algorithm is based on the Stabilized Fast Transversal Filter (SFTF) algorithm, which is a numerically stabilized version of the classical FTF algorithm. The FTF algorithm exploits the shift invariance that is present in the RLS algorithm of a FIR filter. The FTF algorithm is in essence the application of a rotation matrix to a set of filters and in that respect utilizes the Levinson algorithm. In the subsampled updating approach, we accumulate the rotation matrices over some time interval before applying them to the filters. It turns out that the successive rotation matrices themselves can be obtained from a Schur type algorithm which, once properly initialized, does not require inner products. The various convolutions that thus appear in the algorithm are done using the Fast Fourier Transform (FFT). For relatively long filters, the computational complexity of the new algorithm is smaller than the one of the well-known LMS algorithm, rendering it especially suitable for applications such as acoustic echo cancellation.

1 Introduction

Fast Recursive Least Squares (RLS) algorithms such as the Fast Transversal Filter (FTF) algorithm [1] exploit a certain shift invariance structure in the input data vector to reduce the computational complexity from $O(N^2)$ for RLS to $O(N)$ for FTF $(N$ being the FIR filter length). In [3],[4],[5], we have pursued an alternative way to reduce the complexity of RLS adaptive filtering algorithms. The approach consists of sub-sampling the filter adaptation, i.e. the LS filter estimate is no longer provided every sample but every $L \geq 1$ samples (subsampling factor $L$). This strategy has led us to derive two new RLS algorithms that are the FSU RLS and FSU FTF algorithms which present a reduced complexity when dealing with long filters. Here, we extend the FSU FTF idea to the Stabilized FTF algorithm (SFTF) which is a numerically stabilized Version of the FTF. The starting point is an interpretation of the SFTF algorithm as a rotation applied to the vectors of filter coefficients. Using the filter estimates at a certain time instant, we compute the filter outputs over the next $L$ time instants. Using what we shall call a SPTF-Schur algorithm, it will be possible to compute from these multi-step ahead predicted filter outputs the one step ahead predicted filter outputs in an efficient way. These quantities will allow us to compute the successive rotation matrices of the SFTF algorithm for the next $L$ time instants. Because of the presence of a shift operation in the SFTF algorithm, it turns out to be most convenient to work with the $z$-transform of the rotation matrices and the filters. Applying the $L$ rotation matrices to the filter vectors becomes an issue of multiplying polynomials, which can be efficiently carried out using the FFT. The subsampled updating technique turns out to be especially applicable in the case of very long filters such as occur in the acoustic echo cancellation problem. The computational gain it offers is obtained in exchange for some processing delay, as is typical of block processing.

In order to formulate the RLS adaptive filtering problem and to fix notation, we shall first recall the RLS algorithm.

2 The RLS Algorithm

An adaptive transversal filter $W_{N,k}$ forms a linear combination of $N$ consecutive input samples $\{x(i-n), n = 0, \ldots, N-1\}$ to approximate (the negative of) the desired-response signal $d(i)$. The resulting error signal is given by

$$
\epsilon_N(i|k) = d(i) + W_{N,k}^H X_N(i) = d(i) + \sum_{n=0}^{N-1} W_{N,k}^{n+1} x(i-n)
$$

(1)

where $X_N(i) = [x_H(i) x_H(i-1) \cdots x_H(i-N+1)]^H$ is the input data vector and superscript $^H$ denotes Hermitian (complex conjugate) transpose. In the RLS algorithm, the set of $N$ transversal filter coefficients $W_{N,k} = [W_{N,k}^1 \cdots W_{N,k}^N]$ are adapted so as to minimize recursively the following LS criterion

$$
\xi_N(k) = \min_{W_{N,k}} \left\{ \sum_{n=1}^{k} \lambda^{k-n} \|d(i) + W_N^H X_N(i)\|^2 \right\}
$$

(2)

where $\lambda \in (0,1]$ is the exponential weighting factor, $\|v\|^2 = v^H v$, $\|v\| = \|v\|_F$. Minimization of the LS criterion leads to the following minimizer

$$
W_{N,k} = -(P_{N,k}^{-1} R_{N,k}^{-1})^H
$$

(3)
where

\[ R_{N,k} = \lambda R_{N,k-1} + X_N(k)X_N^H(k) \]
\[ P_{N,k} = \lambda P_{N,k-1} + X_N(k)d^H(k) \]

are the sample second order statistics. Substituting the time recursions for \( R_{N,k} \) and \( P_{N,k} \) from (4) into (3) and using the matrix inversion lemma for \( R_{N,k}^{-1} \), we obtain the RLS algorithm:

\[ \tilde{C}_{N,k} = -X_N^H(k)\lambda^{-1}R_{N,k-1}^{-1} \]
\[ \gamma_N^{-1}(k) = 1 - \tilde{C}_{N,k}X_N(k) \]
\[ R_{N,k}^{-1} = \lambda^{-1}R_{N,k-1}^{-1} - \tilde{C}_{N,k}^H\gamma_N(k)\tilde{C}_{N,k} \]
\[ \epsilon_N^P(k) = \epsilon_N(k|k-1) = d(k) + W_{N,k-1}X_N(k) \]
\[ \epsilon_N(k) = \epsilon_N(k|k) = \epsilon_N^P(k)\gamma_N(k) \]
\[ W_{N,k} = W_{N,k-1} + \epsilon_N(k)\tilde{C}_{N,k} \]

where \( \epsilon_N^P(k) \) and \( \epsilon_N(k) \) are the a priori and a posteriori error signals (resp. predicted and filtered errors in the Kalman filtering terminology) and one can verify (or see [1]) that they are related by the likelihood variable \( \gamma_N(k) \) as in (9). \( \tilde{C}_{N,k} \) is the Kalman gain of order \( N \) at time \( k \).

### 3 The Stabilized Fast Transversal Filter Algorithm

The computational complexity of the FTF algorithm is \( 7N \) in its most efficient form [1]. However, the FTF suffers from numerical instabilities that are due to long-term round-off error propagation. In [2], a stabilization procedure has been introduced to overcome the instability problem. This has led to the Stabilized FTF algorithm (SFTF) with computational complexity \( 8N \). In what follows we shall consider the single-channel case. However the generalization to the multichannel case can easily be done. The SFTF algorithm can be described in the following way, which emphasizes its rotational structure:

\[
\begin{bmatrix}
\tilde{C}_{N,k}^1 \\
A_{N,k} \\
B_{N,k} \\
W_{N,k}
\end{bmatrix}
= \Theta_k
\begin{bmatrix}
0 \\
\tilde{C}_{N,k-1}^1 \\
A_{N,k-1} \\
B_{N,k-1} \\
W_{N,k-1}
\end{bmatrix}
\]

\[
\begin{align*}
e_N^P(k) &= A_{N,k-1}X_{N+1}(k) \\
\epsilon_N(k) &= \epsilon_N^P(k)\gamma_N(k-1) \\
\gamma_{N+1}(k) &= \gamma_N(k-1) - \tilde{C}_{N+1,k}e_N^P(k) \\
\gamma_{N,k} &= \gamma_{N+1}(k) + \tilde{C}_{N+1,k}r_N^P(k) \\
r_N^P(k) &= -\lambda B_{N,k-1}\tilde{C}_{N+1,k} \\
r_N^P(k) &= B_{N,k-1}X_{N+1}(k) \\
\sigma_N^{-1}(k) &= \lambda^{-1}\sigma_N^{-1}(k-1) - \tilde{C}_{N+1,k}\sigma_N^{-1}(k) \tilde{C}_{N+1,k}^H \\
r_N^{(j)}(k) &= K_jr_N^P(k) + (1-K_j)r_N^F(k) \\
\beta_N(k) &= \lambda\beta_N(k-1) + \sigma_N^{-1}(k) \tilde{C}_{N,k} \\
\gamma_N(k) &= \lambda^2\beta_N(k)\sigma_N^{-1}(k)
\end{align*}
\]

where \( A_{N,k} \) and \( B_{N,k} \) are the forward and backward prediction filters, \( \epsilon_N^P(k) \) and \( \epsilon_N(k) \) are the a priori and a posteriori forward prediction errors, \( r_N^P(k) \) and \( r_N^F(k) \) are the a priori and a posteriori backward prediction errors, \( \tilde{C}_{N+1,k} \) are the output and backward prediction error variances. \( K_j = 1.5 \) and \( K^2 = 2.5 \) are the optimal feedback gains that ensure the stability of the dynamics of the accumulated round-off errors [2]. \( \Theta_k \) is a \( 4 \times 4 \) rotation matrix given by

\[
\Theta_k = \Theta_k^1 \Theta_k^2 \Theta_k^3 \Theta_k^4
\]

where the four \( 4 \times 4 \) matrices \( \Theta_k^i \) for \( i = 1, 2, 3, 4 \) are

\[
\Theta_k^1 = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
a & 0 & 0 & 1
\end{bmatrix}
\]
\[
\Theta_k^2 = \begin{bmatrix}
1 & 0 & c & 0 \\
0 & 1 & 0 & d \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

with \( a = \epsilon_N(k), b = r_N^{(1)}(k), c = -\tilde{C}_{N+1,k}^N, d = \epsilon_N(k) \) and \( \tilde{C}_{N+1,k}^N = \lambda^{-1}\sigma_N^{-1}(k-1). \)

In order to compute the rotation matrices, one must obtain the a priori errors \( \epsilon_N^P(k), r_N^{(1)}(k) \) and \( \epsilon_N^P(k) \) which are the outputs at time \( k \) of the filters \( A_{N,k-1}, B_{N,k-1} \) and \( W_{N,k-1} \).

### 4 The SFTF-Schur Algorithm

Now we introduce subsampled updating and from the filters at time instant \( k-L \), we want to obtain the filters at time instant \( k \). This will require the rotation matrices and hence the a priori errors in that time range. We shall show that these quantities can be computed without generating (completely) the intermediate filter estimates using a SFTF-Schur algorithm. Let us introduce the negative of the filter output

\[ \hat{A}_N^P(k) = d(k) - \epsilon_N^P(k), \quad \hat{A}_N(k) = d(k) - \epsilon_N(k) \]

Consider now the following set of filtering operations

\[
F_L(k) \triangleq \begin{bmatrix}
\eta_{N+1,L,k}^H \\
\epsilon_{N+1,L,k}^P \\
r_{N+1,L,k}^F \\
\tilde{C}_{N+1,k}
\end{bmatrix}
= \begin{bmatrix}
0 \\
\tilde{C}_{N,k-1} \\
A_{N,k-1} \\
B_{N,k-1} \\
W_{N,k-1}
\end{bmatrix}
\begin{bmatrix}
X_{N+1,L,k}^H \\
X_{N+1,L,k}^P \\
X_{N+1,L,k}^F \\
X_{N+1,L,k}
\end{bmatrix}
\]

where

\[
X_{N+1,L,k} = [X_N(k-L+1) \cdots X_N(k)]^H
\]

is the \( L \times (N+1) \) Toeplitz input data matrix. \( F_L(k) \) is a \( 4 \times L \) matrix, the rows of which are the result of the filtering of the data sequence \( \{x(j)\}, j = k-N-L+1, \ldots, k \) by the four filters of the SFTF algorithm. \( \eta_{N,L,k} \) is the output of the Kalman gain and \( e_{N,L,k}^P \) and \( r_{N,L,k}^H \) are respectively the vectors of forward and backward prediction errors

\[
e_{N,L,k}^P = \begin{bmatrix}
e_{N,k-1}^P(k-L) \\
\epsilon_{N,k-L}^F(k-L) \\
\vdots \\
e_{N,k-L}^P(k-L)
\end{bmatrix}, r_{N,L,k}^H = \begin{bmatrix}r_{N,k}^H(k-L+1\{k-L) \\
\vdots \\
r_{N,k}^H(k-L)
\end{bmatrix}
\]

(17)
The last row of \( F_L(k) \) corresponds to the (multi-step ahead predicted) adaptive filter outputs

\[
\begin{bmatrix}
\hat{d}_H^P(k-L+1|k-L) \\
\vdots \\
\hat{d}_H^P(k|k-L)
\end{bmatrix} = 

\begin{bmatrix}
\delta_H^P(k-L+1|k-L) \\
\vdots \\
\delta_H^P(k|k-L)
\end{bmatrix}
\]  

(18)

The first column of \( F_L(k) \) is

\[
F_L(k) \ u_{L,1} = 
\begin{bmatrix}
1 - \tau_{N}^{-1}(k-L) \\
\epsilon_N^P(k-L+1) \\
r_{N}^P(k-L+1) \\
-\hat{d}_N^P(k-L+1)
\end{bmatrix}
\]  

(19)

where \( u_{L,1} \) is the \( L \times 1 \) vector with 1 at the \( n^{th} \) position and 0 elsewhere. In order to obtain \( \Theta_{k-L+1} \), one needs to compute \( r_{N}^{P}(k-L+1) \) and hence \( \tilde{C}_{N,L}^{-1} \). In fact, it turns out that the different \( \tilde{C}_{N,L}^{-1} \) for \( j = 1, \ldots, L \) can be obtained by carrying out the SFTF recursions on the last \( L-j \) entries of the filters.

For \( j = 1, \ldots, L \) do

\[
\begin{align*}
m &= N - L + j \\
K &= k - L + j \\
\tilde{C}_{N+1,k}^{-1} &= \tilde{C}_{N+1,k}^{-1}(K) = \tilde{C}_{N+1,k}^{-1} - \lambda^{-1}(K-1) \tilde{e}_N^H(K) A_{N+1,k}^{-1} \\
A_{N+1,k}^{-1} &= A_{N+1,k}^{-1} - \lambda \tilde{e}_N^H(K) B_{N+1,k}^{-1} \\
B_{N+1,k}^{-1} &= B_{N+1,k}^{-1} + r_{N}(k) \tilde{C}_{N+1,k}^{-1}
\end{align*}
\]

(20)

Counting only the most significant term as we often do, the computational complexity of these recursions is \( 2L^2 \). So with the quantities in \( F_L(k) \) and the recursions (11) and (20), it is possible to construct \( \Theta_{k-L+1} \). Now we rotate both expressions for \( F_L(k) \) in (15) with \( \Theta_{k-L+1} \) to obtain \( \Theta_{k-L+1} \).

\[
\begin{bmatrix}
\tilde{C}_{N,L}^{-1} 0 \\
A_{N,L+1} \\
B_{N,L+1} \\
W_{N,k+1}
\end{bmatrix} = 
\begin{bmatrix}
\tilde{C}_{N,k} 0 \\
A_{N,k} \\
B_{N,k} \\
W_{N,k}
\end{bmatrix} \begin{bmatrix}
z^{-1} \\
1 \\
A_{k-1} \\
B_{k-1} \\
W_{k-1}
\end{bmatrix}
\]

(23)

Hence (11) can be written in the \( z \)-transform domain as

\[
\begin{bmatrix}
\tilde{C}_{k} 0 \\
A_{k} \\
B_{k} \\
W_{k}
\end{bmatrix} = \Theta_{k} \begin{bmatrix}
z^{-1} \\
1 \\
A_{k-1} \\
B_{k-1} \\
W_{k-1}
\end{bmatrix}
\]

(24)

It appears natural to introduce

\[
\Theta_{k} = \Theta_{k} \begin{bmatrix}
z^{-1} \\
1 \\
1 
\end{bmatrix}
\]

(25)

Now, in order to adapt the filters at time \( k \) from the ones at time \( k-L \), we get straightforwardly

\[
\begin{bmatrix}
\tilde{C}_{k} 0 \\
A_{k} \\
B_{k} \\
W_{k}
\end{bmatrix} = \Theta_{k} \begin{bmatrix}
z^{-1} \\
1 \\
A_{k-1} \\
B_{k-1} \\
W_{k-1}
\end{bmatrix}
\]

(26)

where

\[
\Theta_{k} = \Theta_{k} \Theta_{k-1} \cdots \Theta_{k-L+1}(z)
\]

(27)

As mentioned before, the successive rotation matrices can be obtained via the SFTF-Schur algorithm with a computational complexity of \( 4.5L^2 \) operations, which takes into account the fact that a rotation matrix in factored form as in (13) only contains five non-trivial entries. Now also remark that \( \Theta_{k} \) has the following structure

\[
\Theta_{k}(z) = \begin{bmatrix}
\ast & \ast & \ast & 0 \\
\ast & \ast & \ast & 0 \\
\ast & \ast & \ast & 1
\end{bmatrix}
\]

(28)
Table I: the FSU SFTF Algorithm

<table>
<thead>
<tr>
<th>Computation</th>
<th>Cost per L samples</th>
</tr>
</thead>
</table>
| \[
\begin{bmatrix}
\eta_{N,k,L}^R, L,k \\
\eta_{N,k,L}^I, L,k \\
\n\hat{\eta}_{N,k,L}^R, L,k \\
\hat{\eta}_{N,k,L}^I, L,k
\end{bmatrix}
\begin{bmatrix}
0 \\
\hat{C}_{N,k,L} \\
A_{N,k,L} \\
B_{N,k,L} \\
W_{N,k,L,0}
\end{bmatrix}
\] \(X_{N+1,L,L}^R (5 + 4 N_L^L) \cdot \text{FFT}(2L) + 8N\) |

2 SFTF-Schur Algorithm:
Input: \(\eta_{N,L,k}, \eta_{N,L,k}^R, \eta_{N,L,k}^I, \hat{\eta}_{N,L,k}^R, \hat{\eta}_{N,L,k}^I\)
Output: \(\Theta_{k} (z), \quad i = L - 1, \ldots, 0\)
\(4.5L^2\)

3 \(\Theta_{k,L} (z) = \prod_{i=0}^{L-1} \Theta_{k-i} (z)\)
\(7.5L^2\)

4 \[
\begin{bmatrix}
\hat{C}_k (z) \\
A_k (z) \\
B_k (z) \\
W_k (z)
\end{bmatrix}
\begin{bmatrix}
\hat{C}_{k-L} (z) \\
A_{k-L} (z) \\
B_{k-L} (z) \\
W_{k-L} (z)
\end{bmatrix}
\] \((12 + 4 N_L^L) \cdot \text{FFT}(2L) + 24N\)

Total cost per sample \((17 + 8 N_L^L) \cdot \text{FFT}(2L) + 32 N_L^I + 12L\)

where the stars stand for polynomials in \(z^{-1}\) of degree at most \(L\). Taking into account these two remarks, the accumulation of the successive rotation matrices to form \(\Theta_{k,L} (z)\) takes \(7.5L^2\) operations. As a result of the structure displayed in (28), the product in (26) represents 12 convolutions of a polynomial of order \(L\) with a polynomial of order \(N\). These convolutions can be done using fast convolution techniques. In the case we consider, in which the orders of the polynomials are relatively large, we will implement the convolutions using the FFT technique. In that case the complexity for the update of each one of the four filters is \(3(1 + 2 N_L^L) \cdot \text{FFT}(2L) + 5(N + 1)\) (multiply/add) operations plus \(6(N + 1)\) additions (\(\text{FFT}(m)\) denotes the computational complexity for computing a FFT of length \(m\), and we assume that \(L\) is a power of 2 and \(N_L^L\) is an integer). The computation of \(P_k (k)\) in (15) can also be done with the FFT and one should compute the FFTs of the filters only once. In the Overlap-Save method, the data matrix is decomposed into \(N_L^L\) blocks of \(L \times L\) Toeplitz matrices, which are then embedded into \(L \times L\) Toeplitz matrices. Note that at time \(k\), only the most recent \(2L\) samples of the input signal, corresponding to the new \(L \times L\) block in the data matrix, have to be Fourier transformed. The other parts have been computed at previous instants (see [5] for more details). The resulting FSU SFTF algorithm is summarized in Table I.

6 Concluding Remarks

The complexity of the FSU SFTF is \(O((8 N_L^L + 17) \cdot \text{FFT}(2L) + 32 N_L^I + 12L)\) operations per sample. This can be very interesting for long filters. For example, when \((N,L) = (4095,256), (8191,256)\) and the FFT is done via the split radix \((\text{FFT}(2m) = m \log_2(2m)\) real multiplications for real signals) the multiplicative complexity is respectively \(1.2N\) and \(0.6N\) per sample. This should be compared to \(8N\) for the SFTF algorithm, the currently fastest stable RLS algorithm, and \(2N\) for the LMS algorithm. The number of additions is somewhat higher. The cost we pay is a processing delay which is of the order of \(L\) samples. We have simulated the algorithm and have verified that it works. In [3,6], we have introduced the FSU RLS algorithm, an alternative algorithm with a very similar computational complexity, but a very different internal structure. These developments lead us to conjecture that perhaps a lower bound on computational complexity has been reached for RLS algorithms when the subsampled updating strategy is applied and when the filters to be adapted are relatively long.

REFERENCES

A Set of Algorithms Linking NLMS and RLS Algorithms

Mohsen MONTAZERI and Pierre DUHAMEL†

France Telecom-CNET/PAB/RGF/RCN
38-40 rue du Général Leclerc, 92131 Issy les Moulineaux, France
† Télécom Paris, Département Signal, 46 rue Barrault, 75634 Paris Cedex 13, France
Tel: ++331 45817365, Fax: ++331 45887935

Abstract. This paper proposes a modified projection algorithm with increased robustness when modelling noise is present in the reference signal. Projection algorithms form a set of adaptive filters which contains as external cases the Normalized LMS (NLMS) and the Recursive Least Squares (RLS) algorithms. The modified algorithms allow a large set of tradeoffs between convergence rate, residual error, tracking capacity, and arithmetic complexity. Several fast algorithms are feasible: we propose here a version which computes only the modelization error (hence can be used in applications such as acoustic echo cancellation), and requires $2L + 15N$ multiplications ($L$ is the filter length, $N < L$). An analysis of the adaptive behavior of the algorithm is outlined. It is further illustrated by simulations and compared to the LMS and RLS algorithms and to the Fast Newton Transversal Filter in a realistic context.

1. Introduction

A major difficulty encountered in acoustic echo cancellation is the requirement for very long adaptive filters (the filter length $L$ is often chosen between 256 and 4000). Thus, the fast versions of the RLS algorithm are disqualified because of their complexity. On another side, the convergence rate of the LMS algorithm is very slow when the input signal is correlated and non-stationary, which is the case in this application. Hence, variations of the LMS, such as frequency domain adaptive filters or subband adaptive filtering schemes are often used. However, they often require large blocks, hence introduce medium to large processing delay (not allowed in a hand free phone application) and often lack tracking capacity, another point which is very important since the system to be identified is non-stationary.

Another approach would be to simplify RLS algorithms, as was done in the recently proposed Fast Newton Transversal Filter (FNTF) algorithm [5]. The FNTF reduces the computational complexity of the FRLS algorithm to $2L + 12N$ [5] with a hardly noticeable degradation of its performance for medium size filters ($L \approx 256$), this degradation increasing with the filter size [5]. The FNTF algorithm can be seen as a generalisation of both NLMS and RLS algorithms.

Another generalization of both NLMS and RLS algorithms is the so-called Affine Projection Algorithm (APA) [1], [2]. The general APA is based on a multiple dimension projection per tap update. Fast versions have been derived [2], [4]. We show in this paper that APA have poor performances when modelling noise is present in the reference signal (and especially when the input signal is non-stationary), which is the case in actual applications.

This paper proposes a modified APA solving this problem. The resulting algorithm has a convergence rate comparable to that of FRLS algorithms even when the input signal is non-stationary and noise is added to output signal, with a computational complexity about four times smaller. The new algorithm is derived in section 2. Section 3 provides a fast version, and its adaptive behavior is analyzed in section 4. Section 5 provides simulations, comparing the modified APA to NLMS, stabilized FTF and FNTF algorithms.

2. The Proposed Algorithm

2.1. Projection algorithms and regularized ones

Consider the so-called projection algorithm given in [1], [2]:

$$E_n' = Y_n - X_n' H_n$$

$$H_{n+1} = H_n + \alpha X_n (X_n' X_n)^{-1} E_n'$$

(1)

(2)

where $H_n$ is the filter weight vector at time $n$. The $L \times N$ matrix $X_n = (X_n, X_{n-1}, \ldots, X_{n-N+1})$ is made from the $N$ last input vectors and $Y_n$ (respectively $E_n'$) is the vector of the $N$ last samples of the reference signal $y_n$ (resp. the modelisation error). Usually, $N < L$, and $0 < \alpha < 1$. Throughout the paper, all sequences are assumed to be real-valued.

This algorithm has a number of attractive properties: its convergence rate is very close to that of the RLS algorithm and fast algorithms for the sample by sample version require $2L + 21N$ multiplications [4]. Fast algorithms for the block version [2], [3] even require fewer operations than the standard LMS algorithm. However, a number of problems arise in the very circumstances met in practical cases: when the input signal is non-stationary and noise is added to reference signal, this algorithm has a large residual error. It has been shown in [2] that for large time index $n$, the contribution of noise to the Mean-square error $J_n$ can be approximated by:

$$J_n \approx \sigma^2 + \alpha^3 \sigma^2 \times \text{tr} \left[ X_n' X_{n-1} (X_{n-1}' D_{n-1})^{-1} (X_{n-1}' D_{n-1})^{-1} X_{n-1}' X_n \right]$$

(3)

where $\sigma^2$ is the variance of the modelling noise. It is seen from (3) that the noise gets amplified significantly when the
sample covariance matrix $X_n^T X_{n-1}$ is badly conditioned. This happens for example when the input signal is non-stationary and very harmonic, like speech. This problem is reduced by using regularized algorithms [3] [4]:

$$
E_n = Y_n - X_n^T H_n
$$

$$
H_{n+1} = H_n + \alpha X_n (X_n^T X_n + \delta I)^{-1} E_n
$$

where the scalar $\delta$ is the regularization parameter.

2.2. Derivation of the modified projection algorithms

Another remedy to the conditioning problem would be to replace $X_n^T X_{n-1}$ by another estimate of the correlation matrix with lower statistical variations and better conditioning factor. The following exponential window estimate has the desired property:

$$
R_n = \lambda R_{n-1} + L(1-\lambda) X_n^T X_n
$$

where $0 < \lambda < 1$, is the forgetting factor, and $X_n = (x_n, x_{n-1}, \ldots, x_{n-N+1})^T$ ($R_0$ being initialized to a positive diagonal value).

Simulations showed that this estimation of the covariance matrix results in conditioning factors altogether much smaller and more stable than the other two estimates. This should improve the convergence characteristics of the corresponding algorithm in presence of noise.

However, just replacing the matrix in (2) would not be reasonable, without deriving the algorithm with respect to some criterion. This can be done by using a generalized version of the one found in [6]:

$$
\min_{H_{n+1}} \{ \| Y_n - X_n^T H_{n+1} \|^2 + \| H_{n+1} - H_n \|^2_{Q_n} \}
$$

where $\| \cdot \|_{Q_n} = (\cdot)^T P_n (\cdot)$ and $P_n, Q_n$ are Hermitian positive definite matrices, defined by $P_n = \frac{1}{1-\lambda^2} R_n - \frac{1}{\lambda} X_n^T X_n$, $Q_n = I$ (identity matrix) of size $N \times N$, which results in:

$$
E_n = Y_n - X_n^T H_n
$$

$$
H_{n+1} = H_n + \alpha X_n R_n^{-1} E_n
$$

2.3. Special cases

Rewriting the equations (8) and (9) in the special case where $N = 1$, we obtain the NLMS algorithm where the estimation of the power of $z_n$ is done by an exponential window. When $N = L$ after multiplying equation (2) by $X_n X_n^T$, provided $\alpha = 1$, we obtain:

$$
\hat{e}_n = y_n - X_n^T H_n
$$

$$
H_{n+1} = H_n + \alpha X_n R_n^{-1} E_n
$$

When replacing the matrix $(X_n X_n^T)$ by the matrix $R_n$ defined in (6), (10) and (11) clearly correspond to an exponential window RLS algorithm.

3. Computational complexity

Using the same technique as explained in [4], a Fast version of APA is obtained. This algorithm computes only the modelling errors, and not the actual filter coefficients, which is not a problem in acoustic echo cancellation. This algorithm requires $2L + 15N + 6$ multiplications.

Fast Algorithm

1. Initialization:

   (a) $S_0 = 0$, $E_0 = 0$, $E_0^{-1} = 0$, $H_0 = 0$, $E_0^{-1} = 0$

   (b) Initial conditions of step 2: $A_0 = 0$, $B_0 = 0$, $E_0 = \frac{\eta^2}{\alpha^2}$, $E_0^{-1} = \frac{\eta^2}{\alpha^2}$

2. Run a stabilized exponential window FTF algorithm of size $N$, $N \ll L$ to update: $E_n, \hat{e}_n, A_n, B_n$

3. Compute

$$
S_n = S_{n-1} + \hat{e}_n (X_n^T X_{n-1})^{-1} \hat{e}_n N
$$

$$
\hat{e}_n = y_n - X_n^T \hat{H}_n
$$

$$
e_n = \hat{e}_n - \alpha S_n (X_n^T X_{n-1})^{-1}
$$

$$
E_n = \left( (1 - \alpha) (E_{n-1}^T)^{-1} \right)^{-1} E_n
$$

$$
A E = \left( (1 - \alpha) (E_{n-1}^T)^{-1} \right)^{-1} E_n
$$

$$
E_n^{m+1} = \frac{0}{1 - (1 - \alpha) E_n^m} + \frac{A E}{(1 - \alpha) E_n^m} \left( \begin{array}{c}
1
- A_n
\end{array} \right)
$$

$$
B E = \frac{E_n^m}{(1 - \alpha) E_n^m} \left( \begin{array}{c}
B_n
1
\end{array} \right)
$$

$$
E_n = E_n^{m+1} + \frac{0}{E_n^m} \left( \begin{array}{c}
B_n
1
\end{array} \right)
$$

$$
H_{n+1} = H_n + \alpha X_n \hat{H}_{n+1} E_n^{m+1}
$$

where vector $(\hat{v}_n)_j$ has size $j = i$ and consists of the set of the $i$th to the $j$th elements of vector $\hat{v}_n$.

4. Convergence Analysis

This section is devoted to the convergence study of the proposed algorithm. It is undertaken on an adaptive identification scheme, corresponding to the acoustic echo cancellation case. The desired scalar output $y_n$ is supposed to be the output of some "ideal" system $H^*$, fed by the same input as the adaptive filter and corrupted by additive noise $b_n$:

$$
y_n = X_n^T H^* + b_n
$$

The constant vector $H^*$ and the modelling adaptive filter $H_n$ are both assumed of length $L$, and $b_n$ is a zero mean Gaussian independent sequence with variance $\sigma^2$. All the analysis is undertaken under the assumption that the $L \times L$ matrix $X_n^T X_n$ has rank $L$.

4.1. Average tap-weight behavior

Subtracting the weight vector $H_n$ from both sides of (9), the proposed algorithm is rewritten in terms of the weight-error vector $\Delta H_n = H_n - H^*$ as:

$$
\Delta H_{n+1} = (1 - \alpha X_n R_n^{-1} \Delta H_n) \Delta H_n + \alpha X_n R_n^{-1} B_n
$$

745
Taking the mathematical expectation of both sides of (13), we may write, using the assumption made for the sequences $x_n$ and $b_n$,
\[ E[\Delta H_{n+1}] = (I - M_n) E[\Delta H_n] \]  
where $M_n = \alpha X_n R_{n}^{-1} X_n^T$. The size of matrix $M_n$ is $L \times L$ and its rank is $N$, provided the rank of $R_n$ is $N$. Therefore $L - N$ of its eigenvalues are zero. The $N$ other ones can be evaluated as follows. Let $\mu_n(i)$ for $i = 1, \ldots, N$ denote the eigenvalues matrix $M_n$. The eigenvalues of matrix $A_n = I - M_n$ are,
\[ \rho_n(i) = \begin{cases} \frac{1}{1 - \mu_n(i)} & i = 1, \ldots, N \\ \frac{1}{i - 1} & i = N + 1, \ldots, L \end{cases} \]  
It is easily shown that $\mu_n(i) \geq 0$ for $i = 1, \ldots, N$ provided that $R_n$ is nonnegative definite. Therefore, a sufficient condition for the mean convergence of the proposed algorithm is
\[ \mu_n^\text{max} \leq 2 \]  
where $\mu_n^\text{max}$ is the largest eigenvalue of $M_n$. Another, more tractable, sufficient condition for the convergence of algorithm is,
\[ \sum_{i=1}^{N} \mu_n(i) = \alpha \text{tr} (X_n^T X_n R_{n}^{-1}) \]  
Refering to (16) and (17), one can also choose
\[ 0 < \alpha \text{tr} (X_n^T X_n R_{n}^{-1}) < 2 \]  
as a sufficient condition for the convergence of algorithm in the mean, thus resulting in:
\[ 0 < \alpha < \frac{2}{\text{tr} (X_n^T X_n R_{n}^{-1})} \]  
which is more restrictive than (16).

4.2. Mean-Squared-error Behavior
Due to lack of space, we do not address the details the problem, and give the following results:
- If $0 < \alpha < \frac{2}{\text{tr} (X_n^T X_n R_{n}^{-1})}$, and provided that the input signal matrix $X_n$ has full rank, then the filter converges in the mean square.
- The following approximate relationship present the contribution of noise in the mean-square error $J_n$ for large time index $n$:
\[ J_n \approx \sigma^2 + \alpha^2 \sigma^2 \times \text{tr} (X_n^T X_{n-1} R_{n-1}^{-1} R_{n-1}^{-1} X_{n-1}^T X_{n-1}) \]  
for large $n$.

5. Simulations
This section compares the performance of the proposed algorithm, on the one hand to that of the regularized projection algorithms, and on the other hand to that of the NLMS, numerically stable PTF and FNTF algorithms in the context of acoustic echo cancellation. The system to be identified is a measured acoustic response of a conference room, of a duration of 128 ms, the sampling frequency being 16 kHz, truncated to $L = 256$ or $L = 1024$ samples, according to the length of the identifying filter. All plots show the mean-squared modelling error versus the number of iterations. For the purpose of smoothing the curves, error samples are averaged over 256 points. Fig. 1 shows a simulation run on a first speech signal, a white noise with 20dB SNR (Signal to Noise Ratio) is added to the output. This noise is subtracted from the drawn sequence in order to make more explicit the misadjustment due to noise. The adaptive filter has length $L = 256$, as well as the system to be identified, and the window size is $N = 16$. It is observed that the performance of the proposed algorithm is clearly improved over that of regularized projection algorithm.

Fig. 2 shows a simulation run on a second, longer speech signal, a white noise with 20dB SNR being added to the output. Both the adaptive filter and the system to be identified have length $L = 1624$. It is seen that the proposed algorithm converges much faster than NLMS and slightly slower than the numerically stable PTF, despite its complexity about four times smaller. The proposed algorithm converges faster than the FNTF algorithm (Fig. 2c), with a comparable complexity.

6. Conclusion
A modified projection algorithm has been introduced, which has not only the good performance of the projection algorithm, that is the fast convergence, the good tracking capability [2], but also has a good performance when the input signal is non-stationary and noise is added to the output signal (one case where the original projection algorithm has a large misadjustment). The proposed algorithm has a slightly slower convergence compared to the numerically stable PTF algorithm with an arithmetic complexity about four times smaller. Compared to the NLMS algorithm, it has superior performance with respect to convergence rate and tracking, while requiring slightly more arithmetic operations. The proposed algorithm is very similar to the FNTF algorithm proposed in [5], but seems to track better the system variations, and to converge faster for very long adaptive filters.

References
Figure 1. Comparison of various APA with speech as input, filter length $L = 256$, window size $N = 16$, $\alpha = 0.005$ and SNR=20dB: a) regularized projection algorithm (regularization parameter $\delta = 0.005\sigma_x^2$, $\sigma_x^2$ the power of $x_n$), b) new algorithm (forgetting factor $\lambda = 1 - \frac{1}{1024}$).

Figure 2. Comparison of various algorithms with speech as input, filter length $L = 1024$ and SNR=20dB: a) NLMS (adaptation gain $\mu = 0.1$, $\lambda = 1 - \frac{1}{1024}$), b) stabilized FTF (forgetting factor $\lambda = 1 - \frac{1}{1024}$), c) FNTF (length of prediction $N = 64$, forgetting factor $\lambda = 1 - \frac{1}{1024}$), d) new algorithm ($\alpha = 0.005$, window size $N = 64$, forgetting factor $\lambda = 1 - \frac{1}{1024}$).
An a priori Error Bound for the Steiglitz-McBride Method in Reduced Order Cases

Mamadou MBOUP† and Phillip A. REGALIA‡

† Université René Descartes - Paris V, UFR Mathématiques et Informatique
45, rue des Saints Pères, 75270 Paris cedex 06, France.
‡ Institut National des Télécommunications, Département Électronique et Communications
9, rue Charles Fourier, 91011 Evry cedex, France.

Abstract: In this paper, we derive a simple a priori error bound for the Steiglitz-McBride method (SMM) in undermodelled cases. The bound is obtained in the context of the Hankel-norm criterion which provides a natural interpretation of the degree of undermodelling, in the framework of model reduction. This allows us to explain the closeness of the SMM convergence points to the global minimum of the mean-square error surface, observed in many simulation studies. Examples are included.

1 Introduction

The Steiglitz-McBride method [1] has proved very useful in adaptive IIR filtering [2], [3]. One important feature of the algorithm is that it does not track output error gradients. Hence, the presence of local minima in the mean-square error (MSE) surface has no direct bearing on the convergence properties. Like most identification schemes, correct convergence applies for the sufficient order case, provided the additive output noise is white [2].

However, more realistic is the reduced order case. A priori knowledge on the unknown system is rarely available. Moreover, the systems encountered in practice are generally of very large degree, if not infinite. For example, an infinite number of parameters is required [4] to exactly model the acoustical phenomena in an echo path. In such cases, the selected McMillan degree for the adjustable model $G(z)$ is likely inferior to the unknown McMillan degree of the system $H(z)$. In this situation, perfect identification is no longer achievable and the best one can expect is that the algorithm finds the best model $G(z)$ in some sense. In $L_2$-norm, a meaningful criterion is minimizing the MSE and then, “best model” refers here to the optimal $L_2$ approximant to $H(z)$. However, the MSE surface is generally multimodal in the undermodelled case and thus, gradient based methods may easily be trapped by some local minimum. Although the Steiglitz-McBride method is free from such traps, it is not designed to find the global minimum since it has nothing to do with the minimization of the MSE. Nonetheless, it has been observed in many simulation studies [6] that the method always give a model $G(z)$, lying in a close vicinity of the global minimum point of the MSE surface.

This striking feature of the Steiglitz-McBride method has not been explained so far, unless by unproved conjecture [6]. In this paper, we derive a simple a priori error bound for the Steiglitz-McBride method in the undermodelled case. The bound is obtained from the analytical descriptions (see [7]) of the reduced order stationary points of the SMM. It is given in Theorem 1 below, in terms of the Hankel singular values of $H(z)$. A meaningful sense is thus attached to it, in the context of the Hankel-norm criterion which is widely used in model reduction problems [8]. Moreover, it is shown that the conditions for the closeness of the SMM stationary points to the global minimum of the MSE surface follow as a straightforward consequence of our a priori bound.

We recall the example given in [6] to illustrate these conditions. A comparison with the Hankel-norm approximant is also given.

2 Steiglitz-McBride Method

2.1 Model Structure

The basic scheme is depicted in figure 1, where the input $\{u(n)\}$ and output $\{y(n)\}$ sequences, from the unknown system $H(z')$, are prefiltered using a state space representation:

$$\begin{bmatrix} x(n+1) \\ w(n) \end{bmatrix} = Q(\Theta_k) \begin{bmatrix} x(n) \\ u(n) \end{bmatrix}$$

and

$$\begin{bmatrix} \xi(n+1) \\ s(n) \end{bmatrix} = Q(\Theta_k) \begin{bmatrix} \xi(n) \\ y(n) \end{bmatrix}.$$
2.2 Algorithm

Suppose the prefilter parameter vector is set to an initial value \( \Theta_0 \). Then find a vector \( b_{k+1} = [a_0^{(k+1)}, \ldots, a_M^{(k+1)}] \) and a set of \( M \) rotation angles \( \Theta_{k+1} \) to minimize the variance of an equation error signal

\[
E[\alpha(n)^2] = E\left\{ q^T(\Theta_{k+1}) \begin{bmatrix} \xi(n+1) \\ s(n) \end{bmatrix} - b_{k+1} \begin{bmatrix} x(n+1) \\ w(n) \end{bmatrix} \right)^2
\]

(4)

Then update the prefilter parameters to \( \Theta_{k+1} \) for the next experiment, and determine \( \Theta_{k+2} \). At any convergence point, the successive experiments yield \( \Theta_{k+1} = \Theta_k = \Theta \) so that the left hand branch of figure 1 reduces to

\[
\dot{y}(n) = b^T \begin{bmatrix} V_0(z) \\ V_M(z) \end{bmatrix} u(n) = G(z)u(n)
\]

(5)

while the right hand branch

\[
q^T(\Theta)Q(\Theta) \begin{bmatrix} \xi(n) \\ y(n) \end{bmatrix} = y(n)
\]

(6)

reduces to a through path. Hence, at any convergence point, equation (4) becomes an output error variance

\[
E[\alpha(n)^2] = E\{[y(n) - \dot{y}(n)]^2\}
\]

\[
= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_u(\omega)|H(\omega) - G(\omega)|^2 d\omega
\]

(7)

where \( S_u(\omega) \) is the power spectral density function of the input sequence, assumed stationary. If the input signal is a unit variance white noise, then the output error reads \( E[\alpha(n)^2] = ||H(z) - G(z)||^2 \).

This procedure is known as an off-line version of the Steiglitz-McBride method. On-line versions of the method are also available and moreover, it is shown that the set of candidate convergence points of the two variants coincide in the transfer function space \( [3],[5],[8] \). Therefore, we are free to choose whichever variant proves most convenient towards establishing theorem 1 below.

3 The a priori error bound

We assume that the unknown system is causal and stable so that its transfer function can be written as:

\[
H(z) = \sum_{k=0}^{\infty} h_k z^{-k}, \quad |z| > 1.
\]

(8)
\( H(z) \) is said to be the symbol of the associated infinite Hankel form:

\[
\Gamma_H = \begin{bmatrix}
  a_1 & a_2 & a_3 & \cdots \\
a_2 & a_3 & a_4 & \cdots \\
a_3 & a_4 & a_5 & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{bmatrix}.
\]  

The Hankel singular values, \( \{\sigma_k\}_{k=0}^{\infty} \), of \( H(z) \) are, by definition, the singular values of \( \Gamma_H \). They are set, by convention, in decreasing order: \( \sigma_1 \geq \sigma_2 \geq \cdots \), where \( \sigma_1 = \|\Gamma_H\| \) is the Hankel norm of \( H(z) \). Suppose that the adaptive filter transfer function has McMillan degree \( M \) and that the measurement errors are white, we then have:

**Theorem 1.** Let \( G(z) \) be any convergent point of the Steiglitz-McBride method. If the input sequence is white, then

\[
\| H(z) - G(z) \|_2 \leq \sigma_{M+1}. \tag{10}
\]

Note that if \( H(z) \) is rational of McMillan degree \( M \) (sufficient order case), then by the Kronecker’s theorem [9], \( \sigma_{M+1} = 0 \), resulting in an exact identification as known before [2].

### 4 Proof of Theorem 1

Rewrite equation (4) as:

\[
E[\alpha(n)^2] = [q^T(\Theta_{k+1}) - b^T_{k+1}] \mathcal{R}(\Theta_k) \begin{bmatrix}
  q(\Theta_{k+1}) \\
  -b_{k+1}
\end{bmatrix}
\]

where \( \mathcal{R}(\Theta_k) = \begin{bmatrix}
  \mathcal{R}_{\xi\xi}(\Theta_k) & \mathcal{R}_{\xi z}(\Theta_k) \\
  \mathcal{R}_{z\xi}(\Theta_k) & \mathcal{R}_{zz}(\Theta_k)
\end{bmatrix} \), with

\[
\mathcal{R}_{\xi\xi}(\Theta_k) = E\left\{ \begin{bmatrix}
  \xi(n+1) \\
  s(n)
\end{bmatrix} \begin{bmatrix}
  \xi(n+1) \\
  s(n)
\end{bmatrix}^T \bigg| \Theta = \Theta_k \right\}
\]

\[
\mathcal{R}_{\xi z}(\Theta_k) = E\left\{ \begin{bmatrix}
  \xi(n+1) \\
  s(n)
\end{bmatrix} \begin{bmatrix}
  x(n+1) \\
  w(n)
\end{bmatrix}^T \bigg| \Theta = \Theta_k \right\}
\]

\[
\mathcal{R}_{z\xi}(\Theta_k) = E\left\{ \begin{bmatrix}
  x(n+1) \\
  w(n)
\end{bmatrix} \begin{bmatrix}
  \xi(n+1) \\
  s(n)
\end{bmatrix}^T \bigg| \Theta = \Theta_k \right\}
\]

\[
\mathcal{R}_{zz}(\Theta_k) = E\left\{ \begin{bmatrix}
  x(n+1) \\
  w(n)
\end{bmatrix} \begin{bmatrix}
  x(n+1) \\
  w(n)
\end{bmatrix}^T \bigg| \Theta = \Theta_k \right\}
\]

The minimization of eq. (11) with respect to \( b_{k+1} \) gives:

\[
R_{zz}(\Theta_k) b_{k+1} = R_{\xi z}(\Theta_k) q(\Theta_{k+1}). \tag{12}
\]

Substituting (12) into (11) gives the reduced error surface \( E[\alpha(n)^2] = q^T(\Theta_{k+1}) R_{\xi\xi}(\Theta_k) q(\Theta_{k+1}) \) where the Shur complement

\[
R_{\xi\xi}(\Theta_k) \triangleq R_{\xi\xi}(\cdot) - R_{\xi z}(\cdot) R_{zz}(\cdot) R_{z\xi}(\cdot) \bigg| \Theta = \Theta_k
\]

is positive defined [10]. Since \( q(\Theta_{k+1}) \) is a unit norm vector, the reduced error is then seen as a Rayleigh quotient of \( R_{\xi\xi}(\Theta_k) \). The minimum point is therefore attained if and only if \( q(\Theta_{k+1}) \) is an eigenvector associated to the minimum eigenvalue \( \lambda_{\text{min}} \{ R_{\xi\xi}(\Theta_k) \} \) of \( R_{\xi\xi}(\Theta_k) \), and the following lemma holds:

**Lemma 1.** The candidate convergence points of the Steiglitz-McBride method are characterized in lattice form by

\[
\mathcal{R}_{\xi\xi}(\Theta) q(\Theta) = \lambda_{\text{min}} \{ \mathcal{R}_{\xi\xi}(\Theta) \} q(\Theta).
\]

The minimum mean square output error (MSOE) is

\[
E[\{ y(n) - \hat{y}(n) \}^2] = \lambda_{\text{min}} \{ \mathcal{R}_{\xi\xi}(\Theta) \}
\]

Further characterizations of the convergence points are given in [7] where the minimum MSOE was also obtained as a Rayleigh quotient:

\[
E[\{ y(n) - \hat{y}(n) \}^2] = q^T(\Theta) C^T(\Theta) \Gamma_H^2 C(\Theta) q(\Theta)
\]

where \( C(\Theta) \) is an \( \infty \times (M+1) \) matrix whose \( \ell \)th column is the vector obtained from the impulse response of the function \( V_{\ell-1}(z^{-1}) V(z) \). It is not difficult to verify that these functions are causal and orthogonal, and hence we have \( C^T(\Theta) C(\Theta) = \Gamma_{M+1}^2 \). The two expressions (13) and (14) of the minimum MSOE are reconciled by the following lemma, proved in [10]:

**Lemma 2.** \( \mathcal{R}_{\xi\xi}(\Theta) = C^T(\Theta) \Gamma_H^2 C(\Theta) \).

Finally, from \( C^T(\Theta) C(\Theta) = \Gamma_{M+1} \) and setting the eigenvalues of \( \mathcal{R}_{\xi\xi}(\Theta) \) in decreasing order, we have from the Poincaré’s separation theorem [11]:

\[
\lambda_k \{ \mathcal{R}_{\xi\xi}(\Theta) \} = \lambda_k \{ C^T(\Theta) \Gamma_H^2 C(\Theta) \} \leq \lambda_k \{ \Gamma_H^2 \} = \sigma_k^2,
\]

for \( k = 1, \ldots, M+1 \). If we set \( k = M+1 \), we have:

\[
\| H(z) - G(z) \|_2^2 = \lambda_{\text{min}} \{ \mathcal{R}_{\xi\xi}(\Theta) \} \leq \sigma_{M+1}^2
\]

to complete the proof of Theorem 1.

### 5 Closeness to optimal \( L_2 \) and to Hankel-norm approximants

Let \( G_1(z) \), \( G_2(z) \) and \( G_3(z) \) be three stable and causal rational functions of McMillan degree \( M \), denoting respectively any convergence point of the Steiglitz-McBride method, the global minimum point of the MSE surface and the Hankel-norm approximant to \( H(z) \).

It is well known that \( \| H(z) - G_2(z) \|_2 \leq \sigma_{M+1} \) (see e.g. [12]) and from the remarkable result of AAK [9]:

\[
\min_{\| \Gamma_H - \Gamma_G \| = \sigma_{M+1}} \| \Gamma_H - \Gamma_G \| = \sigma_{M+1}
\]

we have \( \| H(z) - G_3(z) \|_2 \leq \| H(z) - G_2(z) \|_2 = \sigma_{M+1} \).

Therefore, the three approximants \( G_1(z) \), \( G_2(z) \) and \( G_3(z) \), all lie within a \( \sigma_{M+1} \)-ball in \( L_2 \) norm to \( H(z) \). As such, if \( \sigma_{M+1}/\sigma_1 \) is small, the three approximants are obviously close to each other. On the other hand, they may be quite different if the ratio \( \sigma_{M+1}/\sigma_1 \) is close to one. This appears clearly in the following example.
We consider the example studied by Fan et al. in [6] for the second order transfer function:

$$H(z) = \frac{0.05z^{-1} - 0.4z^{-2}}{(1 - 0.8303z^{-1})(1 - p_1 z^{-1})} - 1 < p < 1,$$

with a first order approximant. Significant discrepancies between $G_1(z)$ and $G_2(z)$ were observed in [6] in the range $-0.85 < p < -0.6$. Figure 2 shows the Hankel singular values spread $\sigma_2/\sigma_1$ versus the pole $p$. Note that the range $-0.85 < p < -0.6$ corresponds precisely to the values of $p$ for which $\sigma_2/\sigma_1$ is close to one. For example, for $p = -0.786$, we have $||H(z)||_2 = 0.5279$, $\sigma_1 = 0.6984 \ldots$ and $\sigma_2 = 0.6955 \ldots$ yielding a very poor relative error in Hankel-norm. This indeed suggests that $H(z)$ cannot be well approximated by a first order transfer function.

Figure 2: Hankel singular values spread $\frac{\sigma_2}{\sigma_1}$ vs pole $p$.

On the other hand, for $p = 0.3011$, we have $||H(z)||_2 = 0.8633$ and $\sigma_2/\sigma_1 = 0.4694/1.8505 = 0.2637 \ll 1$. For this case, the three approximants given by:

$$G_1(z) = \frac{-0.3111z^{-1}}{1 - 0.9057z^{-2}}; \quad G_2(z) = \frac{-0.3110z^{-2}}{1 - 0.9058z^{-1}};$$

$$G_3(z) = \frac{-0.3310z^{-1}}{1 - 0.9059z^{-1}},$$

are very close to each other, with $G_1(z)$ and $G_2(z)$ indistinguishable up to the fourth significant digit.

6 Conclusion

The a priori error bound of Theorem 1 is the first formal evidence that the Steiglitz-McBride method is capable of satisfactory system approximation properties for the undermodelled case. The bound is expressed in terms of the Hankel singular values of the unknown system. This provides a natural criterion for gauging the degree of undermodelling. In particular, a significant value of the ratio $\sigma_{M+1}/\sigma_1$ suggests that the unknown system cannot be well approximated by an $M^{th}$ order transfer function. On the other hand, a small value of $\sigma_{M+1}$ implies that the unknown system is "close" to an $M^{th}$ order transfer function so that any convergence point of the SMM yields a "good" approximant to the system. Moreover, this approximant must be close to both the corresponding Hankel norm and $L_2$ norm approximants. We should remark that, although the a priori error bound of Theorem 1 is a very encouraging result, it applies in its present form only for white noise inputs. A more general bound, taking into account spectral weighting for correlated inputs, would of course be desirable. Further descriptions of the convergence points of the SMM are developed in [7] in terms of spectral weighting.

References


A Cyclic Method for Signal-Selective DOA Estimation

Giacinto GELLI, Luciano IZZO and Luigi PAURA

Dipartimento di Ingegneria Elettronica, Università di Napoli Federico II,
via Claudio, 21 I-80125 Napoli, Italy, Tel/Fax: +39-81-7683156/7683149,
E-Mail: izzo@nalis.dis.unina.it

Abstract. The problem of estimating the directions of arrival of multiple wavefronts impinging on a sensor array is addressed. The proposed technique belongs to the class of the methods (cyclic methods) that are able to discriminate in favor of the signals of interest against noise and interfering signals by exploiting their different cyclostationarity properties. The advantage of using the new method is twofold: a relative ease of implementation with respect to previously proposed cyclic high-resolution methods, and a significant resolution improvement over the cyclic beamforming technique, which exhibits a comparable complexity.

1. Introduction

The estimation of the directions of arrival (DOA's) of multiple wavefronts radiated from remote sources and impinging on an array of sensors is of great interest in a number of applications, including radar, sonar, communications, geophysics, and astronomy. The recently proposed cyclic approach to the problem (see [1] and references therein) assumes that the signals of interest (SOI's) exhibit cyclostationarity [2], a property that arises mainly from periodic processing operations (e.g., sampling, modulation, coding) performed on the SOI's. Shortly, a signal exhibits cyclostationarity with a cycle frequency $\alpha$ if it is correlated with a possibly conjugate frequency-shifted version of itself for a value $\alpha$ of the frequency shift.

The methods based on the new approach (cyclic methods) can automatically discriminate in favor of the SOI's against noise and interfering signals on the basis of their known different cyclostationarity properties. This signal selection capability allows cyclic methods to operate correctly in many situations where conventional (i.e., non cyclic) methods, which assume a stationary model for the SOI's, perform poorly or fail. In fact, unlike conventional methods, cyclic methods can work well in environments with unknown or changing noise characteristics, and operate successfully also when the number of impinging signals is greater than the number of sensors. Moreover, the presence of (possibly correlated) interfering signals does not affect the performances of the cyclic methods, as long as the cyclostationarity properties of SOI's and interfering signals are different. The only information required by all cyclic methods is the a priori knowledge or estimation of a cycle frequency common to all SOI's that is not shared by noise and interfering signals.

A number of the cyclic algorithms so far proposed, such as Cyclic MUSIC [3], Cyclic ESPRIT [4] and the method proposed in [5] (which can be regarded as an extension of the Pisarenko-Berni method), are generalizations of well-known conventional techniques. Such methods perform high-resolution signal-selective DOA estimation on the basis of the eigendecomposition of the array cyclic correlation matrix (see Section 3) of the received signals. However, they are not able to operate properly when highly correlated SOI's are present. The cyclic least squares method [6] overcomes such a limitation at the cost of a time-consuming multidimensional search performed over the signal DOA's.

The recently proposed cyclic beamforming method [7] is a non eigenstructure-based technique that can represent a feasible solution to the DOA estimation problem whenever both signal selectivity and ease of implementation are primary purposes and, moreover, the capability of working in a multipath propagation scenario is needed. Furthermore, unlike the previously mentioned eigenstructure-based cyclic methods, the cyclic beamforming technique does not require the a priori knowledge of the number of SOI's. Its major drawback is the low angular resolution of the SOI's sharing the same cycle frequency, which can be improved only by increasing the number of sensors.

In order to improve resolution while retaining the computational simplicity of the cyclic beamforming, the present paper proposes a new non eigenstructure-based method for DOA estimation. Such a method, which is based on a constrained minimization procedure that requires only one matrix inversion, can achieve an angular resolution considerably finer than the one exhibited by the cyclic beamforming technique. Moreover, since it is not based on the eigendecomposition of the array cyclic correlation matrix, the proposed method does not require any a priori knowledge about the number of SOI's. Its main drawback with respect to the cyclic beamforming technique is that a high degree of correlation among the SOI's can lead to a significant performance degradation or even to a complete failure.

2. Background on Cyclostationarity

A scalar complex time series $s(t)$ is said to contain
a first-order periodicity, i.e., a finite additive sine wave component, with frequency \( \alpha \neq 0 \) if and only if the infinite time average

\[
M_\alpha \triangleq \langle s(t) e^{-j2\pi \alpha t} \rangle,
\]

which represents the complex amplitude of the sine wave component, is finite but not zero.

When \( s(t) \) does not exhibit first-order periodicities, in many cases there exists an absolutely integrable kernel \( k(\cdot, \cdot) \) such that at least one of the time series

\[
y_1(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s(\lambda) k(-\lambda, t-\mu) s^*(\mu) d\lambda d\mu,
\]

\[
y_2(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s(\lambda) k(-\lambda, t-\mu) s(\mu) d\lambda d\mu,
\]

resulting from quadratic time-invariant transformations, contains a first-order periodicity with frequency \( \alpha \), namely, the spectrum of \( y_1(t) \) and/or \( y_2(t) \) presents spectral lines at frequencies \( f = k \omega + \omega \alpha \). In such a case \( s(t) \) is said to contain a second-order periodicity. More specifically, \( s(t) \) exhibits cyclostationarity [cyclostationarity] with cycle frequency \( \alpha \) if \( y_1(t) \) [\( y_2(t) \)] contains a first-order periodicity with frequency \( \alpha \). Cycle frequencies of interest for communication signals include keying rates and their harmonics, doubled carrier frequencies, and sums and differences of these.

It can be shown [2] that \( s(t) \) exhibits cyclostationarity with cycle frequency \( \alpha \) if and only if the cyclic autocorrelation function

\[
R_{s^2}(\tau) \triangleq \langle s(t+\tau/2) s^*(t-\tau/2) e^{-j2\pi \alpha t} \rangle
\]

exists and is not zero for some values of the lag parameter \( \tau \). Analogously, \( s(t) \) exhibits cyclostationarity if and only if such a property holds for the cyclic conjugate correlation function

\[
R_{s^2}(\tau) \triangleq \langle s(t+\tau/2) s(t-\tau/2) e^{-j2\pi \alpha t} \rangle.
\]

It is worthwhile to note that, for a given value of \( \tau \), \( R_{s^2}(\tau) \) represents the complex amplitude of the sine wave component contained in the lag product time series \( s(t+\tau/2) s^*(t-\tau/2) \) [\( s(t+\tau/2) s(t-\tau/2) \)].

Finally, if one deals with vectors of complex-valued time series, two matrices are necessary for the second-order characterization. More specifically, given an \( M \)-column vector \( x(t) \), the \( M \times M \) cyclic correlation matrix \( R_{XX}(\tau) \) and the \( M \times M \) cyclic conjugate correlation matrix \( R_{X\bar{X}}(\tau) \) must be considered:

\[
R_{XX}(\tau) \triangleq \langle x(t+\tau/2) x^*(t-\tau/2) e^{-j2\pi \alpha t} \rangle,
\]

\[
R_{X\bar{X}}(\tau) \triangleq \langle x(t+\tau/2) x^*(t-\tau/2) e^{-j2\pi \alpha t} \rangle,
\]

where the superscripts \( H \) and \( \bar{\cdot} \) denote Hermitian [conjugate transpose] and transpose operations, respectively.

3. The Proposed DOA Estimation Method

Consider a passive array consisting of \( M \) isotropic sensors and assume that \( D_o \) SOIs, exhibiting (possibly conjugate) cyclostationarity with a common cycle frequency \( \alpha \), impinge on the array. The radiating sources are located in the far-field of the array so that the wavefronts received by the array can be modeled as planewaves. If one also assumes that sensors and sources are coplanar, then the position of each source is described by a single parameter, i.e., the direction of arrival of the planewave.

If the frequency band of interest is sufficiently narrow so that the amplitude and phase modulations of any SOI do not change significantly during the time it takes for the wavefront to propagate across the array and the sensor characteristics are essentially constant across the bandwidths of the SOI's, then the \( M \)-column vector \( x(t) \) of the received analytic signals can be expressed as

\[
x(t) \approx A(\theta) s(t) + i(t).
\]

In eq. (8), \( x(t) \) is the \( D_o \)-column vector of the zero-mean analytic SOI's impinging on the array, \( s(t) \) is the \( M \)-column vector of the zero-mean analytic signals modeling noise and interference and, finally, \( A(\theta) \) is the \( M \times D_o \) matrix of the steering vectors \( a(\theta_k) \) \( k = 1, 2, \ldots, D_o \), with \( \theta_k \) denoting the DOA of the \( k \)th SOI. The functional dependence of the steering vectors on the DOA's is assumed to be known and unambiguous.

Under the assumptions that the interfering and noise signals are uncorrelated with the SOI's and do not exhibit (possibly conjugate) cyclostationarity with the considered cycle frequency \( \alpha \), the array cyclic correlation matrix (ACCM) [array cyclic conjugate correlation matrix (ACCCM)] of the received signal vector is given by

\[
R_{XX}(\tau) \triangleq A(\theta) R_{s^2}(\tau) A^H\theta(\theta),
\]

where the superscript [\( \ast \)] denotes conjugation to be considered in the conjugate cyclostationarity case only, and \( R_{XX}(\tau) \) is the \( D_o \times D_o \) cyclic correlation matrix of the SOI's. Such a result allows one to predict, for the cyclic approach (based on the proprieties of an estimate of \( R_{XX}(\tau) \)), satisfactory performances even in strongly adverse interference environments. In particular, the cyclic methods can provide a good accuracy also when the interfering signals, whose number can be greater than the number of sensors, exhibit an arbitrary degree of correlation among themselves and arrive from directions arbitrarily close to those of the SOI's. Moreover, the DOA estimation techniques based on the ACCM or the ACCCM do not require the knowledge of the noise statistics, which is generally required in the conventional methods.

To introduce the proposed method, let us consider the lag-product matrix of the received signals

\[
L_{XX}(\tau) \triangleq x(t+\tau/2) x^H(t-\tau/2),
\]

and let us perform a linear spatial filtering to discriminate SOI's coming from DOA's different from the look direction \( \phi \). It results that

\[
y(t) = L_{XX}(\tau) w^*(\phi),
\]

where \( w(\phi) \) is an \( M \)-column weight vector that steers the array towards the direction \( \phi \).

The weight vector \( w(\phi) \) is singled out by minimizing the squared magnitude of the sine wave component with frequency \( \alpha \) contained in \( y(t) \) under the constraint that the gain in the look direction be unity, that is,

\[
P(\phi) \triangleq \min \| \langle y(t) e^{-j2\pi \alpha t} \rangle \|^2 \quad \text{subject to} \quad a^H(\phi) w(\phi) = 1,
\]

\[
w(\phi),
\]

\[
\text{subject to} \quad a^H(\phi) w(\phi) = 1,
\]

\[
w(\phi),
\]

\[
\text{subject to} \quad a^H(\phi) w(\phi) = 1.
\]

\[
\text{subject to} \quad a^H(\phi) w(\phi) = 1.
\]
where $| \cdot |$ denotes the Euclidean norm of a vector. The solution of this minimization problem, easily obtainable by the Lagrange multiplier technique, yields

$$P^\alpha(\phi) = \frac{1}{\alpha^H(\phi)[R_{z|x}(\tau) R_{z|x}^\alpha(\tau)]^{-1}}$$

(13)

This functional can be viewed as an estimator of the squared magnitude of the sinewave component with frequency $\alpha$ contained in the lag-product of the SOI coming from the direction $\phi$. As a consequence, the locations of the maxima of $P^\alpha(\phi)$ can be taken as DOA estimates of the signals exhibiting (possibly conjugate) cyclostationarity with cycle frequency $\alpha$. Note that the value of $\tau$ in eq. (10) must be chosen appropriately according to the cyclostationarity features of the SOI's, in order to assure that the magnitude of the sinewave component with the considered cycle frequency $\alpha$ be not zero.

Obviously, for practical implementation $R_{z|x}^\alpha(\tau)$ must be replaced in eq. (13) by an estimate based on $N$ snapshots. The estimate considered in the simulation experiments is

$$\hat{R}_{z|x}^\alpha(\tau) = \frac{1}{N} \sum_{i=1}^{N} z(t_i + \tau) z^H(t_i) e^{j2\pi \alpha(t_i + \tau/3)}$$

(14)

Finally, it is worthwhile to note that in the proposed method signals exhibiting cyclostationarity with cycle frequency $\alpha$ and coming from directions different from the look direction are handled as interfering signals. Therefore, their contribution to the magnitude of the spectral line with frequency $\alpha$ is minimized. Consequently, when these signals are significantly correlated with the one coming from the look direction, the considered minimization also reduces or annihilates the contribution due to the signal arriving from the look direction. In other words, the method fails when a high degree of correlation among the SOI’s is present.

4. Simulation Results

In this section, results of simulation experiments, aimed at substantiating the effectiveness of the new method and comparing it with the cyclic beamforming technique, are reported. The experiments assume a passive uniform linear array of $M = 5$ isotropic sensors with interelement spacing equal to one half of the SOI carrier wavelength. Unless otherwise specified, the noise is additive white Gaussian noise which is uncorrelated from sensor to sensor, the SOI’s and the interfering signals are uncorrelated BPSK signals transmitted with Nyquist-shaped pulses using a 100% excess bandwidth, and all SOI’s and noise and interfering signals have the same average power. The estimate of $R_{z|x}^\alpha(\tau)$ given in eq. (14) was performed on the basis of $N = 8192$ samples of the received signals and the value $\tau = 0$, which is appropriate for the modulation format of the SOI’s, was selected.

The first experiment illustrates the signal selectivity capability of the proposed algorithm. More precisely, it is shown that the new method allows one to discriminate two SOI’s with the same DOA but different cyclostationarity properties. The SOI’s have DOA’s $\theta_1 = \theta_2 = 0.2$ rad., carrier frequency $f_0 = f_1$ (with $f_2$ the sampling frequency) and baud rates $\alpha_1 = 0.125 f_1$ and $\alpha_2 = 0.175 f_1$. Moreover, the signal-to-noise ratios (SNR’s) for the two SOI’s are $\text{SNR}_1 = 60 \text{dB}$ and $\text{SNR}_2 = -30 \text{dB}$, respectively. The proposed technique is able (Figure 1) to correctly estimate each signal DOA, one at a time, by working, subsequently, at the cycle frequencies $\alpha = \alpha_1$ and $\alpha = \alpha_2$ with the estimate of $R_{z|x}^\alpha(\tau)$.

In the second experiment, the performance of the proposed method for an overloaded array is tested. A total of seven signals impinge on the five-element array. Two SOI’s, with the same carrier frequency $f_0 = f_1$ and the same baud rate $\alpha_0 = 0.125 f_1$, arrive from $\theta_1 = -0.3$ rad. and $\theta_2 = 0.3$ rad., and the AM interfer ing signals are present with carrier frequency $f_1 = f_2$, bandwidth $0.08 f_1$, and the following DOA values (in rad.): $-0.4, -0.2, 0.0, 0.2, 0.4$. Since the SOI’s exhibit cyclostationarity with $\alpha = \alpha_0$ but the AM signals do not, the proposed method can correctly estimate $\theta_1$ and $\theta_2$ by working (Figure 2) with the estimate of $R_{z|x}^\alpha(\tau)$ with $\alpha = \alpha_0$.

In the third experiment, it is shown that the proposed method, unlike all non-cyclic methods, can work well independently of the spatial noise characteristics. Two SOI’s, with the same carrier frequency $f_0 = f_2$ and baud rate $\alpha_0 = 0.125 f_2$, arrive from $\theta_1 = -0.2$ rad. and $\theta_2 = 0.3$ rad. The noise signals are correlated from sensor to sensor, according to a correlation matrix whose $ij$th element is given by $0.8 | i - j |$. Satisfactory performances (Figure 3) for the proposed method are obtained working with $R_{z|x}^\alpha(\tau)$ with $\alpha = 2 f_0$.

In the fourth experiment, the sensitivity of the pro-
posed method to the degree of correlation among the SOI's is tested. Two correlated SOI's are present, with DOA's $\theta_1 = -0.3 \text{ rad.}$ and $\theta_2 = 0.3 \text{ rad.}$ and the same carrier frequency $f_0 = f_3$ and baud rate $c_0 = 0.125 f_3$. Moreover, one interfering signal is present, with DOA $\theta_1 = 0.0 \text{ rad.}$, carrier frequency $f_1 = 0.99 f_3$, and baud rate $c_1 = c_0$. Figure 4 presents curves relative to the proposed method (solid lines) working with $R_{2,2}(\tau)$ with $\alpha = 2 f_0$ for three values of the correlation coefficient $\rho$ between the SOI's. For comparison purposes, the figure also shows the results for the cyclic beamforming method, working with the ACCM with $\alpha = 2 f_0$ in the case of fully correlated SOI's ($\rho = 1$). Note that in Figure 4 (as well as in Figure 5) the values of the cyclic beamforming function $F_B(\phi)$ (see [7]) are reported on the right-hand ordinate scale, whereas those referring to the proposed method are given on the left-hand scale. The results show that the new method yields almost correct estimates for low and moderate values of the correlation coefficient, while it completely fails when the degree of correlation is high. On the contrary, the cyclic beamforming technique, as expected, provides accurate DOA estimates even when the SOI's are fully correlated.

Finally, the last experiment is aimed at showing the resolution gain achievable by the proposed technique with respect to the cyclic beamforming. Two uncorrelated SOI's and one interfering signal are present, whose carrier frequencies and baud rates are the same as in the previous experiment. Figure 5 presents simulation results for the proposed and cyclic beamforming methods, both working with $R_{2,2}(\tau)$ with $\alpha = 2 f_0$. More specifically, the solid lines refer to the case where the SOI’s arrive from DOA’s $\pm 0.05 \text{ rad.}$, while the dotted ones refer to the case where the signal DOA’s are $\pm 0.10 \text{ rad.}$ In either case, the interference DOA is fixed at $\theta_I = 0.0 \text{ rad.}$ The results show that the new method exhibits a good estimation accuracy even when the SOI’s are very closely spaced. The cyclic beamforming technique, vice versa, yields very inaccurate estimates when the DOA separation is $0.2 \text{ rad.}$ and, moreover, fails completely (exhibiting a single peak) when the SOI’s are only $0.1 \text{ rad.}$ spaced apart. The interference rejection capability of the proposed technique is also fully confirmed: no peak is present in correspondence of $\phi = \theta_I = 0.0 \text{ rad.}$

References


THE POTENTIAL GAIN IN USING SPECTRAL INFORMATION IN PASSIVE LOCALIZATION OF SOURCES

Hagit Messer

Department of Electrical Engineering - Systems, Tel Aviv University, Tel-Aviv 69978, Israel
E-mail: messer@eng.tau.ac.il

Abstract. In this paper we study the potential performance gain in passive localization of sources using an array of sensors due to the use of spectral information about the source signals. By using the Cramer Rao lower bound on the localization error, we compare the achievable performance of 5 types of localization algorithms, which differ from each other by the amount of spectral information they employ. We show that by not using complete knowledge of the source spectra, the performance loss is most significant in high resolution conditions. However, as long as the number of sensors in the array is larger than the number of sources to be located, the performance gain of using partial spectral information is very limited and hardly significant.

1. Introduction

In classical array processing the aim is to extract spatial information using samples of the tempo-spatial field. We assume that the field consists of wideband signals, radiated from point sources and that the objectives is to estimate their locations. Although the radiated signals are of no interest, knowledge of their waveshape, or their statistics, is essential for applying optimal localization algorithms. We assume that the source signals are stationary, zero mean Gaussian processes, so their statistics is summarized by their cross-spectral matrix.

Consider an array composed of \( M \) sensors with a known arbitrary geometry, sampling a wavefield generated by \( N \) wideband sources at locations \( \theta_1, \theta_2, \ldots, \theta_N \) with an overlapping spectra of bandwidth \( B \), in the presence of additive noise. The signal received at the \( m \)-th sensor can be described by

\[
\mathbf{s}_m(t) = \sum_{n=1}^{N} s_n(t - \tau_m(\theta_n)) + n_m(t),
\]

for \( m = 1, \ldots, M \), where \( s_n(\cdot) \) is, in the passive case, an unknown signal radiated by the \( n \)-th source, \( n_m(\cdot) \) is the additive noise at the \( m \)-th sensor, and \( \tau_m(\theta_n) \) is the propagation delay associated with the \( n \)-th source and the \( m \)-th sensor.

The observation time interval \( T_0 \) is sectioned into \( K \) subintervals of duration \( T_d \) each, and a Discrete Fourier Transform is applied to each subinterval. We may write

\[
\mathbf{X}_k(f_j) = \mathbf{A}_k(f_j) \mathbf{G}_k(f_j) + \mathbf{N}_k(f_j),
\]

for \( j = 1, \ldots, J, k = 1, \ldots, K \), where

\[
\mathbf{X}_k(f_j) = [X_{1k}(f_j), X_{2k}(f_j), \ldots, X_{Kk}(f_j)]^T,
\]

\[
\mathbf{G}_k(f_j) = [S_{1k}(f_j), S_{2k}(f_j), \ldots, S_{Nk}(f_j)]^T,
\]

\[
\mathbf{N}_k(f_j) = [N_{1k}(f_j), N_{2k}(f_j), \ldots, N_{Nk}(f_j)]^T,
\]

\[
\mathbf{A}_k(f_j) = [g_{\theta_1}(f_j), g_{\theta_2}(f_j), \ldots, g_{\theta_N}(f_j)],
\]

\[
g_{\theta_n}(f_j) = e^{-j2\pi f_j \tau_n(\theta_n)}, e^{-j2\pi f_j \tau_2(\theta_n)}, \ldots, e^{-j2\pi f_j \tau_N(\theta_n)}.
\]

Here \( X_{nk}(f_j), S_{nk}(f_j) \) and \( N_{nk}(f_j) \) are the Discrete Fourier coefficients of \( \mathbf{s}_n(t), \mathbf{G}_n(f) \) and \( \mathbf{N}_n(f) \), respectively, at the \( k \)-th subinterval and frequency \( f_j \). \( A \) is the \( M \times N \) direction matrix, and the vectors \( g_{\theta_n}(f) \) are referred to as the steering vectors of the array. Under standard assumptions (e.g., [2,7]), the Fourier coefficients at different frequencies are approximately uncorrelated and so are the coefficients derived at different subintervals, so the frequency domain covariance matrix of the received data is block-diagonal. The wideband log-likelihood function of the data is then

\[
\mathbf{L}^G = -K \sum_{j=1}^{J} \left\{ \log|\mathbf{R}_c(f_j)| + \text{tr}\left(\mathbf{R}_c^{-1}(f_j)\right) \right\},
\]

where \( \mathbf{R}_c(f_j) \) is the data covariance matrix at frequency \( f_j \)

\[
\mathbf{R}_c(f_j) = \mathbf{A}_k(f_j) \mathbf{R}_f(f_j) \mathbf{A}_k^H(f_j) + \mathbf{N}(f_j),
\]

and \( \mathbf{R}_f \) is the frequency domain sample covariance matrix at the frequency \( f_j \)

\[
\mathbf{R}_f = \frac{1}{K} \sum_{k=1}^{K} \mathbf{X}_k(f_j) \mathbf{X}_k^H(f_j).
\]

In this paper we present results of a study in which we compare the best achievable localization performance using different levels of prior spectral information. We do that by comparing the Cramer-Rao lower bound (CRLB) on the estimation error of the bearing (DOA) of a source where the received data consists of radiation from two or more sources in unknown location in spatially white noise. The bound is derived under different assumptions on the source signals cross-spectral matrix: from the extreme cases where this matrix is assumed to be completely known/unknown via intermediate cases where the signals are assumed to be uncorrelated and/or of flat spectra. If less prior spectral information is used the localisation procedures are, usually, simpler, but their performance is expected to be reduced, relative to the case where the source spectra are completely known.

We assume that the number of sensors is larger than the number of sources to be located (\( M > N \)), and we quantify the achievable localization accuracy of different types of algorithms which employ different levels of prior spectral knowledge. If, however, the number of sources in the field is not smaller than the number of sensors, it can be shown [5] that if absolutely no spectral prior is used, the array cannot resolve more sources than sources. The use of prior spectral information (complete knowledge of the spectrum at the extreme case or even rough knowledge about the smoothness of the spectra) makes it possible to resolve more sources than sensors.
2. The bounds

The Cramer-Rao bound (CRB) on the covariance of any unbiased estimate of a parameter vector \( \theta \) is given by \( \text{CRB}(\theta) = J^{-1} \), where \( J \) is the Fisher information matrix (FIM) [1]. If the observation vector, \( z \), is a zero-mean, complex Gaussian vector, then the entries of the FIM are given by (e.g., [3]):

\[
J_{\theta_i \theta_j} = \text{trace} \left\{ \frac{\partial R_x(\theta)}{\partial \theta_i} R_x^{-1}(\theta) \frac{\partial R_x(\theta)}{\partial \theta_j} R_x^{-1}(\theta) \right\}
\]  
(6)

where \( R_x(\theta) = E\{ z z^T \} \) is the covariance matrix of the observation vector \( z \). In our problem, due to the block diagonal nature of the covariance matrix of the observation in the frequency domain, the FIM entries can be written as:

\[
J_{\theta_i \theta_j} = \sum_{k=1}^{\sqrt{2}} J_0(f_k) = \sum_{k=1}^{\sqrt{2}} \text{trace} \left\{ \frac{\partial R_x(f_k)}{\partial \theta_i} R_x^{-1}(f_k) \frac{\partial R_x(f_k)}{\partial \theta_j} R_x^{-1}(f_k) \right\}
\]  
(7)

where \( R_x(f_k) \) is given in (4). The unknown parameter vector in our problem can be written as \( \theta^T = [\theta_1^T, \theta_2^T] \) where \( \theta_1 \) contains all unknown spatial parameters (e.g., source bearings and ranges) and \( \theta_2 \) contains all unknown spectral parameters. Notice that the FIM is determined by the frequency domain data covariance matrix, which depend on \( \theta_1 \) via the matrix \( A \) only, and on \( \theta_2 \) via the matrix \( R_x \) only. Using this partition, the matrix \( J_0(f_k) \) can be written as:

\[
J_0(f_k) = \begin{bmatrix}
J_1(f_k) & J_{12}(f_k) \\
J_{12}(f_k) & J_2(f_k)
\end{bmatrix}
\]

where:

\[
J_1(f_k)_{ij} = \text{trace} \left\{ \frac{\partial R_x(f_k)}{\partial \theta_i} R_x^{-1}(f_k) \frac{\partial R_x(f_k)}{\partial \theta_j} R_x^{-1}(f_k) \right\}
\]  
(8)

\[
J_{12}(f_k)_{ij} = \text{trace} \left\{ \frac{\partial R_x(f_k)}{\partial \theta_i} R_x^{-1}(f_k) \frac{\partial R_x(f_k)}{\partial \theta_j} R_x^{-1}(f_k) \right\}
\]  
(9)

\[
J_2(f_k)_{ij} = \text{trace} \left\{ \frac{\partial R_x(f_k)}{\partial \theta_i} R_x^{-1}(f_k) \frac{\partial R_x(f_k)}{\partial \theta_j} R_x^{-1}(f_k) \right\}
\]  
(10)

The CRB on the estimation errors of the spatial parameters, which are the only parameters of interest in our problem, is given by [3]:

\[
\text{CRB}(\theta_1) = \{ J_1 - \Delta J \}^{-1} \quad \text{with} \quad \Delta J = 0
\]  
(11)

\( \Delta J \) summarizes the role of the unknown spectral parameters on the achievable estimate error of the spatial parameters. If all spectral parameters are completely known, \( \Delta J \equiv 0 \). Otherwise, it is a non-zero non-negative definite matrix which causes the diagonal elements in the CRB(\( \theta_1 \)) to be smaller than in the case where sources are completely known.

Some well known localization algorithms do not employ source spectral information, even if it is available (e.g., the MUSIC, the deterministic maximum likelihood, [2],[8]). Our aim is to study the performance losses resulted from not using such information. We compare the CRB of (11) for 5 different cases of assumed prior spectral information, which represent the best achievable localization performance of an estimator which employs such prior. The 5 cases are:

1. The matrix \( R_x(f_k), k = 1, ..., J \), is completely known. In that case there are no unknown spectral parameters and \( \Delta J = 0 \).

2. Sources are assumed uncorrelated with flat spectra. Here \( R_x(f_k) \) is a diagonal matrix and its diagonal terms are \( S_k(f_k) = S_{0_k}, k = 1, ..., J \). Therefore, \( \theta_1 = [S_1, S_2, ..., S_N]^T \) is an \( N \)-dimensional vector.

3. The matrix \( R_x(f_k), k = 1, ..., J \), is only assumed to be a diagonal matrix. In that case the sources are assumed uncorrelated, but their individual spectrum is arbitrary. The vector \( \theta_2 \) is an \( N \times J \)-dimensional vector.

4. The matrix \( R_x(f_k) \) is assumed independent of \( f_k \), so sources are assumed to have flat spectra/cross spectra. That is, \( S_k(f_k) = S_{mm}, k = 1, ..., J, n = 1, ..., N \). Therefore, \( \theta_2 \) is an \( N \times J \)-dimensional vector.

We used MATLAB routines for deriving the CRB on the variance of an unbiased estimate of \( \theta_1 \) under the assumed model of the 5 different cases. In the next section we present numerical examples and we discuss the results.

3. Numerical examples - results and conclusions

The bounds for the 5 different cases were numerically derived and were plotted, using the following notation: (1) solid line; (2) dashed line; (3) dotted line; (4) dash-dot; (5) solid line. We assumed 2 sources (\( N=2 \)), and a uniform, linear array, where the separation between successive sensors is \( d \). The sources have equal power so the signal to noise ratio (SNR) in each sensor is the same for both of them. The sources are assumed as wideband signals with flat spectra (the signal cross-spectra matrix is independent on the frequency), centered at \( \omega_0 = \pi c/d \). The sources are located in the far-field. The first is in the broadside of the array (\( 0^\circ \)) and the bearing of the second varies.

In Figure 1 we plot the 5 CRBs on the bearing of the first source as a function of the bearing of the second source for uncorrelated sources with SNR, at each of the 10 frequency bins, of -27 dB (Fig. 1a) and -15 dB (Fig. 1b). The array consists of \( M=6 \) sensors. The upper bound on the achievable performance - the case where no spectral information is employed - and the lower bound on the achievable performance - the case where exact and complete spectral information is employed - are the two solid lines which are clearly noticeable. The difference between them vanishes when the separation between the sources is larger than the beamwidth of the array beampattern. Within its mainlobe, this difference is larger as the SNR decreases. This observation is not new (e.g., [4]). However, it is interesting to see that the use of partial spectral knowledge can hardly contribute to the localization performance: the dashed line, which represents the performance of an algorithm which uses the knowledge that the sources are uncorrelated and that their spectra is flat (but not their spectral level), is noticeable but is very close to the upper bound. The other curves, which represent the achievable performance of algorithms that use less prior knowledge (either the fact that the sources are uncorrelated, or that their spectra is flat) are coincide with the upper bound and are hardly noticeable. In Figures 2a and

\[ \text{The routines are based on M-files written by Dr. M. Doron} \]
2b, where the same bounds are depicted for a smaller array (M=3), the bounds related to the partial spectral knowledge are more noticeable, but still their distance from the upper bound is insignificant.

Next, we consider the case of partially correlated source signals of flat spectra/cross-spectra. In this case only 3 bound are relevant: the one uses all spectral information (1), the one uses only the fact that the matrix $R_{xx}$ is independent of $f_{2}$ (4), and the one that uses no prior spectral information (5). In Figure 3a the three curves are plotted where the correlation coefficient between the source signal is 0.95 and the SNR of each of the sources, at each of the 10 frequency bands, is -30dB, M=3. As in [6] and as in the case of uncorrelated sources, when source separation is smaller than the beamwidth, the difference between the two extreme cases is significant. The middle case, however, practically coincide with the upper bound. For the results of Figure 3b we enlarged the bandwidth to 50 frequency bands and we increased the correlation coefficient to 0.99. Here, the middle curve is more noticeable, but is still close to the upper bound, everywhere but in an extremely low source separation.

The following conclusions can be made:

Under the assumption that the number of sensors in the array, M, is larger than the number of sources in the temporal/spatial field, N, the achievable performance of an algorithm which employs complete knowledge of the spectra is a lower bound on the achievable localization performance while the achievable performance of an algorithm which employs no knowledge of the spectra is an upper bound on the achievable localization performance, as expected. All other cases of localization algorithms which use different levels of partial spectral information falls between these two bounds. However, the performance gain of using spectral information is potentially significant mainly when the separation between sources is smaller than the beamwidth of the array beam-pattern, a situation which is known as an 'high resolution' situation. Even in such conditions, the potential gain is significant mostly in the non-realistic case where the source spectra in completely known. If everything is known but the source levels, the performance gain - even in very low SNR and/or high correlation conditions - is very limited. This suggests that the main factor in improving localization performance in high resolution conditions is the use of prior knowledge of the power of the sources rather than the shape of their spectra/cross-spectra. This conjecture implies also that a successful localization algorithm should involve power measurements, at least in 'high resolution' conditions [4]. Present and future study of the problem discussed in this paper check this conjecture, study the effect of spectral modelling errors on source localization accuracy and the effect of the unbiassness requirement on the mean square error of the estimate, using means different than the CRB.

4. References


Figure 1: Cramer Rao lower bound on the source localization error under different assumptions about its spectrum. 2 uncorrelated sources of equal SNR, equally spaced linear array of 6 sensors. (a) SNR=-27dB; (b) SNR=-15dB.

758
Figure 2: Cramér Röo lower bound on the source localization error under different assumptions about its spectrum. Two uncorrelated sources of equal SNR, equally spaced linear array of 3 sensors. (a) SNR=-24dB; (b) SNR=-12dB.

Figure 3: Cramér Röo lower bound on the source localization error under different assumptions about its spectrum. Two correlated sources of equal SNR=-30dB, equally spaced linear array of 3 sensors. (a) Correlation coefficient = 0.95, bandwidth=10; (b) Correlation coefficient = 0.99, bandwidth=50.
Analysis of LMS–Like Source Tracking
for Robust Adaptive Beamforming

Sofiène Affes, Saeed Gazor, and Yves Grenier

Ecole Nationale Supérieure des Télécommunications
Télécom Paris, Département Signa, 46 rue Barrault, 75034 Paris Cedex 15, France.
E-mail: Affes@sig.ensit.fr, Tel:+331 45817782, Fiz:+331 45887935.

Abstract. The performance of a new algorithm for robust adaptive beamforming via LMS-like target tracking, recently presented in a former work, is analyzed in this paper. It is first proven the necessity of the beamforming tracker to be in a locking range of the tracked source. Mean and covariance convergence proofs and analyses are then made. The stability condition, the convergence time, the misadjustment, and the optimal choice of the adaptation step-size are given. The new proposition we make along this analysis agree with our intuitive expectations. The corresponding theoretical results are all confirmed by simulations.

1. Introduction

To combat the source signal cancellation phenomena encountered in beamforming when in the presence of localization errors [1–4], we presented in [5,6] a robust adaptive beamformer based on target tracking. This new proposed algorithm was shown to gain robustness through an efficient reduction of localization errors, using classical beamforming with a time-corrected and adapted steering vector. In addition, it has a computational complexity of order $O(m)$ where $m$ is the number of sensors, and can be implemented easily.

In this paper, we make a performance and convergence analyses of the new robust adaptive beamformer presented in [5,6], in the case of a linear array and far field propagation.

We first prove the pointing of the beamformer in a locking range of the source to be a necessary initial condition for convergence. This involves the presence of no other sources or jammers in that defined location interval.

We therefore give the iterative equation in time of the localization error, and prove its convergence in the mean sense when in the presence of spatially diffuse noise. We also establish the iterative equation in time of the quadratic localization error, and show its convergence in the covariance sense.

We thus infer the stability condition and the time constant of the algorithm. In addition, we derive the expression of misadjustment and its lower bound with the corresponding optimal adaptation step-size. We finally define the holding range into which the source is assigned to move, so as to avoid the locking break during the source tracking. All the above theoretical results are confirmed by simulations.

2. Mathematical Formulation

We consider the following model of a plane wave propagating signal received by a linear array (see Figure 1):

$$X_t = G_t s_t + N_t,$$  \hspace{1cm} (1)

$$G_t = F(\theta_t),$$  \hspace{1cm} (2)

$$F(\theta) = e^{-jr} [e^{-j\kappa x_{1}}, e^{-j\kappa x_{2}}, \ldots, e^{-j\kappa x_{m}}]^{T},$$  \hspace{1cm} (3)

where $X_t$ is the m-dimensional observation vector, $s_t$ is the desired narrowband signal to be extracted, $N_t$ is an additive noise vector, and $G_t$ is the transfer function (i.e. steering vector) between the emitted source $s_t$ and the m-sensor antenna array. $F$ is a parameterizing function where $	heta = [\kappa \tau]^{T}$. The wavenumber $\kappa = 2\pi \sin(\phi) / \lambda$ where $\phi \in [-\pi, \pi]$ is the DOA, and $\lambda$ is the wave length. $x_{1}, x_{2}, \ldots, x_{m}$ are the sensor positions. $\tau$ represents the phase delay from the origin to within about an integer multiple of $2\pi$, and is obviously restricted to $[0, 2\pi]$. All the signals considered herein are complex, and the subscript $t$ stands for time index.

As well, we make the further assumptions that:

A1: $G, N$ and $s$ are mutually independent.

A2: Channel $G$ is slowly time varying in comparison to the variations of $N$ and $s$. Hence, we are able to estimate $G$ properly, then update it.

A3: On each relatively short time interval, $G$ has a Gaussian distribution with mean $\hat{G}$ and autocorrelation matrix $R_{G}$, both slowly time varying in comparison to the variations of $N$ and $s$.

A4: A possibly erroneous approximation of $\theta_0$, say $\hat{\theta_0}$, is initially provided either by an approximate a priori guess, or by a given localization technique.

A5: $N$ is a white noise with zero mean Gaussian distribution and autocorrelation matrix $R_{N}$. For the sake of clarity and simplicity, we avoid the formulation of the problem in the presence of jammers.

3. Proposed algorithm

Given these assumptions, our approach is to constantly correct the steering vector $\hat{G}$, and time-adapt the classical
beamformer say $W_t$ to the new look-direction. The robust adaptive beamforming proposed in [5,6] can be then summed up by the following steps (see Figure 1):

$$W_t^{n+1} = W_t^n - \eta \frac{1}{N_t} X_t^n \operatorname{diag}[\hat{G}_{t-1}^H] X_t^n,$$

$$\hat{G}_t = \hat{G}_{t-1} + \mu (X_t - \hat{G}_{t-1} s_t) s_t^H,$$

$$\hat{s}_t = W_t^H X_t - W_t^H \left[ P \operatorname{diag}(\hat{G}_{t-1}^H) X_t \right],$$

$$\hat{G}_t = F \left[ \hat{s}_t, \hat{G}_{t-1} \right]^T,$$

where $\operatorname{Im}(\cdot)$ denotes the complex imaginary part, and:

$$P \triangleq \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

$$K \triangleq \begin{bmatrix} \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n \sum_{j=1, j\neq i}^n x_i x_j \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1}.$$  

$P$ is a $(m-1) \times m$ matrix (remark: $P \operatorname{diag}(\hat{G}_{t-1}^H) \hat{G}_{t-1} = 0$).

The iterative equation (5) consists of correcting the steering vector $\hat{G}_{t-1}$ by a gradient stochastic term depending on the classical beamforming output. The target coordinates are then extracted in (6) by a projection of $\hat{G}_t$ over the array manifold. The estimated parameters are finally used in (7) to reconstruct the steering vector of the classical beamformer $\hat{G}_t$ for enhanced signal extraction. The beamformer should thus provide a robust source signal extraction and an efficient method for source tracking. It should be noted that all the steps above involve a number of operations proportional to the number of sensors $m$.

### 4. Mean Convergence analysis

As stated at the end of section 2, an initial steering vector $\hat{G}_0$ is assumed available. Let us assume the corresponding wavenumber $\hat{s}_0$ to be corrupted by an initial estimation error $\Delta \hat{s}_0$ where $\Delta \hat{s}_0 \triangleq \hat{s}_0 - s_0$. In [5], we consider the optimistic case in which the noise is zero. Then it is shown that:

$$1 - 4\mu |s_0|^2 \leq \frac{|\Delta \hat{s}_0|}{|\Delta \hat{s}_0 - \hat{s}_0|} \leq 1.$$  

Hence, the time series $\Delta \hat{s}_n$ converges to zero necessarily when the upper bound factor is smaller than one (remark: we assume that $4\mu |s_0|^2$ is smaller than one and strictly positive with a non zero probability). As it can be seen, when the step-size $\mu$ is in the stability range, $\Delta \hat{s}_n$ converges to the stability point at zero if the initial point verifies the following condition:

$$\Delta \hat{s}_0^2 \leq \frac{12m \sum_{i=1}^m x_i^2}{m} + 3 \left( \sum_{i=1}^m x_i^2 \right)^2 \leq \frac{12m}{m} = 12.$$

We must note that $\Delta \hat{s}_0$ is only depending on the array sensor positioning $x_1, x_2, \ldots, x_m$, and is smaller than the main lobe width of the array. We can also see from the plot of the antenna directivity that $\Delta \hat{s}_n$ must be always smaller than the main lobe width (first null of the array pattern of the array), so as to let the algorithm correct the localization error which make sure that $\hat{s}_n - s_0$ converges to $s_0$. This proves that the main lobe width and $\Delta \hat{s}_n$ are tightly related. Hence, we should assign $\Delta \hat{s}_0$ to be initially in a locking range arbitrarily fixed to $\Delta \hat{s}_n$. Given this condition, we expect $\hat{s}_n$ to be relatively close to $s_0$. The foregoing implies that the necessary condition for convergence even if noise is not zero, is that the localization error be in a locking range actually reducing with increasing noise power.

Now assuming the error $\Delta \hat{s}_n$ to be in the locking range, we will analyze the convergence of the global algorithm. To do so, let us suppose that the target motion to be a white random process independent of $s_1, N_t$ and $\theta$, with a time increment defined by $s_1 = s_{n+1} - s_1$ (e.g. for an immobile source, $s_1 = 0$). In particular, we consider the convergence of the localization error $\Delta \hat{s}_n = \hat{s}_n - s_0$, using the first order Taylor series expansion of $F(\theta)$ around $\theta_{n-1}$ and assuming $W_t^n = 0$ (case of Delay-Sum), the expression of $\Delta \hat{s}_n$ can be derived as follows [see [5] for more details]:

$$\Delta \hat{s}_n = \left[ 1 + \mu \sum_{i=1}^m x_i \operatorname{det} K \right] \left[ \frac{1}{m} \sum_{i=1}^m x_i \operatorname{Im}(F^H(\theta_{n-1}) D N_t s_0^H) \right]$$

$$+ \mu \sum_{i=1}^m x_i \operatorname{det} K \left[ \left( \frac{1}{m} \sum_{i=1}^m x_i \right) s_0^H \right]$$

$$+ \left( \frac{1}{m} \sum_{i=1}^m x_i \right) s_0^H,$$  

where $D \triangleq \operatorname{diag}(\hat{s}_n^H) = \operatorname{diag} \left[ \left( \frac{1}{m} \sum_{i=1}^m x_i \right) s_0^H \right]$. It can be seen in this equation that $\Delta \hat{s}_n$ is the output of an iterative linear time-varying system. The value of $\mu$ should be carefully selected, so that the maximum absolute angular speed always remains in a holding range.

To study the performance of the algorithm, we consider the case of a white noise where $R = \sigma n^2 I$. This major assumption is not actually very restrictive, as long as possibly other point jammers can be present far enough from the desired source. We hence assume an additional hypothesis stating that no other sources are present within the locking range of the desired one. The role of $W_t^n$ assumed here to be 0, is precisely to cancel such point jammers. Using now (11) and the statistical properties of the process, it can be proven that (see [5] for more details):

$$E[\hat{s}_n] = E[\hat{s}_{n-1} - \mu \sigma_0^2] + \mu \sigma_0^2 E[\hat{s}_n].$$  

(12)
where \( \sigma^2 \) is the variance of \( w_i \). So \( \mu \) must satisfy the stability condition \( \mu < \frac{1}{2\sigma^2} \). In this case, the mean of the wavenumber estimate is the output of a first order AR lowpass filter when the target wavenumber is at the input. Equation (12) says that the mean of the wavenumber estimate is equal to the temporal mean of \( E[\kappa_i] \), and proves that the algorithm is able to track the target in the mean. Consequently, the time constant of this filter is given by:

\[
\tau_{E[\kappa]} = \frac{1}{\mu \sigma^2}. \tag{13}
\]

This expression shows that the algorithm is faster with increasing values of the step-size \( \mu \) (see Figure 2)\(^1\).

5. Covariance Convergence Analysis

In [5], taking the origin at the array center and assuming \( E[w_i] = 0 \), the covariance of localization error in \( \Delta \kappa_i \) is derived for moving targets as follows:

\[
E[\Delta \kappa_i] = E[\Delta \kappa_{i-1}] \left\{ 1 - 2\mu \sigma^2 + \mu^2 \left( 2\sigma^2 + \frac{\sigma^2 N}{2m} \right) \right\} + \frac{\mu^2 \sigma^2}{2\det K} (\sigma_N^2 + m \sigma^2) + \sigma^2, \tag{14}
\]

where \( \sigma^2 \) is the variance of \( w_i \). This equation states that the localization error is also the output of a first order AR filter. The stability condition of the algorithm is:

\[
\mu < \frac{1}{2\sigma^2 + \frac{\sigma^2 N}{2m}}. \tag{15}
\]

A second projection method presented in [5,6], gives practically the same performance, but remains however stable for a larger step-size range (see Figure 5). Under such a condition, the localization error in the stationary environment will converge with a time constant equal to:

\[
\tau_{E[\Delta \kappa]} = \frac{1}{2\mu \sigma^2 \left[ 1 - \mu \left( \sigma^2 + \frac{\sigma^2 N}{2m} \right) \right]} \tag{16}
\]

The expression of \( \tau_{E[\Delta \kappa]} \) shows the convergence time of the algorithm to be very small when the step-size \( \mu \) is very small.

\(^1\)All the simulations are performed with \( \sigma_1^2 = 1, \sigma_2^2 = 0.1, m = 16, \lambda = 0.5 \) and the sensor positions \( x_i = i - 8.5 \).

Figure 2. Convergence time of localization error compared with the theoretical value \( \tau_{E[\kappa]} = \frac{1}{\mu \sigma^2} \).

Figure 3. Standard deviation of localization error (theoretical and experimental) for different values of \((S/N)\) in dB.

or when it is close to the stability bound \( (\sigma^2 + \sigma^2 N)^{-1} \). Moreover, the best time constant is obtained when \( \mu = \frac{1}{2} (\sigma^2 + \sigma^2 N)^{-1} \). Hence, we prove:

\[
\tau_{E[\Delta \kappa]} > 1 + \frac{1}{4m(S/N)}, \tag{17}
\]

where \((S/N) = \sigma^2 / \sigma^2_y\) is the signal to noise ratio. This inequality shows that the convergence time of the algorithm improves with a higher signal to noise ratio or sensors number.

Regarding the misadjustment of the localization error (i.e. the steady state covariance of \( \Delta \kappa \)), equation (14) implies that:

\[
E[\Delta \kappa_i] = \frac{\sigma^2}{2\mu \sigma^2 + \mu \sigma^2 N + \frac{\sigma^2 N}{2m} + \sigma^2}, \tag{18}
\]

This equation justifies the intuitive expectation regarding target tracking capability relation with \((\sigma^2 / \sigma^2_y)\), \(S/N\), \(\mu\) and sensor positioning \(x_i\). It particularly proves the covariance to be increasing with noise input power (see Figure 3) and/or \( \frac{\sigma^2}{\sigma^2_y} \) ratio (see Figures 4 and 5).

So, in the steady state, the covariance of localization error converges to the sum of two terms respectively resulting from the variations of noise \( \sigma^2 \) say \( E_{\text{no}}[\Delta \kappa_{\sigma^2}] \), and the variations of source position \( \sigma^2 \), say \( E_{\text{no}}[\Delta \kappa_{\sigma^2}] \). Clearly, the first term \( E_{\text{no}}[\Delta \kappa_{\sigma^2}] \geq \frac{\mu \sigma^2 N + \sigma^2}{4 \sum_{i=1}^{m} x_i^2} \) shows the localization error even for immobile targets (see Figure 3). The second term \( E_{\text{no}}[\Delta \kappa_{\sigma^2}] = \tau_{E[\Delta \kappa]} \sigma^2 \) is always greater than \( \frac{1}{2m} \sigma^2 \) in the case of a moving target. Thus, the convergence time will be reflected on the tracking behavior of the algorithm by a permanent estimation delay which increases the localization error in \( \kappa \).

These terms show the effect of the step-size in the compromise to be found between a faster convergence rate and a smaller misadjustment of steady state localization. Under the stability condition (15), a lower bound for the mis-
adjustment of localization error can be easily found:

$$E[\Delta \kappa^2_0] \geq \frac{\sigma_N \sigma_w}{\sigma_s} \sqrt{\frac{1 + \frac{1}{m} (S/N)}{2 \sum_{i=1}^{m} \xi_i^2}}.$$  \hspace{1cm} (10)

The equality is approximately obtained when the step-size value is:

$$\mu_{\text{opt}} \approx \frac{\sigma_w}{\sigma_s \sigma_N} \sqrt{\frac{2 \sum_{i=1}^{m} \xi_i^2}{1 + 1/m(S/N)}},$$  \hspace{1cm} (20)

on condition that $\sigma_w^2 \left( \sum_{i=1}^{m} \xi_i^2 \right) \left( (S/N) + \frac{1}{m} \right) \leq 0.5$.

As shown in Figures 4 and 5, for greater values of $\mu$ when $0.5 < \mu < 1$, and under the stability condition, the error covariance $\text{Cov}_{\Delta \kappa}(\Delta \kappa)$ is proportional to $\mu^2$ instead of $\mu$, as expected by the exact theoretical expression. Hence, an optimum step-size value can be computed to find the best compromise between the tracking speed and the localization error.

Equation (19) also shows that the lower bound of the performance improves when the signal to noise ratio increases or when the variations of source position decrease. Another result is that the performance of the algorithm for target tracking depends on the dispersion of sensor positions, or equivalently on the array geometry.

Given these results for the stability and the optimality of the algorithm, the DOA variations should be however in a holding range defined below, so that the target remains locked. Hence, the algorithm is able to track the target if the following condition is fulfilled:

$$\frac{\sigma_N \sigma_w}{\sigma_s} \sqrt{\frac{1 + \frac{1}{m} (S/N)}{2 \sum_{i=1}^{m} \xi_i^2}} < \beta \Delta \kappa^2,$$

where $\Delta \kappa$ is the locking range defined earlier and $\beta$ is a given constant. This condition requires the source to move inside the locking range, within a time interval at least smaller than the optimal convergence time.

6. Conclusion

We analyzed in this paper the robust adaptive beamforming algorithm presented in [5,6].

It was first shown that the localization error, even for the initialization, must be kept in a given locking range.

The proofs of convergence in the mean and the covariance were made at a given stability condition over the adaptation step-size. The established expression of the location misadjustment proves that the target tracking improves to some extent when the signal to noise ratio increases, or when the variations of source position decrease. It also shows that the performance depends on the dispersion of the sensor positions, or equivalently on the array geometry. Moreover, it reveals that the convergence rate is limited by the output SNR.

The generalizations of the present algorithm to the multitarget tracking and the wideband case are studied [7,8].

References

A Parametric Approach to Extended Source Localization

Shahrokh VALAEI\textsuperscript{1,1}, Peter KABAL\textsuperscript{1,1}, and Benoit CHAMPAGNE\textsuperscript{1}

\textsuperscript{1}Dept. of Elect. Eng., McGill University, Montreal, Quebec, CANADA H3A 2A7
\textsuperscript{1}INRS-\textsuperscommunications, Universit\textsupersital du Qu\textsupersit\textsupersrite, Verdun, Quebec, CANADA H3E 1H6

Abstract: Point source modeling is frequently used in array processing. Although this assumption is good for many applications, there are some situations where point source modeling is unrealistic. For instance, in a multi-beam echo sounder, a reflected signal from the sea floor appears as a spatially extended source. In this paper we investigate distributed sources. The approach is based on the assumption that the correlation kernel of the distributed source belongs to a family of parametric functions. We generalize the MUSIC algorithm to a distributed signal parameter estimator (DSPE). The DSPE localizer minimizes a scalar product between an estimated basis for the noise subspace and the array manifold. We study two cases corresponding to completely correlated and totally uncorrelated signal distributions. We also discuss limitations of the application of ordinary beamformer techniques to spatially distributed signals by computing the array gain. It is shown that the array gain is upper bounded by a value which depends on the extension width of the source. Thus increasing the number of the sensors in a beamformer does not necessarily increase the resolution.

1. Introduction

A common assumption in array processing is that the signals are generated by sources that are highly localized in space (point sources). This modeling of a source is not a valid assumption in some applications. In a multi-beam echo sounder the penetration of the signal into the sea floor and reflection from different layers creates a spatially distributed source [1]. In radar, if the target is spread in range, the reflection of the signal from the target is observed as a spatially extended signal [2]. When a signal is distributed in space, it is usually modeled as a spatially colored noise [3]. The spatial extension of signal might occur in other applications such as reflection of sound in a reverberant room and communications through the reflection from the ionosphere and the troposphere.

In past, extended sources were treated as a cluster of point sources [1]. For such a model the dimensionality of the signal subspace grows with the number of point signals. To obtain a unique solution for this model, the number of point signals (the dimension of the signal subspace) should be smaller than the number of sensors (the dimension of the observation space). Thus, if the number of sensors is not large enough, the clustered point sources approximation will not provide a precise model of a distributed source.

Recently, we have presented a new method for the localization of spatially distributed sources [4]. This method is based on a parametric model for the spatial correlation function of the signal and has been applied to localization of coherently distributed (CD) and incoherently distributed (ID) source. The present paper is a continuation of this previous work. In this paper we provide additional details about the derivation of the method. In particular, we show that the new method is a generalization of the MUSIC algorithm. Furthermore, we study the performance of a conventional beamformer for distributed sources by finding the array gain for a uniform linear array.

2. Distributed source modeling

Assume a scenario with \( q \) spatially distributed narrowband sources arriving at an array of \( p \) sensors. The array output vector in the frequency domain can be formulated by

\[ x = \sum_{i=1}^{q} \int_{-\pi}^{\pi} a(\theta) s_i(\theta; \psi_i) d\theta + n, \tag{1} \]

where \( a(\theta) \) is the \( p \times 1 \) noise vector, \( n \) is the \( p \times 1 \) noise vector, and \( s_i(\theta; \psi_i) \) is the angular signal density of the \( i \)-th source which is also a function of the angle of arrival \( \theta \) and the parameter vector \( \psi_i \). The two limits of the directions-of-arrival (DOAs) for a uniform source distribution, and the angle of maximum power and the -3 dB extension width for a bell-shaped distribution are examples of the parameters.

The time samples of the noise vector are modeled as zero-mean, independent, circular, complex Gaussian random variables and uncorrelated from the signals. It is assumed that the noise is spatially white. The white noise assumption can be relaxed if the correlation matrix of the noise is known but for a scalar. With these assumptions the correlation matrix of the array output is given by

\[ R_x = \sum_{i=1}^{q} \sum_{j=1}^{q} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} a(\theta)a(\theta') s_i(\theta; \psi_i) s_j(\theta'; \psi_j) d\theta d\theta' + \sigma_n^2 I \tag{2} \]

where \( \sigma_n^2 \) is the noise power and

\[ p_{ij}(\theta; \theta'; \psi_i, \psi_j) = E[s_i(\theta; \psi_i) s_j^*(\theta'; \psi_j)] \tag{3} \]

is the angular cross-correlation kernel parameterised in terms of the unknown parameter vectors \( \psi_i \) and \( \psi_j \). The superscripts \( H \) and \( * \) represent the Hermitian transposition and the complex conjugation, respectively. In this paper we assume that the sources are uncorrelated from each other which results in a simplification of \( p_{ij}(\theta; \theta'; \psi_i, \psi_j) \) to

\[ p_{ij}(\theta; \theta'; \psi_i, \psi_j) = p(\theta; \theta'; \psi_i) \delta_{ij} \tag{4} \]

where \( \delta_{ij} \) is the Kronecker delta and the same parametric model is used for all sources. This constraint can be relaxed if we model correlated sources as a single source with a new parametric correlation kernel which is the addition of the angular cross-correlation kernel of the sources.

If the signal components corresponding to different directions-of-arrival (DOAs) are uncorrelated, the angular cross-correlation kernel can be further simplified to

\[ p(\theta; \theta'; \psi) = p(\theta; \psi) \delta(\theta - \theta') \tag{5} \]
where $\delta(\theta)$ is the Dirac delta function. We refer to such a source as incoherently distributed (ID). This model can be applied to scattering media.

In some situations the signal components corresponding to different DOAs might be completely correlated with each other. Such a signal is called coherently distributed (CD) and has a separable kernel given by

$$p(\theta, \theta'; \psi) = n_i g(\theta; \psi) g^*(\theta'; \psi)$$

where $n_i$ is a scalar representing the power of the $i$th source observed at the reference point of the array, the superscript $*$ shows the complex conjugation, and $g(\theta; \psi)$ is a complex deterministic angular signal density defined in the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and normalized according to

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} g(\theta; \psi) d\theta = 1.$$  

(7)

In practice, the signal components at different angles might be partially correlated. Partially correlated signals can be localized using the same method as for ID signals.

3. The array gain

Beamformers improve the array output SNR by steering a beam in the direction of a signal. Because of the ease of implementation, these methods are practically important. However, they have relatively low resolution. A large number of sensors must be used in a conventional beamformer to achieve a high resolution. For point sources the array gain can be improved by increasing the number of sensors. Here, we show that for distributed sources the spatial correlation function of the signal is upper bounded by an exponentially decreasing function. Then, we derive the array gain and show that it is bounded and thus cannot increase linearly with the number of sensors. For the specific case of CD sources, we show that the array gain attains a maximum and then decreases exponentially as the number of sensors becomes very large.

The gain of an array of sensors is defined as the ratio of the SNR at the array output to the SNR at a single sensor [9]. Assuming that the noise is spatially white and that a conventional beamformer is used, the array gain is given by

$$G_a = \frac{\mathbf{R}_a \mathbf{n}^H}{\mathbf{R}_a} \mathbf{R}_a$$

(8)

where $\mathbf{n}$ is the location vector of the array steered towards the direction of interest and $R_a$ is the correlation matrix of the array output in a noise-free environment.

3.1. The CD source case

Assume that the array output can be observed along a continuous linear aperture and denote the observation at point $z$ by $x(z)$. For a single CD source in a noise-free environment we have

$$x(z) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} e^{j2\pi z \sin \theta} g(\theta; \psi) d\theta$$

(9)

where $\gamma$ is a zero-mean complex Gaussian random variable and $g(\theta; \psi)$ is the normalized deterministic angular signal density. Assuming that the source is uniformly distributed around $\theta_0$, i.e.,

$$g(\theta; \psi) = \begin{cases} \frac{1}{2\Delta}, & |\theta - \theta_0| \leq \Delta, \\ 0, & \text{otherwise}, \end{cases}$$

the observation vector can be written as

$$x(z) = \gamma \frac{1}{2\Delta} \int_{\theta_0}^{\theta_0+\Delta} e^{j2\pi z \sin \theta} d\theta.$$  

(11)

For a small $\Delta$, it is straightforward to show that

$$x(z) \approx \gamma e^{j2\pi z \sin \theta_0} \sin(\frac{2\Delta}{\lambda} \cos \theta_0).$$

From (12) we arrive at the following result.

Property 1. For a uniform CD source with a small extension, the spatial correlation function at the points $z_1$ and $z_2$ in a noise-free environment is bounded by

$$|E[x(z_1)x^*(z_2)]| \leq K z_1^{-1} z_2^{-1}$$

(13)

where $K$ is a positive scalar.

An example of the correlation between two points on a linear array for a uniform CD source is depicted in Fig. 1. It is assumed that the first point is the phase reference of the array. The second point varies along the array. The envelope of the correlation function exponentially decreases with the separation between the two points. Thus, as the aperture length of the array increases, the correlation between widely separated sensors decreases and the corresponding signals cannot be coherently added to increase the SNR. This suggests that the array gain does not increase linearly with the number of sensors.

For a uniform linear array with half the wavelength spacing between sensors, the component of the observation (12) at the position of the $l$th sensor is

$$x_l = \gamma e^{j2\pi l \sin \theta_0} \sin(l \Delta \cos \theta_0).$$

(14)

Assuming that the power of the source is unity and $\theta_0 = 0$, the array gain is given by

$$G_a = \frac{1}{p} \left( \sum_{l=0}^{p-1} \sin(l \Delta) \right)^2.$$  

(15)

Note that for $\Delta = 0$ the array gain is equal to $p$ which is the gain of a point source scenario. For $\Delta \neq 0$ and large $p$, the sum in (15) is approximated by $\pi/2$ which reveals that the array gain decreases with a rate of $1/p$ which is revealed to the array gain decreases with a rate of $1/p$. The array gain for a CD source as a function of the number of sensors $p$ is depicted in Fig. 2. The array gain has a maximum which depends on the extension width. Increasing the number of the sensors beyond the maximum point decreases the array gain. We have found that at the maximum point, the array length $p_{\text{MAX}}$ can be approximated as

$$p_{\text{MAX}} \Delta \approx 40$$

(16)

where $\Delta$ is the extension width measured in degrees.

3.2. The ID source case

For an ID source the spatial cross-correlation function at the two observation points $z$ and $z'$ is given by [4]

$$E[x(z)x^*(z')] = \frac{1}{\Delta} e^{j2\pi (z - z') \sin \theta_0} \sin(\frac{2}{\lambda} (z - z') \Delta \cos \theta_0).$$  

(17)

From this we can easily arrive at the following result.

Property 2. For a uniform ID source with a small extension, the spatial correlation function decreases exponentially with the separation and is upper bounded by

$$|E[x(z)x^*(z_2)]| \leq K_{z_1}^{-1} z_2^{-1}$$

(18)

where $K$ is a positive scalar.

Since the spatial correlation function decreases exponentially with distance, the array gain cannot increase linearly with the number of sensors. For a uniform linear array with half the wavelength spacing between sensors, the spatial cross-correlation function between the $l$th and the $k$th sensors is given by

$$E[x_l x_k^*] = e^{j\pi (-l - k) \Delta} \sin[(l - k) \Delta \cos \theta_0].$$

(19)

Assuming that $\theta_0 = 0$, the array gain is given by

$$G_a = \frac{1}{p} \left( \sum_{l=0}^{p-1} \sum_{k=0}^{p-1} \sin[(l - k) \Delta] \right).$$

(20)
Again, it is seen that for $\Delta = 0$, we obtain the same result as a point source case. With a change of variable the array gain can be represented as

$$G_a = \frac{1}{p} \left[ p + 2 \sum_{r=1}^{p-1} (p - r) \text{sinc}(r\Delta) \right].$$  \hspace{1cm} (21)

The array gain for an ID source is depicted in Fig. 3. For a fixed extension width, the maximum array gain for the ID source is higher than that for the CD source.

4. A generalization of MUSIC

Let us denote by $L^2(-\frac{\pi}{2}, \frac{\pi}{2})$ the Hilbert space of all complex valued square integrable functions defined over the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$. The inner product and the norm in this space are defined by

$$\langle s_i, s_j \rangle_c = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} s_i^*(\theta) s_j(\theta) d\theta$$  \hspace{1cm} (22)

$$||s_i||_c = \sqrt{\langle s_i, s_i \rangle_c}$$  \hspace{1cm} (23)

where the subscript $c$ indicates the continuous nature of the waveform. According to (1), the observation vector $X$ at the array output can be expressed as

$$X = \sum_{i=1}^{q} Ls_i(\cdot; \psi_i) + n$$  \hspace{1cm} (24)

where $L$ is a linear operator that maps $L^2[-\frac{\pi}{2}, \frac{\pi}{2}]$ into a complex observation vector space $G^p$ with dimensionality $p$ such that

$$L: L^2[-\frac{\pi}{2}, \frac{\pi}{2}] \rightarrow G^p$$  \hspace{1cm} (25)

$$Ls = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} a(\theta)s(\theta)d\theta.$$  \hspace{1cm} (26)

The inner product and the norm in $G^p$ are defined by

$$\langle x_i, x_j \rangle_d = x_i^H x_j$$  \hspace{1cm} (27)

$$||x_i||_d = \sqrt{\langle x_i, x_i \rangle_d}$$  \hspace{1cm} (28)

where the subscript $d$ indicates the discrete nature of the function.

The adjoint operator $L^*: G^p \rightarrow L^2[-\frac{\pi}{2}, \frac{\pi}{2}]$ satisfies

$$\langle L^*s, x \rangle_c = \langle s, L^*x \rangle_c.$$  \hspace{1cm} (29)

For the linear operator (26), we have

$$\langle Ls, x \rangle_d = [L^*H]x = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} x^*(\theta) a^H(\theta)d\theta x = \langle s, a^Hx \rangle_c.$$  \hspace{1cm} (30)

Thus the adjoint is given by

$$L^*X = a^H(\theta)X.$$  \hspace{1cm} (31)

As a starting point, we extend the definition of the signal and noise subspaces to distributed sources. Note that the source signal $s_i(\theta; \psi_i)$ in (24) is a random signal which is also a function of the DOA $\theta$ and the parameter vector $\psi_i$. By the source subspace we mean the span of all realizations of the source signals $s_i(\theta; \psi_i), i = 1, \ldots, q$ where the $\psi_i$s are fixed. This subspace is denoted by $S$ and is defined as

$$S = \text{Span}\{s_i(\theta; \psi_i) : i = 1, \ldots, q, \text{ and all realizations}\}.$$  \hspace{1cm} (32)

The source subspace $S$ is a subspace of $L^2[-\frac{\pi}{2}, \frac{\pi}{2}]$. The range of the linear operator $L$ under $S$ is the signal subspace and is represented by

$$R = \{Ls : \text{ all } s \in S\}.$$  \hspace{1cm} (33)

The orthogonal complement of $R$ is the noise subspace and is denoted by $R^\perp$. It can be shown that the range of the adjoint operator $L^*$, when the domain is restricted to the noise subspace $R^\perp$, is the orthogonal complement of $S$ which is represented by $S^\perp$. The above concept of the signal and noise subspaces can be reconciled with the conventional definitions for the point source case [5].

We now derive a MUSIC type algorithm for distributed source parameter estimation. In [4] we have defined the concept of the effective dimension of the signal subspace. For distributed sources the signal component might extend to the whole observation space. For such signals the dimensionality of the signal subspace is $p$ the number of sensors. The effective dimension of the signal subspace has been defined as the dimension of the subspace which contains 95 percent of the total energy of the signal. For localization of distributed sources we use the effective dimension of the signal subspace to estimate the noise eigenvectors.

Suppose that $R^\perp$ has dimension $p - q_e$ where $q_e$ is the effective dimension of the signal subspace, and we have a basis for $R^\perp$, say $e_{q+1}, \ldots, e_{q+q_e}$, and let $E_n = [e_1, \ldots, e_{q+q_e}]$. Since $e_i$s are in $R^\perp$, their transformation under $L^*$ will be in $S^\perp$, i.e.

$$L^*e_i = a^H(\theta)e_i \in S^\perp, \quad i = 1, \ldots, p - q_e.$$  \hspace{1cm} (34)

Thus for any $s(\theta) \in S$ we have

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} a^H(\theta)E_n s(\theta)d\theta = 0.$$  \hspace{1cm} (35)

In (32) the source subspace $S$ was defined as the span of the functions $s_i(\theta; \psi_i)$. Hence (35) can be written as

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} a^H(\theta)E_n s_i(\theta; \psi_i)d\theta = 0$$  \hspace{1cm} (36)

for all realizations of $s_i(\theta; \psi_i)$, and for $i = 1, \ldots, q$. Since $s_i(\theta; \psi_i)$ is a random function, this is equivalent to

$$E\left[\left\|\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} a^H(\theta)E_n s_i(\theta; \psi_i)d\theta\right\|^2\right] = 0.$$  \hspace{1cm} (37)

for $i = 1, \ldots, q$. Using (3) and (4) this equation can be expressed as

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} a^H(\theta)E_n p(\theta, \theta'; \psi_i)E_{\Delta}^H a(\theta')d\theta d\theta' = 0$$  \hspace{1cm} (38)

for $i = 1, \ldots, q$. We propose that the parameter vector be estimated by locating the peaks of

$$\psi = \arg \max_{\psi} \frac{1}{\sqrt{\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} a^H(\theta) E_n p(\theta, \theta'; \psi) E_{\Delta}^H a(\theta')d\theta d\theta'}}.$$  \hspace{1cm} (39)

We call this method the distributed signal parameter estimator (DSPE). The spectrum of the DSPE algorithm is searched in an $m$-dimensional space for $q$ local maxima where $m$ is the dimension of the parameter vector $\psi$. The maximum points correspond to the estimates of the parameter vector.

5. Some simulation results

In this section we show how the DSPE technique can localize distributed sources. Here we only consider a single ID source with a uniform correlation kernel. The source is arriving from 10 degrees with an extension of 2 degrees at a uniformly spaced array of 20 sensors. Fig. 4 shows the spectrum of the DSPE localization. The prominent peak in this spectrum corresponds to the distributed source. As it is seen the method successfully localizes the source. In this simulation we have used 2 eigenvalues for the signal subspace. The spectrum of the corresponding MUSIC algorithm can be read in the figure by putting $\Delta = 0$. For a complete set of simulations see [5].
6. Summary

This paper is a continuation of a previous work for distributed source localization [4]. It has been assumed that the spatial correlation kernel of the signals is chosen from a parametric class of functions. We have generalized the MUSIC algorithm so that it can localize distributed sources. The new technique has been called a distributed source parameter estimator (DSPE). The new algorithm can localize the extended sources by minimizing a scalar product between the back-transform of the noise eigenvectors and a basis for the source subspace. Furthermore, we have studied the performance of a conventional beamformer by finding the array gain for a uniform linear array. We have shown that the gain of a conventional beamformer is bounded with increasing the number of sensors.

References


Fig. 1 Spatial cross-correlation for a uniform CD source. (The first sensor is positioned at the phase reference point.)

Fig. 2 Array gain for a uniform CD source for different distribution widths, $\Delta$, in degrees.

Fig. 3 Array gain for a uniform ID source for different distribution widths, $\Delta$, in degrees.

Fig. 4 The spectrum of the DSPE algorithm. The angles are in degrees.
Maximum likelihood source separation for discrete sources

Adel Belouchrani  Jean-François Cardoso

École Nationale Supérieure des Télédécommunications
Télécom Paris, Département Signal, 46 rue Barrault, 75013 Paris Cedex 13, France.
E-mail: adel@sig.enst.fr, cardoso@sig.enst.fr, Tel: (33) 1 45 81 75 47, Fax: (33) 1 45 88 79 35.

Abstract. This communication deals with the source separation problem which consists in the separation of a noisy mixture of independent sources without a priori knowledge of the mixture coefficients. In this paper, we consider the maximum likelihood (ML) approach for discrete source signals with known probability distributions. An important feature of the ML approach in Gaussian noise is that the covariance matrix of the additive noise can be treated as a parameter. Hence, it is not necessary to know or to model the spatial structure of the noise. Another striking feature offered in the case of discrete sources is that, under mild assumptions, it is possible to separate more sources than sensors. In this paper, we consider maximization of the likelihood via the Expectation-Maximization (EM) algorithm.

1. Introduction

When an array of m sensors samples the fields radiated by n narrow band sources its output is classically modeled as an instantaneous spatial mixture of a random vector made of m one-dimensional components, possibly corrupted by additive noise. The source separation problem consists in identifying such a spatial from its output only. The key point here is that source separation algorithms do not assume any a priori information on the propagation, the array shape, and other physical parameters which determine the linear system. While much attention has been paid to the blind identification of convolutional mixtures, source separation concerns itself only with multiplicative mixtures.

Source separation algorithms are based on the crucial assumption of mutual statistical independence of the source signals. Various techniques have been proposed. A first class of them are based on the block processing of higher order cumulants [1, 2, 3]. Another class is based on non-linear spatial adaptive filters [4, 5, 6, 7, 8]. Both these classes assume non-Gaussian source signals but do not require the exact knowledge of their distributions. In contrast, when the source distributions are known in advance, significant improvement can be gained by taking this information into account. In this case, it is possible to implement a maximum likelihood approach to solve the source separation problem. The noiseless case is addressed in [9] for continuously distributed source signals. The ML approach in the case of discrete sources shows a very different structure. In particular, the spatial covariance of the additive noise can be easily included in the parameters to be estimated providing estimates which are "blind" to the noise as well as to the linear system.

This paper shows how ML estimates for the separation of discrete sources can be implemented using the Expectation-Maximization (EM) algorithm. Striking features of this approach with respect to previously reported works are i) the spatial structure of the noise needs not to be known in advance since the noise covariance matrix can be treated as a parameter ii) under mild assumptions, the proposed method allows the separation of more source signals than sensors iii) As expected with discrete sources, substantial improvement with respect to cumulant-based blind identification algorithms is observed.

The paper is organized as follows. The data model and the relevant hypothesis are stated in section 2. The proposed ML approach is described in section 3 and is illustrated by numerical simulations in section 4.

2. Model and assumptions

Consider an array of m sensors receiving signals from n narrow band sources. The array output denoted \(X(t)\) is a \(m \times 1\) random vector corrupted by independent additive noise denoted \(N(t)\); it is classically modelled as:

\[
X(t) = AS(t) + N(t),
\]

where the \(n \times n\) matrix \(A\) is the mixture matrix and \(S(t)\) is a \(n \times 1\) random vector whose p-th component, denoted \(S_p(t)\), is the signal emitted by the \(p\)-th source at time \(t\). It is assumed to be independent of \(S_q(t')\) if \(p \neq q\) or \(t \neq t'\).

We further assume that each source transmits symbols taken from a known alphabet with known a priori probabilities. Hence, the vector of symbols \(S(t)\) takes its values in a finite set of vectors, denoted \(S\), with size \(|S|\). Each composite symbol vector \(S_T\) in \(S\) has size \(n \times 1\), and its a priori probability is denoted \(p(S_T)\). For convenience, we denote

\[
S_T \overset{\text{def}}{=} [S(1), \cdots, S(T)] \in S^T, \quad (2)
\]

\[
X_T \overset{\text{def}}{=} [X(1), \cdots, X(T)] \in \mathbb{C}^{n \times T}. \quad (3)
\]
The additive noise $N(t)$ is zero mean Gaussian, temporally white with an unknown covariance matrix $R_N$. The parameters to be estimated are globally denoted as 

$$\theta \overset{df}{=} (A, R_N).$$

3. Maximum likelihood approach

3.1. Conditional likelihood estimation

If the emitted symbols were available, it would be a simple matter to compute the ML estimate of $\theta = (A, R_N)$ because the density probability of $X_T$ given a particular sequence $S_T$, denoted $p(X_T|S_T; \theta)$, is easily maximized under our hypothesis. The solution consists in evaluating the statistics:

$$R_{XX} = T^{-1} \sum_{t=1}^{T} X(t)X(t)^*$$
$$R_{XS} = T^{-1} \sum_{t=1}^{T} X(t)S(t)^* = R_{SS}^T$$
$$R_{SS} = T^{-1} \sum_{t=1}^{T} S(t)S(t)^*$$

and computing estimated parameters by

$$\hat{A} = R_{XX}R_{SS}^{-1}$$
$$\hat{R}_N = R_{XX} - R_{XS}R_{SS}^{-1}R_{SX}.$$  

3.2. Incomplete likelihood

In our problem, the emitted symbols are not available: the likelihood $p(X_T; \theta)$ of the observations is the marginal distribution

$$p(X_T; \theta) = \sum_{S_T} p(X_T, S_T; \theta) = \sum_{S_T} p(X_T|S_T; \theta) p(S_T)$$

where $p(X_T, S_T; \theta)$ denotes the joint likelihood of $X_T$ and $S_T$, and $p(S_T)$ is the a priori probability of $S_T$. This is a sum over $|S|$ terms, but the independence assumption reduces it to a sum over $T|S|$ terms. In particular, the log-likelihood of the observations,

$$L(\theta) \overset{df}{=} \log p(X_T; \theta),$$

has the following partial derivative

$$\frac{\partial L}{\partial \theta_i} = \sum_{t=1}^{T} \sum_{S_T} p(S_T|X(t); \theta) \frac{\partial}{\partial \theta_i} \log p(X(t)|S_T; \theta)$$

which can be evaluated by

$$p(S_T|X(t); \theta) = \frac{p(X|S_T; \theta)p(S_T)}{p(X; \theta)} = \frac{p(X|S_T; \theta)p(S_T)}{\sum_{S_T} p(X|S_T; \theta)p(S_T)}.$$  

Hence the cost of computing the gradient of the log-likelihood essentially is $T|S|$ times the cost of computing $p(X|S_T; \theta)$ and its log-derivative.

3.3. The EM algorithm

In this paper, we consider the Expectation-Maximization (EM) approach to likelihood maximization [16]. It is based on the auxiliary function:

$$Q(\theta, \theta') \overset{df}{=} \sum_{S_T} \log p(X_T|S_T; \theta)p(S_T|X_T; \theta'),$$

which shows the key property that

$$Q(\theta, \theta') \geq Q(\theta', \theta') \implies L(\theta) \geq \langle \theta' \rangle.$$  

It follows that a sequence $\theta(t)$ of parameters iteratively computed as

$$\theta(t+1) = \arg\max_{\theta} Q(\theta, \theta(t))$$

produces a monotonic increase in the likelihood.

Under the assumption of independence at different time lags, the derivative of the auxiliary function (13) is

$$\frac{\partial Q}{\partial \theta_i} (\theta, \theta') = \sum_{t=1}^{T} \sum_{i=1}^{|S|} p(S_T|X(t); \theta') \frac{\partial}{\partial \theta_i} \log p(X(t)|S_T; \theta)$$

in striking similarity to (11). In particular, it is clear from (16) that a fixed point of the EM algorithm is a stationary point of the likelihood. What makes the algorithm of simple implementation here is equation $\frac{\partial^2 Q}{\partial \theta_i^2} (\theta, \theta') = 0$ admits for the model under consideration a unique solution which can be computed in closed form. It is obtained by evaluating the statistics:

$$R_{XS}^{(i)} = R_{XX}^{(i)} - R_{XS}^{(i)}R_{SS}^{-1}R_{SX}$$
$$R_{SS}^{(i)} = R_{SS} - R_{XS}^{(i)}R_{SS}^{-1}R_{SX}$$

and reestimating new parameters $\theta^{(i+1)} = (A^{(i+1)}, R_N^{(i+1)})$ according to

$$A^{(i+1)} = R_{XS}^{(i)}R_{SS}^{-1}$$
$$R_N^{(i+1)} = R_{XX} - R_{XS}^{(i)}R_{SS}^{-1}R_{SX}.$$  

Note that the above procedure does not involve the (pseudo)inverse of the mixing matrix $A$. As a matter of fact, the ML approach allows the estimation of noise signals when sensors, as illustrated in the next section.

4. Simulation results

The numerical performance evaluation presented below has been obtained in the following setting. The performance indices are evaluated over 500 Monte Carlo runs. The source signals are i.i.d. QAM4 symbols with unit variance. The p-th element of the q-th column of $A$ is $\exp(2\pi r q a_p)$, so that the q-th column of $A$ is parameterized by the ‘electric’ angle $a_q$. Of course, we do not use here this a priori information. Similarly, the noise covariance matrix is $R_N = \sigma_l^2 I_n$ but is estimated without using this constraint. The algorithm is initialized using the JADE algorithm, a
cumulant-based estimation technique described in [3, 11]. The plotting convention is: solid line for the ML algorithm; dotted for ML identification assuming that the emitted symbols are available (this is for reference); dash-dotted line for the cumulant-based algorithm JADE; dashed line at level 1/4T: this is the theoretical lower bound [12] for the rejection level of source separation algorithms based on prewhitening (such as JADE).

Figure 1 displays a mean rejection level (MRL) defined as

\[
MRL = \frac{1}{n(n-1)} \sum_{i \neq j} |(\hat{A}^i \hat{A})_{ij}|^2. \tag{21}
\]

It is plotted in dB versus \( \sigma \) in dB for \( m = 5 \) sensors and two sources at \( \alpha_1 = 0 \) and \( \alpha_2 = 0.04 \) and \( T = 30 \) samples.

Figure 2 shows the variation of \( \|\hat{A} - A\|^2 / \|A\|^2 \) versus the number of sensors, \( m \) ranging from 2 to 5 (in integer values) for \( T = 200 \) samples, \( \sigma = -5 \) dB and 3 sources at \( \alpha_1 = 0.1, \alpha_2 = 0.2, \alpha_3 = 0.3 \). We use here a different index since MRL is not defined for \( n > m \).

Figures 3 and 4 illustrate the convergence during a sample run with 3 sensors, 3 sources, 100 samples and \( \sigma = -10 \) dB. Figure 3 shows the improvement from the initial estimate (top row) till convergence (last row). Each column corresponds to a particular source. The symbols are displayed in the complex plane after artificially removing the additive noise. Figure 4 displays the improvement of the MRL across 3 iterations in the same sample run as in fig. 3. In this case, the initial estimate returned by JADE was of good quality so that a fast (3 iterations) convergence is observed. The improvement with respect to the cumulant based technique still is significant.

![Figure 1. Performance vs noise level.](image1)

![Figure 2. Performance vs number of sensors.](image2)

![Figure 3. A sample run of the EM algorithm.](image3)

![Figure 4. Mean rejection level vs number of iterations.](image4)
Conclusion

We have considered maximum likelihood estimation for a mixture of discrete sources in Gaussian noise. The Expectation-Maximization technique can be used to this effect and takes a simple form thanks to the i.i.d. assumption. The ML approach assumes that the a priori distribution of the emitted symbols is known in advance. The benefits with respect to other techniques not using this kind of information are manifold: i) improvement of the quality of the separation ii) ability to deal with additive noise of unknown covariance, since the covariance matrix can be included in the parameters to be estimated iii) possibility of estimating more sources than sensors. A stochastic version of the previous technique is under study and will be presented in a forthcoming communication.

References

INTERPOLATION OF DISTORTED TOWED ARRAYS
FOR SOURCE BEARING ESTIMATION

Alain Marsal, Sylvie Marcos

Laboratoire des Signaux et Systèmes, CNRS-ESE,
Plateau de Moulon, F-91192 Gif-sur-Yvette Cedex, France
Phone: 33 1 69 41 80 40 Fax: 33 1 69 41 30 60
E-mail: marcos@lss.supelec.fr

Abstract. We here propose an interpolation technique to solve the problem of source localization with a distorted array. This technique needs to know the sensor location only for a number of sensors equal to the number of sources. This is quite interesting for long arrays where there are many more sensors than sources. The proposed technique is obtained with the Propagator, a linear operator, which has already been successfully used for array calibration. The point here is that we can interpolate a real array of unknown shape in such a way that we can perform source bearing estimation with a virtual array of desired shape without estimating the sensor location of the actual array (i.e. without calibration).

1 Introduction

Source bearing estimation is of major interest in array signal processing. It consists in finding the angles of arrival of multiple signals transmitted by far-field sources which impinge on an array of sensors. It turns out that some methods of source bearing estimation require that the sensor array possesses a particular arrangement of the sensors or a particular shape. For example, in the ESPRIT method [1] the array should consist of two subarrays which are translational equivalent. Another example is the root-MUSIC algorithm [2] which requires that the array is uniform and linear (UL) in order to search for the source bearings by rooting a polynomial instead of searching for the maxima in a pseudospectrum. Note that the sensor array must also be of the UL type for the spatial smoothing methods dealing with fully correlated sources [3].

In some applications, the sensor array is, however, given and cannot be designed as wanted. It is then interesting to be able to construct from the actual array a virtual or a synthetic array possessing the desired characteristics. The use of virtual or synthetic arrays has already been considered in the source bearing estimation problem with manoeuvring towed arrays [4] where a single virtual array is constructed from the successive positions of the actual array. Also, in the case of wide-band array processing, some methods require to have several different arrays designed in such a way that the array manifolds at each frequency bin are equal [5]. In order to avoid having multiple physical arrays, the different arrays may be virtually constructed from the actual array.

The design of a virtual array is obtained by an interpolation technique [4, 6, 7, 8]. Given a desired shape for the array, the interpolator coefficients are obtained by a minimization of the interpolation error between the actual array manifold and the virtual array manifold within an angular sector. Using these interpolators, a sequence of cross spectral matrices (CSM) of the signals received on the virtual array, corresponding to each sector, is synthesized. The composite matrix can then be used to apply the desired source bearing estimation technique.

Although this interpolation technique is valid for any arrangement of the sensors, it requires the shape and the sensor location of the actual array to be known. In some applications of array processing the sensor array may, however, be distorted and of unknown shape. This is the case of long flexible towed arrays in underwater acoustics' applications.

The purpose of this communication is to present an interpolation technique which only requires the knowledge of the location of a number of sensors equal to the number of sources. In section 2, we introduce the notations for the problem and we briefly recall the definition of the Propagator. In section 3, we recall the main techniques for the interpolation of arrays [4, 7], their required hypothesis and their practical limitations. In section 4, we propose the Propagation-based technique. This technique is finally illustrated in section 5, in the case of a manoeuvring distorted array.

2 Definition of the Propagator

Consider an array of $M$ sensors on which $P$ incident waves impinge ($M > P$). At a given frequency and for a particular snapshot, the usual model of the $M$-
dimensional vector of the received signals can be written as
\[ x = As + b, \]  
where \( s \) is a \( P \)-dimensional vector containing the complex amplitudes of the sources, \( b \) is a noise vector and \( A \) is the \( (M, P) \)-dimensional matrix of the steering vectors. Under the hypothesis that matrix \( A \) is full rank, \( P \) rows of \( A \) are linearly independent.

Let us suppose here that the first \( P \) rows of \( A \) are linearly independent. The definition of the Propagator [9] relies on the division of the steering matrix \( A \) according to
\[ A = \begin{bmatrix} A^{(1)} \\ A^{(2)} \end{bmatrix}, \]
where \( A^{(1)} \) is a matrix having \( P \) rows. Under the hypothesis that \( A^{(1)} \) is non-singular, we define the propagator as the unique linear operator given by
\[ P^* A^{(1)} = A^{(2)}, \]
where \( \star \) denotes the hermitian–transposition.

The propagator can be estimated either from the data matrix
\[ X = [x(1), x(2), \ldots, x(K)], \]
or from the estimated CSM
\[ \hat{\Gamma} = \frac{1}{K} XX^* \]
where \( K \) is the number of snapshots. By splitting matrices \( X \) and \( \hat{\Gamma} \) according to
\[ X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} \quad \text{and} \quad \hat{\Gamma} = [G, H] \]
where \( X^{(1)} \) contains the \( P \) first rows of \( X \) and where \( G \) contains the \( P \) first columns of \( \hat{\Gamma} \), a Least Square (LS) estimation of the propagator is given by
\[ \hat{\Phi}_{\text{data}} = (X^{(1)}X^{(1)*})^{-1}X^{(1)}X^{(2)*} \quad \text{or} \quad \hat{\Phi}_{\text{com}} = (G^*G)^{-1}G^*H. \]

### 3 Interpolation of arrays

The interpolation technique mainly consists in searching a matrix \( B \) such that [4]
\[ BA = A_V \]
where \( A_V \) is the steering matrix corresponding to the desired virtual array. If rough estimates of the source bearings are known, say \( \hat{\theta} \), the interpolation matrix is the one which minimizes [7]
\[ \left\| A_V(\hat{\theta}) - BA(\hat{\theta}) \right\|_F^2 \]
where \( \| . \|_F \) denotes the Frobenius norm of a matrix.

If no rough estimates of the directions of arrival (DOAs) are available, the angular space is divided into sectors within which an interpolation matrix can be obtained minimizing
\[ \left\| A_{V_l} - B_l A \right\|_F^2 \]
where the index \( l \) stands for the sector \( l \), and where the matrices \( A_l \) and \( A_{V_l} \) are no longer the steering matrices but the actual and virtual array manifolds associated with sector \( l \), respectively.

In both above-mentioned cases, the CSM of the received signals on the virtual array can then be derived from the CSM measured on the real array and from the interpolation matrix. The desired source bearing estimation can then be applied.

Assume now that only the location of the first \( P \) sensors is known (without loss of generality). It follows from the definition of the Propagator (3) that the interpolation matrix (9) can be obtained as a function of \( P \), \( A^{(1)} \) and \( A_V \) only, where \( P \) is estimated from the CSM of the signals received on the actual array. In the case of the sector interpolation, the manifolds \( A_l^{(1)} \) and \( A_{V_l} \) should be used.

Since the Propagator is defined from a division of the steering matrix and not from a division of the manifold, the sectorization cannot be used in this case. We must have an a priori information on the DOA in order to build a matrix close to the matrices \( A^{(1)} \) and \( A_V \). Krishna and Pau raj [4] used the same a priori information to interpolate a manoeuvring, known shape array (see section 5). This required knowledge may be obtained by a pre-processing (e.g. beamforming) DOAs estimation. As most of the array shape is unknown, we suggest to roughly estimate the DOAs with the smallest moduli dispersion method (see appendix A).

### 4 Interpolation of an array of partially unknown shape

When the array shape is not perfectly known, the interpolation is not feasible. Assume that the location of only the first \( P \) sensors is known (for example when long towed arrays are used, with \( M \gg P \) and where the first part of the array is supposed to follow the direction of the boat).

If we have a rough estimation of the DOAs \( \hat{\theta} \), the definition (3) allows to estimate the block \( A^{(2)}(\hat{\theta}) \), from the block \( A^{(1)}(\hat{\theta}) \) and the \( (M - P, P) \)-dimensional Propagator \( P \) (7). We can also build \( A_V(\hat{\theta}) \). Then we can find a transformation matrix \( B \) as a least square solution of
\[ A_V(\hat{\theta}) = B \begin{bmatrix} A^{(1)}(\hat{\theta}) \\ P^* A^{(1)}(\hat{\theta}) \end{bmatrix}, \]
where \( \| . \|_F \) denotes the Frobenius norm of a matrix.

If no rough estimates of the directions of arrival (DOAs) are available, the angular space is divided into sectors within which an interpolation matrix can be obtained minimizing
\[ \left\| A_{V_l} - B_l A \right\|_F^2 \]
where the index \( l \) stands for the sector \( l \), and where the matrices \( A_l \) and \( A_{V_l} \) are no longer the steering matrices but the actual and virtual array manifolds associated with sector \( l \), respectively.

In both above-mentioned cases, the CSM of the received signals on the virtual array can then be derived from the CSM measured on the real array and from the interpolation matrix. The desired source bearing estimation can then be applied.

Assume now that only the location of the first \( P \) sensors is known (without loss of generality). It follows from the definition of the Propagator (3) that the interpolation matrix (9) can be obtained as a function of \( P \), \( A^{(1)} \) and \( A_V \) only, where \( P \) is estimated from the CSM of the signals received on the actual array. In the case of the sector interpolation, the manifolds \( A_l^{(1)} \) and \( A_{V_l} \) should be used.

Since the Propagator is defined from a division of the steering matrix and not from a division of the manifold, the sectorization cannot be used in this case. We must have an a priori information on the DOA in order to build a matrix close to the matrices \( A^{(1)} \) and \( A_V \). Krishna and Pau raj [4] used the same a priori information to interpolate a manoeuvring, known shape array (see section 5). This required knowledge may be obtained by a pre-processing (e.g. beamforming) DOAs estimation. As most of the array shape is unknown, we suggest to roughly estimate the DOAs with the smallest moduli dispersion method (see appendix A).
Many algorithms are provided to solve the system (11). They are based on a least square optimization, with or without any constraint (see [4, 7]). Once \( B \) is estimated, the CSM of the signals received on the virtual array can be built from the true CSM. Finally, the DOAs can be estimated by any high-resolution method.

The initial DOA can be found by the “Smallest Moduli Dispersion” (SMD) algorithm, defined with the Propagator [10] (see appendix A). This rough DOA estimator appears to be unaffected by any distortion on the \( M - P \) other sensors.

In [10] the SMD was used to calibrate the array in order to improve source bearing estimation. The interest here is that the calibration step is avoided. The SMD is used to interpolate the actual array of a partially unknown shape into a virtual array of the desired shape. Note that the calibration is not a simple task and may lead to non unique solutions. The interpolation technique, in conjunction with the Propagator, not only allows to process the signals received on a virtual array of desired shape, but also represents an alternative to the calibration techniques.

5 Illustration

The performances of the proposed interpolation technique is here illustrated in the case of a manoeuvring array.

Let us consider a towed array manoeuvring along a particular curve. This motion can be sequenced into several positions in which the array can be considered as fixed. However, the number of snapshots thus recorded during each position may not be large enough to allow an accurate spatial analysis.

The solution presented by Krishna and Paulraj in [4] is to interpolate each position of the array onto the same virtual array. This means that all the snapshots recorded during the whole motion can be used for a better analysis.

This algorithm relies on the assumption that the shape of the array is exactly known for each position. This is not completely realistic, as the array tends, in practice, to follow the motion of the boat. If we suppose that only a very few sensors stay very close to their nominal positions, we can implement the Propagator-based interpolation described in section 4. Our algorithm is used not only to perform the interpolation of the different positions of the array into a single virtual array, but also to compensate for the actual array distortions appearing in the motion.

For the simulation, an array of 8 sensors (figure 1) or 50 sensors (figure 2) is moving along half a circle. We consider that the array shape is distorted according to the same arc of a circle during the motion of the boat. Three sources are present at -5, 10 and 25 degrees with respect to the normal of the array. The signal-to-noise ratio is 10 dB per sensor and per source. The total time during which the array is manoeuvring is decomposed into eight sequences each consisting of 25 snapshots.

The figures 1 and 2 exhibit the result of the MUSIC method [11, 12] performed on the first sequence (plain lines), the Krishna—interpolation method (dashed lines) (the array is assumed to be linear during the interpolation whereas in fact it is distorted) and the proposed interpolation technique using the Propagator (dotted lines). The vertical lines indicate the true DOAs.

For the 50 sensors array, MUSIC, applied to the first sequence, fails because the number of sensors is smaller than the number of snapshots. The Krishna algorithm also fails because the distortions are much more important. The Propagator—based algorithm both compensates for the array distortions and interpolates the different positions of the moving array into a single array.

![Figure 1: Pseudo-spectrum after the interpolation of the manoeuvring array (m=8). Music (-----), Krishna algorithm (-----), Propagator-based algorithm (…)](image)

6 Conclusion

In this communication, we have proposed a new array interpolation technique based on the Propagator, a linear operator which can be estimated from the CSM of the signals received on the actual array. This Propagator—based algorithm can interpolate an array with partially unknown shape into a virtual array of the desired shape.

This algorithm is thus an extension to the case of distorted array of the algorithm proposed in [7] for the interpolation of known shape arrays. The information on the array shape, together with the sectorization are replaced by an a priori information on the DOAs.

The Propagator—based interpolation technique can also be considered as an alternative to the array calibration algorithms.
Figure 2: Pseudo-spectrum after the interpolation of the maneuvering array (m=50). Music (---), Krishna algorithm (-----), Propagator-based algorithm (…)

A Appendix: The SMD method

Given \( a(\theta, \gamma) \) a steering vector model which depends on an angular parameter \( \theta \) and given the following partition

\[
\begin{bmatrix}
  a^{(1)}(\theta, \gamma_1) \\
  \vdots \\
  a^{(2)}(\theta, \gamma_2)
\end{bmatrix}
\begin{bmatrix}
  P \\
  M - P
\end{bmatrix},
\]

where \( \gamma = [\gamma_1, \gamma_2] \) consists of the location parameters \( \gamma_1 \) and \( \gamma_2 \) of the \( P \) first sensors and the \( M - P \) last sensors, respectively; \( a^{(1)}(\theta, \gamma_1) \) is the vector model associated to the subarray where propagation parameters \( \gamma_1 \) are supposed to be known, and \( a^{(2)}(\theta, \gamma_2) \) is unknown. One can build, from \( P \) and \( a^{(1)}(\theta, \gamma_1) \), the manifold of the possible wavefronts

\[
\mathcal{V}_P = \{ a(\theta, \gamma) : a^{(2)}(\theta, \gamma_2) = P^*a^{(1)}(\theta, \gamma_1); \ \theta \in [-\pi, +\pi] \}.
\]

Therefore, the unknown steering vectors belong to manifold (13) as long as the location of the \( P \) first sensors (\( \gamma_1 \)) is known. In order to identify the steering vectors \( a_i \) within manifold (13), the method presented in [10] supposes that the components of \( a_i \) have the same modulus. This implies that the gain sensor modulus should be equal. In this case, the unknown steering vectors are identified as belonging to manifold (13) with a minimum of modulus dispersion. More details on this technique are given in [10] and are not mentioned here. The modulus dispersion of the possible steering vectors \( a \) is defined by

\[
\sigma(\theta) = \sqrt{\frac{M}{\sum_{m=1}^{M} |a_m|^2} - \left( \frac{\sum_{m=1}^{M} |a_m|}{\sum_{m=1}^{M} |a_m|} \right)^2},
\]

where \( a_m \) is the \( m^{th} \) component of the vector \( a(\theta, \gamma) \) belonging to manifold (13). Note that the information concerning \( \gamma_2 \) is included in the Propagator \( P \); \( \gamma_1 \) is supposed to be known.

It appears that (14) can be considered as a rough DOA estimator which does not require the knowledge of the propagation model for \( M - P \) sensors (this is quite interesting in underwater acoustics where \( M \) is much larger than \( P \)).

References


On the performance of orthogonal source separation algorithms

Jean-François Cardoso

École Nationale Supérieure des Télécommunications
Télécom Paris, Département Signal, 46 rue Barrault, 75634 Paris Cedex 13, France.
email: cardoso@sig.enst.fr, fax: (33) 1 45887935.

Abstract. Source separation consists in recovering a set of \( n \) independent signals from \( m \geq n \) observed instantaneous mixtures of these signals, possibly corrupted by additive noise. Many source separation algorithms use second order information in a whitening operation which reduces the non trivial part of the separation to determining a unitary matrix. Most of them further show a kind of invariance property which can be exploited to predict some general results about their performance. Our first contribution is to exhibit a lower bound to the performance in terms of accuracy of the separation. This bound is independent of the algorithm and, in the i.i.d. case, of the distribution of the source signals. Second, we show that the performance of invariant algorithms depends on the mixing matrix and on the noise level in a specific way. A consequence is that at low noise levels, the performance does not depend on the mixture but only on the distribution of the sources, via a function which is characteristic of the given source separation algorithm.

1. Introduction.

This paper is concerned with the source separation problem which consists in recovering a set of \( n \) independent signals from \( m \geq n \) observed instantaneous mixtures of these signals. Denoting \( x(t) \) the \( m \times 1 \) vector of observations (sensor outputs) at time \( t \), possibly corrupted by additive noise \( n(t) \), the model is

\[
x(t) = As(t) + n(t) = y(t) + n(t)
\]

(1)

where the \( m \times n \) matrix \( A \) is called the 'mixing matrix' and where the \( n \) independent signals are collected in a \( n \times 1 \) vector denoted \( s(t) \). All the processes are assumed to be zero-mean stationary. The aim of source separation is to find a separating matrix, i.e. a \( n \times m \) matrix such that \( \hat{s}(t) = Bx(t) \) is an estimate of the source signals.

In the complex case, model (1) is the familiar linear model used in narrow band array processing. In this context, it is usually assumed that the columns of \( A \) depend on very few location parameters (such as DOAs) and this dependence is assumed to be known via the 'array manifold'. In contrast, we address here a problem of blind array processing in the sense that matrix \( A \) is completely unstructured: it is supposed to be a fixed full column rank matrix but no other assumptions are made.

The 'blind' approach is strongly motivated when i) one is interested in recovering the source signals (like in communication applications) but not in locating the emitting sources and ii) whenever the array manifold is unavailable or is expected to significantly depart from its model. Source separation is calibration-free and, by essence, insensitive to modelling errors. Blind source separation relies essentially on the assumption that the signals to be separated are mutually independent; this is a strong but often plausible assumption, which may be exploited using either the non-normality (if any) or the spectral differences (if any) of the source signals.

Several block-oriented source separation algorithms have been proposed in [1-7] which are based on a two step approach: whitening and rotating the observations. This paper is concerned about the performance of such algorithms. The idea here is to determine the general performance of this approach without specifying a particular algorithm. Although this paper addresses a statistical issue, emphasis is not statistical rigor but rather on exhibiting the algebraic mechanism by which the prewhitening mechanism affects the performance.

2. Performance and blind identification

2.1. Blind identification

A source separation algorithm also is a blind identification technique and may be represented as

\[
\hat{A} = A(X_T)
\]

(2)

where an estimate \( \hat{A} \) is computed from a block of \( T \) samples by some algorithm which is represented here as a function \( A \) of the \( m \times T \) matrix\( X_T \):

\[
X_T = [x(1), \ldots, x(T)].
\]

(3)

Before going further, it is important to notice that the function \( A \) may not be perfectly defined because of some indeterminations of the blind identification problem. This is because the exchange of a complex factor between each source signal and the corresponding column of \( A \) leaves the observations unchanged. It follows that without any loss of generality, the source signals can be assumed to have unit variance. With this normalization convention (we insist that this is not an assumption), the covariance matrix of \( s(t) \) is the identity matrix:

\[
R_s = E(s(t)s^\ast(t)) = I_n.
\]

(4)

\[1\] We use the same notational convention for \( S_T, N_T \), etc...
This convention still leaves undetermined the phase of each column of $A$ as well as their ordering since the ‘labelling’ of the source signals is immaterial. In the following, two estimates $\hat{A}$ and $\hat{A}'$ of $A$ are considered as equivalent if
\[ \hat{A}' = \hat{A}J \]
when $J$ is any $n \times n$ matrix with only one non-zero unit-norm entry in each row and each column. We call such a matrix a ‘quasi identity’; it is a unitary matrix. A source separation algorithm (hence a function $A$) is considered well defined if it is defined up to right multiplication by some quasi-identity.

2.2. Rejection rates
We shall characterize the quality of an estimate $\hat{A}$ of $A$ by the closeness of matrix $\hat{A}^\theta A$ to some quasi-identity matrix where superscript $\#$ denotes pseudoinversion. Indeed if $\hat{A}^\theta$ is used to estimate the source signals from the observations, then
\[ \hat{g}(t) = \hat{A}^\theta X(t) = (\hat{A}^\theta A)\hat{m}(t) + \hat{A}^\theta \hat{n}(t) \]
so that the variance of the $q$-th signal at the $p$-th output of the separator is given by
\[ P_{pq} = |(\hat{A}^\theta A)_{pq}|^2 \]
since our convention is that each source has unit variance. Hence, examination of the distribution $P_{pq}$ provides a very intuitive measure of the performance of the identification.

Note that changing $\hat{A}$ into $AJ$ for $J$ some quasi-identity matrix amounts to changing $P_{pq}$ into $P_{p\sigma(q)}$ where $\sigma$ is some permutation of $\{1, \ldots, n\}$. We assume in the following that the permutation has been removed (for instance on the basis of some additional a priori information) so that $P_{pq}$ actually is the variance of the $q$-th signal in the estimate of the $p$-th signal. The phase indetermination is invisible in $P_{pq}$. Also note that when $\hat{A}$ is close enough to $A$, then $P_{pq}$ is close to 1; the quantities $P_{pq}$ are readily normalized and can be directly read as ‘rejection rates’, ‘interfer-to-signal ratio’, etc…

2.3. Invariance
Assume for a while that $n = m$ and that no noise is present. Some source separation algorithms yields the estimate of $A$ as the solution of
\[ \frac{1}{T} \sum_{i=1}^{T} G(\hat{A}^{-1} x(t)) = 0 \]
where $G$ is some vector-to-matrix mapping. A slight modification of [8] fails in this class and [9] is an adaptive solver of such an equation. The estimator $A$ associated to (8) is such that,
\[ \hat{A}(CX_T) = CA(X_T) J \]
for any invertible matrix $C$ and some quasi-identity matrix $J$. This property is called here full invariance (we consider later a restricted ‘orthogonal invariance’). For a fully invariant estimator:
\[ \hat{A}^\theta A = A(X_T)^{-1} A = A(AS_T)^{-1} A = J A(S_T)^{-1} \]
for some $J$ so that, up to a permutation, the distribution of $P_{pq}$ depends only on $S_T$ in the noiseless case. We conclude that in absence of noise, the performance of a fully invariant estimator does not depend on mixing matrix $A$.

This very brief discussion on invariant algorithms is intended to introduce the next two sections: these ideas can be, to some extent, generalized to source separation algorithms based on prewhitening even though they are not fully invariant and are supposed to operate in noisy situations.

3. Invariant orthogonal estimators
A few preliminary definitions are needed. The signal subspace is the range of $A$ and the noise subspace is the orthogonal subspace. The orthogonal projectors onto these subspaces, respectively denoted as $\Pi$ and $\Pi^\perp$, are given by
\[ \Pi = AA^H = A(A^H A)^{-1} A^H, \quad \Pi^\perp = I_m - \Pi \]

We also define the following covariance matrices
\[ R_x = Ex(x(t) x(t))^*, \quad R_y = Ey(y(t) y(t))^*, \quad R_n = En(n(t) n(t))^*, \]
which, by independence of signal and noise, are related by
\[ R_x = R_y + R_n = AR_AR^H + R_n = AA^H + R_n. \]

3.1. A two step procedure
Source separation algorithms based on prewhitening compute estimates of $A$ in the form
\[ \hat{A}^\theta = \hat{U}^H \hat{W} \]
where $\hat{W}$ is a $n \times m$ matrix called the ‘whitening matrix’ and is obtained from an estimate of $R_y$:
\[ \hat{W} = \hat{W}^H(\hat{R}_y) \]
and $\hat{U}$ is a $n \times n$ unitary matrix computed from the whitened data:
\[ \hat{U} = U(\hat{W}X_T) \]
More specifically, $\hat{W}$ is any $n \times m$ matrix verifying
\[ \hat{W}^H \hat{R}_y \hat{W} = 0, \quad \hat{W} = 0 \]
where $\hat{R}_y = I_n - \hat{U} \hat{U}^H$ and $\hat{U}$ is the orthogonal projector on the range of $\hat{R}_y$.

The idea here is recover the source signals by whitening (second-order decorrelation) and then rotating them to further satisfy a stronger independence criterion.

Our approach here is not to specify the particular algorithm $U$ used to find the ‘missing rotation’ because we are interested in properties which are shared by all algorithms based on prewhitening. However, we assume that $U$ satisfies an orthogonal invariance property:
\[ \forall V \text{ unitary}, \quad U(VZ_T) = VU(Z_T) J \]
for almost any realization of $Z_T$ and for some (irrelevant) quasi-identity matrix $J$. Such a property arises naturally in our context. For instance the maximum contrast estimation [4] of $U$ according to
\[ \hat{U} = \arg \max_{U \text{ unitary}} \sum_{i=1}^{n} |C(i^\perp) U^H z(t)|^2 \]
is easily seen to be an invariant estimator. In fact, any reasonable estimator of $U$ obtained by optimizing under unitary constraint a functional of the (empirical) distribution of $U^H x$ enjoys the orthogonal invariance property.
3.2. The noiseles case

As a first step in investigating performance, we show that in the limit where the noise can be neglected, invariant orthogonal estimators become fully invariant in the sense of equation (9).

Note that if no noise is present, the signal subspace can be determined exactly (as the range of $\hat{R}_s$ for instance); source separation algorithms can then operate entirely in this $n$-dimensional subspace. As a consequence, we can deal with the case $n = m$ without loss of generality. Hence in this subsection (only!), we take $n = m$ so that all the matrices of interest are invertible. Denoting $\hat{R}_s = T^{-1} \sum_{t=1}^{\infty} s(t)s(t)^*$, eq. (15) becomes

$$I_n = \hat{W} \hat{R}_s \hat{W}^H = \hat{W} A \hat{R}_s A^H \hat{W}^H.$$  \hspace{1cm} (16)

It follows that $\hat{W}$ satisfies $\hat{W} = V \hat{R}_s^{-1/2}$ where $\hat{R}_s^{-1/2}$ is the positive hermitian square root of $\hat{R}_s$ and $V$ is some undetermined unitary matrix. Then

$$\hat{A}^\# A = U(\hat{W} X)^H \hat{W} = U(V \hat{R}_s^{-1/2} S_1)^H V \hat{R}_s^{-1/2},$$  \hspace{1cm} (19)

and by the orthogonal invariance property (16):

$$\hat{A}^\# A = J U(\hat{R}_s^{-1/2} S_1)^H \hat{R}_s^{-1/2}$$  \hspace{1cm} (20)

for some quasi-identity matrix $J$. This shows that the value taken by $\hat{A}^\# A$ depends only on $S_1$, the particular realization of the source signals and not on the mixing matrix. In fact, property (20) implies that orthogonal invariant algorithms become fully invariant in the noiseless case (again, we disregard the irrelevant quasi-identity matrix).

4. Orthogonal estimators in noise

To deal with the noisy situation and $m > n$, we consider the polar decomposition of matrix $\hat{W} A$ as

$$\hat{W} A = V H$$  \hspace{1cm} (21)

where $V$ is a $n \times n$ unitary matrix and $H$ is a $n \times n$ hermitian, (almost surely) positive matrix. Combining these symmetry properties with (15), matrix $H$ is found to be the unique $n \times n$ positive hermitian matrix verifying

$$H^2 = A^H \hat{R}_s^\# A.$$  \hspace{1cm} (22)

If $R_y$ is known exactly, i.e. if $\hat{R}_s = R_y = A A^H$, then $H = I_n$ so that the distance of $H$ to the identity is a measure of the whitening error.

Some algebra upon (12), (15), (16) and (21) produces

$$\hat{A}^\# A = J U(H(S_1 + A^\# \hat{N})^H H$$  \hspace{1cm} (23)

$$\hat{N}_T \equiv (I_m + (\hat{P}_T)^* \hat{P}_T) N_T.$$  \hspace{1cm} (24)

for some quasi-identity matrix $J$.

This is the noisy equivalent of (20) and forms the basis to understanding how the invariance properties are affected by the additive noise. The key point here is that matrix $A$ has 'almost' disappeared form (23): it only enters in a noise term $A^\# \hat{N}_T$ and also affects the distribution of $H$.

A pairwise lower bound. Orthogonal algorithms rely on second-order prewhitening and are safe in this respect. However this very procedure introduces a lower bound on the separating performance. The idea here is that errors in the whitening step cannot be compensated by any unitary matrix $\hat{U}$. The required lemma is that $H$ being an hermitian matrix, there for $p \neq q$ and any unitary matrix $\hat{U}$:

$$|(\hat{U} H)_{pq}|^2 + |(\hat{U} H)_{qp}|^2 \geq \frac{|(H^2)_{pq}|^2}{(H^2)_{pp} + (H^2)_{qq}}.$$  \hspace{1cm} (25)

Hence even the most clever choice of $U$ in (23) cannot bring $P_{pq} + P_{qp}$ closer to zero than indicated by the right hand side of (25). This later term depends only on $H$ and is then independent of a specific orthogonal algorithm. See (29) for an explicit evaluation in the i.i.d. case.

5. Asymptotics

To get further insights, we specify how $\hat{R}_s^\#$ is computed from $\hat{R}_s$ and go to the asymptotic domain.

5.1. Distribution of the whitening errors

We consider the case of white Gaussian noise with covariance $R_\epsilon = \sigma I$ which allows for a simple estimation of $R_s$ in the following standard fashion, based on an eigendecomposition of $\hat{R}_s$. The estimated noise variance $\hat{\sigma}$ is first obtained as the average of the $m - n$ smallest eigenvalues of $\hat{R}_s$. Then, denoting $\lambda_1, \ldots, \lambda_n$ and $h_1, \ldots, h_n$ the $n$ largest eigenvalues and associated eigenvectors of $\hat{R}_s$, $\hat{R}_y^\#$ is computed as $\hat{R}_y^\# = \sum_{i=1}^{n} (\lambda_i - \hat{\sigma})^{-1} h_i h_i^*$.

The first order expression of $\hat{R}_y$ as a function of $\hat{R}_s = R_s + \delta R_s$ may be computed by the standard perturbation technique. It is:

$$A^H \hat{R}_y^\# A = I - A^\# \delta R_s A^H + \frac{T_T \hat{P}_T^2}{m-n}.$$  \hspace{1cm} (26)

Inserting $\delta R_s = T^{-1}X_T X_T^T - R_s$, in the above shows, after minor rewriting, that $H^2 = A^H \hat{R}_s A$ depends on matrices $A^\# A^{-1}$, $S_T$, $A^\# N_T$ and $\sigma^{-1/2} \xi^2 N_T$ which are mutually independent under the current assumptions on the noise. The distribution of $A^\# N_T^{-1}$ depends only on the covariance matrix of $A^\# n$ which is $A^\# (A^H A)^{-1}$, while the distribution of $\sigma^{-1/2} \xi^2 N_T$ does not depend on any parameter. We conclude that asymptotically, in spatially and temporally white Gaussian noise, the distribution of $H$ depends only on the distribution of the sources and on matrix $A^\# (A^H A)^{-1}$.

5.2. The lower bound.

The asymptotic rejection rate $I_{pq}$ is defined by

$$I_{pq} \equiv \lim_{T \to \infty} T E P_{pq} = \lim_{T \to \infty} T E (\hat{A}^\# A)_{pq}^2.$$  \hspace{1cm} (27)

Assume i.i.d. signals and noise so that the variance of $\hat{R}_s$ decreases as $T^{-1}$ and $(H^2)^{pp} \approx 1$. According to (25)

$$P_{pq} + P_{qp} \geq \frac{1}{2} \left(\frac{|(H^2)_{pq}|^2}{(H^2)_{pp} + (H^2)_{qq}}\right)$$  \hspace{1cm} (28)

so that $I_{pp} + I_{pp} \geq \lim_{T \to \infty} T E [(H^2)^{pp}/2$ which is easily computed in the i.i.d. case using (26) and turns out not to depend on the distributions of the source signals:

$$I_{pq} + I_{qp} \geq \frac{1}{2} \left\{ (1 + \sigma_{pp})(1 + \sigma_{qq}) + \frac{\sigma_{pp}^2 \sigma_{qq}^2}{m-n} \right\}$$  \hspace{1cm} (29)
where $\rho_{pq}$ is the $(p,q)$-th entry of matrix $(A^H A)^{-1}$. In particular, in the low noise limit, the mean rejection rates are lower bounded by a numerical constant:

$$\lim_{\sigma \to -\infty} \frac{I_{pq}}{2} + \frac{T_{pq}}{2} \geq \frac{1}{4}. \tag{30}$$

See a similar 1/4 factor due to whitening in an adaptive algorithm [9]. This bound is tight: it is reached for instance by the JADE algorithm [5, 10]. Note that $1/4T$ corresponds to a 26 dB rejection for 1000 i.i.d. samples which is quite acceptable in many applications.

### 5.3. Rejection rates

Herein, we assume circular i.i.d. signals and noise and large $T$. The distribution of $\hat{A}^H A$ can be characterized from (23). Note that matrix $(III)^-\hat{\Pi}^2$ goes to zero as $\hat{\Pi}$ converges to $\Pi$ since $III^+ = 0$ (it represents the amount of noise ‘leaking’ from the noise subspace into the signal subspace due to errors in $\hat{\Pi}$) so that for large $T$, $\hat{N}_T$ can be safely approximated by $N_T$ in (23). By the same argument as above, we conclude that, for any orthogonal invariant algorithm in spatially and temporally white Gaussian noise, the asymptotic distribution of $\hat{A}^H A$ depends only on the distributions of the sources and on matrix $\sigma(A^H A)^{-1}$.

It follows that the asymptotic rejection rates take the form

$$I_{pq} = I_{pq}(\sigma(A^H A)^{-1}, D) \quad \forall p \neq q \tag{31}$$

where $D$ represents the distributions of the sources. The functions $I_{pq}$ depend on the specific orthogonal algorithm $U$ used for the estimation of $U$, but our point here is that the dependence of the performance on the physical context is always via matrix $\sigma(A^H A)^{-1}$ which then quantifies the ‘hardness’ of the source separation problem. In particular, the significance of a given noise level may be determined by inspecting the entries of $\sigma(A^H A)^{-1}$.

As a worked out example, the JADE algorithm [5, 10] has an asymptotic performance given by

$$I_{pq} = \sum_{s=1}^{S} \sigma_s^2 (k_s^2 + k_s^2)^{-2} \tag{32}$$

$$g_{pq}^{(0)} = k_s^4 + l_p k_s^2 + l_q k_s^2 \tag{33}$$

$$g_{pq}^{(1)} = \rho_{pq} (k_s^2 + k_s^2) + \rho_{pp} (k_s^2 + 5k_p^2 + 6) + l_p k_s^2 \tag{34}$$

$$g_{pq}^{(2)} = \cdots$$

where $l_p \defeq E|s_p(t)|^4 - E^2|s_p(t)|^2$, $k_p \defeq E|s_p(t)|^2 - 2E|s_p(t)|^2$, and $p_p = (A^H A)^{-1}$. The performance index shows terms with the noise variance raised to the 4th power because the $U$ function of JADE is based on 4th-order cumulants. Only the first two terms are given here, but the following ones also involve the coefficient $|s_p|$. The lowest degree term shows the necessity of a pairwise bound since, if the $q$-th source is Gaussian (so that $k_q = 0$) and the $p$-th source has a constant modulus (so that $l_p = 0$), then $I_{pq} = 0$! This means that a Gaussian source, in this case, experiences super-rejection (numerical experiments show that its residual variance in the estimate of the $p$-th source decreases as $T^{-2}$.)

### Conclusion

The most common class of orthogonal invariant algorithms for the blind separation of independent sources has been studied in terms of performance, quantified by rejection rates $T_{pq}$. Emphasis was on the dependence of the performance on the physical parameters: the mixing matrix and the noise level. The following results were obtained:

- It exists a pairwise lower bound on the performance, introduced by the orthogonal approach.
- In the noiseless case, orthogonal invariant algorithms are fully invariant; hence, their performance does not depend on the mixing matrix; the lower bound is purely numerical for i.i.d. source signals.
- In the i.i.d. noisy case, performance depends only on the distribution of sources and on matrix $\sigma(A^H A)^{-1}$ which characterizes the ‘hardness’ of source separation. The lower bound has a simple expression (29) and does not depend on the distribution of the sources.

Proofs have been only sketched but detailed calculations are available upon request. We believe that further general results could be obtained along the lines of this paper. In particular, the dependence on $\sigma(A^H A)^{-1}$ may probably be specified: for instance, expression (34) shows that the linear term in $x^2$ in $I_{pq}$ depends on the $(p,p)$-th and $(q,q)$-th entry of $\sigma(A^H A)^{-1}$. Generalizing this would be more informative than the general dependence outlined in this paper.

### References


A State Space Method for Direction Finding of Wide-Band Emitters

Filip VANPOUCKE, Marc MOONEN

Katholieke Universiteit Leuven, Dept. of Electrical engineering - ESAT, 3001 Leuven, Belgium,
Tel/Fax: +32 16 20 09 31 / 22 18 55, E-mail: {filip.vanpoucke, marc.moonen}@esat.kuleuven.ac.be

Abstract. In this paper we elaborate on a novel efficient state space method for estimating the angles-of-arrival of multiple wide-band emitters. The method is applicable under the conditions that the sensor array consists of doublets and that the wide-band signals can be modeled as the output of a linear time-invariant system driven by white noise. We show that the number of doublets needed is lower bounded by the maximum multiplicity of the modes present in the signals and we investigate the performance of the algorithm by means of simulations.

1 Introduction

Estimating the directions-of-arrival (DOA) of wide-band emitters by means of a sensor array has many applications, e.g., in communications. A first group of wide-band algorithms reduces the wide-band estimation problem to a series of related narrow-band problems. A second group of algorithms models the sensor signals as the output of a multiple-input-multiple-output linear system driven by white noise. In this class the algorithm of Ottersten and Kailath [1] is especially appealing because of its low computational requirement. It generalizes the ESPRIT algorithm [2] for narrow-band sources. It assumes a sensor array which consists of two identical subarrays displaced by a known constant vector, but no further knowledge of the array characteristics is required. Due to this special structure, a highly nonlinear optimization to map signal subspaces onto DOA's can be replaced by matrix decompositions only. The algorithm is based on a z-domain description. The system poles and corresponding residue matrices are estimated by means of an overdetermined Yule-Walker method.

In this paper an alternative, numerically more reliable algorithm using state space descriptions, is presented. The algorithm shares some ideas with the recently developed subspace algorithms for system identification [3]. In section 2 the data model is presented. Section 3 is devoted to the novel state space direction finding method. Section 4 evaluates the algorithm by means of computer simulations.

2 Data model

Figure 1 shows the sensor array consisting of two identical subarrays, displaced over a known vector of length \( \Delta \). The sensors are pairwise identical (i.e., doublets), but their characteristics may be unknown. The time-difference-of-arrival (TDOA) between the X- and the Y-sensor of a doublet for a signal impinging from the far field under an angle \( \theta_i \), is given by

\[
\tau_i = \frac{\Delta \cdot \sin(\theta_i)}{c},
\]

where \( c \) is the propagation speed. Given estimates of the \( \tau_i \)'s, the DOA's are easily obtained from eq. (1).

The block diagram of Figure 2 presents the wide-band signal model in more detail. The \( M \times 1 \) vector \( w_i(t) \) collects the array observations due to the \( l \)-th emitter. These observations are modeled as the output of a continuous-time linear system \( S_l \) of finite order \( n_l \) driven by an independent scalar stochastic process \( v_l(t) \),

\[
\begin{bmatrix}
    z_l(t) \\
    w_l(t)
\end{bmatrix} =
\begin{bmatrix}
    A_l & B_l \\
    C_l & D_l
\end{bmatrix}
\begin{bmatrix}
    z_l(t) \\
    v_l(t)
\end{bmatrix},
\]

where \( A_l(n_l \times n_l), B_l(n_l \times 1), C_l(M \times n_l), D_l(M \times 1) \) are the system matrices of the \( l \)-th emitter and the \( n_l \)-vector \( z_l(t) \) holds the state of \( S_l \). The independent stochastic processes \( v_l(t) \) are zero-mean and almost white, having a finite correlation time which is smaller than the sampling period \( T_s \). The linear system \( S_l \) models the dynamics of the \( l \)-th emitter and the array response, which may depend on the angle \( \theta_l \). Below, we assume that \( A_l \) is given in diagonal form (which is generically possible).
The X-array outputs are collected in the M-vector
\[ x(t) = \sum_{i=1}^{D} w_i(t) + n_x(t), \]
where \( D \) is the total number of emitters and the vector \( n_x(t) \) adds independent white noise. The full state space model for \( x(t) \) is given by
\[
\begin{bmatrix}
\dot{z}(t) \\
z(t)
\end{bmatrix} =
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix}
\begin{bmatrix}
z(t) \\
v(t)
\end{bmatrix} +
\begin{bmatrix}
0_{n_x 	imes 1} \\
n_x(t)
\end{bmatrix},
\]
where the diagonal matrix \( A \ (n \times n) \) and the matrices \( B \ (n \times D), C \ (M \times n), D \ (M \times D) \), \( z(t) \), \( v(t) \) have block components given in eq. (2) \( (n = n_1 + \cdots + n_D) \).

The Y-array observations are given by
\[ y(t) = \sum_{i=1}^{D} w_i(t - \tau_i) + n_y(t), \]
where \(-\Delta \tau \leq \tau_i \leq \Delta \tau\). It is straightforward to show that \( y(t) \) can also be written as
\[ y(t) = C' \cdot e^{A(t-\tau_0)} \cdot \Phi \cdot z(t_0) + f_y(t_0,t) + n_y(t), \quad (3) \]
where \( t_0 \leq t - \tau_i \) for \( 1 \leq i \leq D \), \( \Phi(n \times n) \) is a diagonal matrix with block entries \( \phi_{ii} = e^{-\Delta \tau \tau_i} \), \( f_y(t_0,t) \) is a term summing the input contributions
\[ f_y(t_0,t) = \sum_{l=1}^{D} \int_{t_0}^{t-\tau_l} h_l(t - \tau_l - \sigma) \cdot v_l(\sigma) \cdot d\sigma, \]
and \( h_l(t) \) is the impulse response of the \( l \)-th system
\[ h_l(t) =
\begin{cases}
C_l \cdot e^{A(t-\tau_0)} \cdot B_l, & t > 0 \\
D_l, & t = 0.
\end{cases}
\]

An expression for \( x(t) \) similar to eq. (3) is easily obtained
\[ x(t) = C' \cdot e^{A(t-\tau_0)} \cdot z(t_0) + f_x(t_0,t) + n_x(t), \quad (4) \]
where
\[ f_x(t_0,t) = \sum_{i=1}^{D} \int_{t_0}^{t} h_i(t - \sigma) \cdot v_i(\sigma) \cdot d\sigma. \]

Comparing eqs. (3) and (4) we see that the terms due to the state at time \( t_0 \), only differ in the occurrence of a matrix \( \Phi \). This mathematical property is due to the doublet structure of the sensor array. Since the X- and the Y-array are identical, they share the same system matrices. The TDOA's are only visible in the exponent of the block components of \( \Phi \), reflecting the exponential decay of the state \( z(t_0) \). This observation will be exploited in the next section.

3 State space direction finding

In this section we show how the TDOA's can be estimated as a function of the corresponding generalized eigenvalues of two related matrix pencils. We need to define block Hankel matrices, consisting of data vectors \( x_k = x_k(T_j) \), e.g.,
\[
X_{ij} = \begin{bmatrix}
x_k & x_{k+1} & \cdots & x_{k+j-1} \\
x_{k+1} & x_{k+2} & \cdots & x_{k+j} \\
\vdots & \vdots & \ddots & \vdots \\
x_{k+i-1} & x_{k+i} & \cdots & x_{k+i+j-2}
\end{bmatrix}_{M \times j}
\]
\[
\frac{df}{d\sigma} = \begin{bmatrix}
x_k & x_{k+1} & \cdots & x_{k+j-1} \\
x_{k+1} & x_{k+2} & \cdots & x_{k+j} \\
\vdots & \vdots & \ddots & \vdots \\
x_{k+i-1} & x_{k+i} & \cdots & x_{k+i+j-2}
\end{bmatrix}
\]

Such block Hankel matrices, formed with the sensor outputs, can be written as a linear combination of a state matrix and an input matrix plus noise terms (see eqs. (3,4)). Define an integer \( q \geq 0 \) such that \( q \cdot T_s \geq \tau_i \) for \( l = 1, \cdots, D \). The following important (noiseless) expressions for \( X_{ij} = X_{i+j+2l-1} \) and \( Y_{ij} = Y_{i+j+2l-1} \) are trivially obtained through element-wise substitution of eq. (3,4) (e.g., [4])
\[
X_{i+j+2l-1} = \Gamma_i \cdot A^q \cdot Z_i + F_{x_i,i}(q),
\]
\[
X_{i+j+2l+2i} = \Gamma_i \cdot A^{q+1} \cdot Z_i + F_{x_i,i+1}(q), \quad (5)
\]
\[
Y_{i+j+2l-1} = \Gamma_i \cdot \Phi \cdot A^q \cdot Z_i + F_{y_i,i}(q),
\]
where \( \Gamma_i(M \times n) \) is the extended observability matrix
\[
\Gamma_i = \begin{bmatrix}
C^T & (CA)^T & \cdots & (CA^{j-1})^T
\end{bmatrix}^T, \quad A = e^{AT_i},
\]
\[ Z_i(n \times j) \text{ contains a state sequence } \]
\[ Z_i = \begin{bmatrix}
z_i & z_{i+1} & \cdots & z_{i+j-1}
\end{bmatrix}
\]
and the \((k,l)\)-th entry of \( F_{x_i,i}(q) \) is \( F_{x_i,i}(q) = f_x((t + l - 1)T_s, (i + q + k + l - 2)T_s) \).

We can eliminate the terms due to the inputs in eq. (5) by projecting these matrices onto the row space of the past observations matrix \( X_{0:i-1} \), i.e.,
\[
X_{i+j+2l-1}^p = X_{i+j+2l-1}/X_{0:i-1}
\]
\[
X_{i+j+2l+2i}^p = X_{i+j+2l+2i}/X_{0:i-1}
\]
\[
Y_{i+j+2l-1}^p = Y_{i+j+2l-1}/X_{0:i-1}
\]

781
where $A/B = \frac{1}{2} AB^T (BB^T)^{-1} B$. Under the assumption that the inputs and the measurement noise are independent zero-mean (almost) white stochastic processes, and that current states are uncorrelated with current and later inputs, we can show [5] that asymptotically

$$
\begin{pmatrix}
X_{i+j+1}[t+2i-1]
\vdots
X_{i+j+1}[t+2i-1]

Y_{i+j+1}[t+2i-1]
\vdots
Y_{i+j+1}[t+2i-1]
\end{pmatrix}
= G \cdot \begin{pmatrix}
\Gamma_i \\
\Gamma_i \cdot A \\
\Gamma_i \cdot \Phi
\end{pmatrix}
\Rightarrow \begin{pmatrix}
3M \times n \\
3M \times n
\end{pmatrix}
\tag{7}
$$

where $G = A^T \cdot Z_0 / X_{0i}[t]$. After projection, the three matrices share a common column space of dimension $n$ - the signal subspace $\mathcal{R}(\Gamma_i)$, and a common row space - the source subspace $\mathcal{R}(G^T)$, where $\mathcal{R}(Q)$ denotes the range of a matrix $Q$, i.e., the space spanned by the columns of $Q$. To reconstruct the signal subspace, it is clearly necessary that $M \cdot i \geq n$.

The final operation towards estimating the $\eta_i$'s is the computation of the non-zero generalized eigenvalues of the following two low-rank matrix pencils

$$X_{i+j+1}[t+2i-1] - \lambda X_{i+j+1}[t+2i-1] \tag{8}$$

$$Y_{i+j+1}[t+2i-1] - \mu Y_{i+j+1}[t+2i-1] \tag{9}$$

Given the eigenvalues $\lambda = e^{\eta_i}$, and $\mu = e^{-\eta_i}$ corresponding to the same eigenvalue $p$ of $A$, the corresponding time-difference-of-arrival $\eta_i$ of source $l$ can be computed as

$$\eta_i = -T_s \cdot \ln \mu / \ln \lambda.$$

To match the generalized eigenvalues of the two matrix pencils, it suffices to notice that by construction both matrix pencils in eqs. (8,9) share the same eigenvectors [6]. Therefore the $k$-th eigenvalues $\lambda_k$ and $\mu_k$ will both correspond to the same pole $p$ of the continuous-time system.

4 Relation doubles vs. emitters

In order to estimate all modes and their corresponding direction-of-arrival the rank of the pencils in eqs. (8,9) must be $n$. The diagonal matrices $A$ and $A^T$ always have rank $n$. Since the $n \times i$ matrix $G$ is a projected stochastic state sequence, its generic rank will be $n$ too. The rank of the matrix pencils thus equals the rank of the observability matrix $\Gamma_i$. The well-known Popov-Belevitch-Hautus (PBH) eigenvector test states that a matrix pair \{C, A\} is observable if and only if there exists no eigenvector of $A$ lying in the null space of $C$.

$$-\{3x \neq 0\} \text{ such that } A \cdot x = \lambda x \text{ and } C^\ast \cdot x = 0.$$ 

Below we examine this condition in detail. Let $K$ be the maximum geometric multiplicity of an eigenvalue $p_k$ of $A$ and let $S_1$ denote the corresponding eigenspace. Let $S_2$ denote the null space of $C(M \times n)$. The dimension of $S_2$ is $n - M$ if $C$ has full rank $M$. This property of ambiguity can be imposed by proper sensor array design. According to the dimension theorem of Grassmann

$$\dim(S_1) + \dim(S_2) \leq n \quad \Rightarrow \quad M \geq n - \dim(S_1 \cap S_2) \tag{9}$$

Therefore, the number of antenna doubles $M$ must be at least equal to the maximum geometric multiplicity of an eigenvalue $p$ of $A$. If no wide-band doublets are observed to have the same pole, then one sensor doublet suffices to estimate all DOAs. In the special case of narrow-band direction finding, all emitters share the same pole $p = j\omega$, where $\omega$ is the common carrier frequency. This implies that in this case there must be at least as many doubles as narrow-band emitters. The same relation doubles vs. emitters also holds for the wide-band algorithm of Ottersten [1]. There it is derived based on the rank of certain residue matrices.

5 Simulations

In this section we compare 3 algorithms. Algorithm 1 is a straightforward implementation of the state space direction finding method. First the projection of eq.(6) is performed by computing the QR decomposition of $X_0[t]$, and postmultiplying the matrices $X_{i+j+1}[t+2i-1]$, $Y_{i+j+1}[t+2i-1]$ and $Y_{i+j+1}[t+2i-1]$, with the transpose of the $Q$-factor of $X_0[t]$. To determine the signal subspace, the singular value decomposition of $X_{i+j+1}[t+2i-1]$ is truncated to rank $n$. The dimensions of $X_{i+j+1}[t+2i-1]$, $X_{i+j+1}[t+2i-1]$ and $Y_{i+j+1}[t+2i-1]$ are then compressed from $M \times i \times n \times n$ to $n \times n$ by pre- and postmultiplying the matrices with the first $n$ left- and right singular vectors. Finally the matrix pencil in eq. (8) is solved as a generalized eigenvalue problem and the eigentransformations are applied in eq. (9).

Algorithm 2 is a refined version of Algorithm 1. Due to the noise the estimated signal and noise subspaces of $X_{i+j+1}[t+2i-1]$, $X_{i+j+1}[t+2i-1]$ and $Y_{i+j+1}[t+2i-1]$ will no longer coincide. In [6] a method is described to approximate the 3 matrices by 'nearest' commuting matrices up to first order which only requires solving a set of linear equations. In algorithm 2 the compressed matrices are first approximated before finally computing the generalized eigenvalues.

Algorithm 3 is the $z$-domain based wide-band algorithm of Ottersten et al. [1].

In the simulation a Uniform Linear Array with $m = 4$ sensors and $\Delta = c \cdot T_s$ is employed. The $X$-subarray consists of sensors 1 to 3, whereas the $Y$-array consists of sensors 2 to 4. There are two second-order emitters with overlapping spectra. The first emitter has a TDOA $\tau_1 = 0.2 T_s$ and a transfer function $H_1(s) = \frac{0.894}{s^2 + 0.25 s + 2}$. The
second emitter has $\tau_2 = 0.5 \, T_x$ and $H_2(s) = \frac{s+0.9}{s+1.5}$. There were $N = 300$ snapshots taken and 100 runs with independent noise performed. The Hankel matrices have $i = 6$ block rows. For Algorithm 3 also correlation matrices up to lag $i = 6$ are used. The delay parameter $q$ is set to its minimum, i.e., 1. The estimated TDOA’s are computed as

$$\hat{\tau} = \mathcal{R}_C \left( \frac{-\ln \hat{\mu}}{\ln \hat{\lambda}} \right).$$

From Figures 3 and 4 it is clear that in this scenario Algorithm 1 has a high failure rate but a low mean square error (MSE) at higher SNR’s. The incorporation of the extra approximation step in Algorithm 2 makes the state space method outperform the z-domain based method, at a negligible increase (1%) of computation. Therefore, it is highly desirable to incorporate the extra approximation step.

Here Algorithm 2 is clearly the best choice, with the additional advantage that it requires 30% less computation than Algorithm 3. A broader simulation study is however needed to investigate the influence of all parameters on the performance of the algorithms.

6 Conclusion

A state space method has been presented to locate multiple wide-band emitters, using a sensor array consisting of doublets on condition that the wide-band signals can be modeled as the output of a linear system driven by white noise. Two matrix pencils consisting of Hankel matrices, were constructed. The first matrix pencil yielded information about the spectral content of the signals, i.e., the poles of the overall system. The second matrix pencil yielded information about the spectral content as well as the source location of the signals. Furthermore, it was shown that the number of sensor doublets is lower bounded by the maximal geometric multiplicity of a mode. Simulations indicated that the state space method may outperform the z-domain based wide-band method.

Acknowledgements

This research is partly sponsored by ESPRIT BRA 6632 of the EC. Filip Vanpoucke is a research assistant of the Belgian N.F.W.O. (National Fund for Scientific Research).

References


783
A New Algorithm For Direction Of Arrival Tracking

Mustafa KOŞAROĞLU  Yalçın TANIK

Electrical Eng. Department, Middle East Technical University, Ankara, 06531 TURKEY
E-mail:A09573@vm.cc.metu.edu.tr

Abstract. In this work, the adaptive tracking of the Direction of Arrivals (DOAs) is considered for high resolution techniques. The spatial autocorrelation matrix is updated with some forgetting through past and the changes in the subspaces are examined. The optimum choice of the forgetting parameters that yield the minimum RMS DOA error in conjunction with the MUSIC algorithm is presented. Additionally, one-step ahead correlation matrix updating scheme that can produce zero lag error is investigated. The threshold cyclic Jacobi method is adopted to track the sources adaptively. Compared to some other techniques, it does not require extra computations to satisfy the orthonormality constraint of the eigenvectors. By accumulating q observations, computations can be reduced approximately by a factor q. Considering the inaccuracy caused by the Jacobi method, a tight upper bound on the DOA error is derived. DOAs are found iteratively by Taylor series approach.

1. Introduction
There are numerous studies that have appeared in the literature for estimating the DOAs. Among them, the subspace based oriented techniques (MUSIC, Min-Norm, Root-MUSIC, etc.) draws special attention. They do not need any information related to the emitted signal. Once an estimation of the autocorrelation matrix obtained from the observation vectors is available, they are reported to work quite well. This estimation may need long observation intervals, thus creating difficulties to track time varying DOAs. This work tackles with this problem and the work done is grouped under three sections; (a) The DOA tracking problem is investigated in terms of the chosen forgetting parameters' effects, (section 3), (b) The well known Jacobi method to find the eigenvectors and the eigenvalues of a real symmetric matrix is adopted for subspace tracking (section 4), and (c) The DOA changes are derived from the subspace changes by the Taylor series approach (section 5). All the analysis is done in conjunction with the MUSIC algorithm, because of its applicability to any array structure and superiority in performance.

2. MUSIC Algorithm
We assume that there are p narrow band sources which are located sufficiently far from the receiving array of L sensors, such that planar wave assumption holds. The observation vectors from the sensors at time t, \( \mathbf{x}(t) \), are used to obtain an estimate of the autocorrelation matrix

\[
R = A(\theta)R_sA^H(\theta) + \sigma_n^2I
\]

where \( R_s \) is the signal correlation matrix, \( \sigma_n^2 \) is the noise power, \( A(\theta) \) is the steering matrix, \( H \) shows the Hermitian operator and \( p \) is the number of sources. Noise samples are assumed to be independent, identically distributed and uncorrelated with the signal. The \( m \)th column of \( A(\theta) \) is the steering vector

\[
\mathbf{a}(\theta_m) = [e^{-j2\pi d\sin\theta_m} \ldots e^{-j2\pi (L-1)d\sin\theta_m}]^T \quad m = 1, \ldots, p
\]

For instance, for a linear array of isotropic sensors placed with a distance of \( d \) \( (0 < f d/c \leq 1/2) \):

\[
y_k(\theta) = (k-1)2\pi f d/c \sin \theta \quad k = 1, \ldots, L
\]

Here, \( f \) is the common frequency of the sources, \( c \) is the speed of light and \( \theta \) is the DOA from the array normal. Assuming that \( R_s \) is nonsingular (i.e., the sources are not fully correlated), the eigenvectors of the correlation matrix (1) are grouped into two matrices. The eigenvectors corresponding to the \( p \) largest eigenvalues define the signal subspace \( E_s \), and the rest defines the noise subspace \( E_n \).

\[
E_s = [\mathbf{e}_1 \ldots \mathbf{e}_p] \quad E_n = [\mathbf{e}_{p+1} \ldots \mathbf{e}_L]
\]

The MUSIC algorithm evaluates the minimum \( p \) values of \( \theta \) where \( D(\theta) \) is minimum, as the DOA estimates [1].

\[
D(\theta) = \mathbf{a}(\theta)^H E_n E_n^H \mathbf{a}(\theta)
\]

3. Choice Of The Forgetting Parameters For DOA Tracking
When the DOAs vary, the subspaces also change. Since the DOA estimation is based on the batch estimation of the eigenvectors from an estimation of the correlation matrix, implementation of the subspace based techniques is not suitable for adaptive processing that is needed to estimate the nonstationary sources and for mobile direction finding systems. One of the widely used methods in such occasions is to update the estimated correlation matrix with a forgetting factor \( \alpha \) [2].
\(0 < \alpha < 1\);  
\[
\hat{R}(\theta(i)) = \alpha \hat{R}(\theta(i-1)) + (1 - \alpha) \hat{\beta}(i) \hat{\beta}^H(i)
\]  
and re-evaluate the eigenstructure with the new observation. Here, \(\alpha\) should be chosen as a compromise between a small DOA estimate error variance and the tracking lag error. However, there has not been a satisfactory analysis for this compromise in the literature.

The analysis for the estimate error, due to noise, assumes a stationary source and the observation vectors are Gaussian, zero mean and independent. If the correlation matrix is obtained as
\[
\hat{R}(\theta) = \frac{1}{N} \sum_{i=0}^{N} \hat{\beta}(i)\hat{\beta}^H(i)
\]
then the DOA Mean Square Error (MSE) is given by [3]
\[
E[(\hat{\theta} - \theta)^2]_{\text{noise}} = \frac{1}{N} f(\theta)
\]
\[
f(\theta) = \frac{\lambda_1 \sigma_\beta^2 \| \hat{\beta}(\theta) \|^2}{(\lambda_1 - \sigma_\beta^2) \left( \frac{\partial \hat{\beta}(\theta)}{\partial \theta} \right)^H \left( \frac{\partial \hat{\beta}(\theta)}{\partial \theta} \right)}
\]
where \(\lambda_1\) is the largest eigenvalue. The entries of (7) have complex Wishart distribution [4] with
\[
E[\hat{R}_{kk}(\theta)] = R_{kk}(\theta)
\]
\[
E[(\hat{R}_{ik}(\theta) - R_{ik}(\theta))^2] = \frac{1}{N} R_{ii}(\theta) R_{kk}(\theta)
\]
As the correlation matrix is replaced by (6)
\[
\hat{R}(\theta(i)) = (1 - \alpha) \sum_{m=0}^{i} \alpha^{i-m} \hat{\beta}(m) \hat{\beta}^H(m)
\]
the DOA MSE can be computed by finding the value of \(\alpha\) that produce the same pdf as the estimate (10,11). As \(i \to \infty\);  
\[
E[\hat{R}_{ik}(\theta(i))] = \alpha R_{ik}(\theta) + (1 - \alpha) R_{ik}(\theta)
\]
\[
E[(\hat{R}_{ik}(\theta) - R_{ik}(\theta))^2] = \frac{1 - \alpha}{1 + \alpha} R_{ii}(\theta) R_{kk}(\theta)
\]
Comparing this result to (11), it can be concluded that when \(N = (1 + \alpha)/(1 - \alpha)\), both estimates are asymptotically equivalent. Consequently, one can obtain the DOA MSE by this replacement in (8). For instance, for a linear array of isotropic sensors
\[
E[(\hat{\theta} - \theta)^2]_{\text{noise}} = \frac{1 - \alpha}{1 + \alpha} f(\theta)
\]
\[
f(\theta) = \frac{6 \sigma_\beta^2}{L \sigma_\beta^2 + 1} \left( \frac{2\pi f_0^2}{L^2 \sigma_\beta^2} \right)^2 \frac{1}{L^2 - 1} \cos \theta
\]
where \(\sigma_\beta^2\) is the source power. If the sources are sufficiently separate, this result is also approximately true for \(p > 1\) [5].

The tracking analysis assumes a moving source having the DOA at time \(i\) given by \(\theta(i) = \theta(0) + i \Delta \theta\), where \(\Delta \theta\) is the incremental DOA between the observation instants. The correlation matrix is obtained as
\[
\hat{R}(\theta(i)) = \alpha \hat{R}(\theta(i-1)) + (1 - \alpha) R(\theta(0)) + i \Delta \theta
\]
Because exact correlation matrices are used, the results are independent of SNR, \(L\), \(p\) (\(p < L\)) and the separation of the sources [6]. The source can be said to be tracked, if the estimated DOA has the same slope \(\Delta \theta\) in the steady state and the lag error is the difference between the true DOA and the estimated DOA. Finally, the two results are added to get the overall RMS error.

\[
\text{RMS}(\hat{\theta} - \theta)_{\text{total}} = \left( E[(\hat{\theta} - \theta)^2]_{\text{noise}} + \right)
\]
\[
\left[ \text{RMS}(\hat{\theta} - \theta)_{\text{lag}} \right]^{1/2}
\]
Thus, we are able to find the optimum \(\alpha\) value where RMS(\(\hat{\theta} - \theta\))_{\text{total}} is minimum for given SNR, \(\Delta \theta\) and \(L\).

The correlation matrix updating in (6) never results in zero tracking lag error, unless there are contributions from the advance observations. Consider the following non-causal, one-step ahead updating scheme to increase the tracking speed.
\[
\hat{R}(\theta(i)) = \alpha \hat{R}(\theta(i-1)) + \beta \hat{\beta}(i) \hat{\beta}^H(i)
\]
\[
+ \gamma \hat{\beta}(i+1) \hat{\beta}^H(i+1)
\]
\[
\hat{R}(\theta(i)) = \sum_{m=0}^{i} \alpha^{i-m} \beta \hat{\beta}(m) \hat{\beta}^H(m)
\]
\[
+ \gamma \hat{\beta}(m+1) \hat{\beta}^H(m+1)
\]
Having the same assumptions before for the additive noise effect, as \(i \to \infty\);  
\[
E[\hat{R}_{ik}(\theta)] = \alpha R_{ik}(\theta)
\]
\[
E[(\hat{R}_{ik} - R_{ik})^2] = \frac{1 - \alpha}{1 + \alpha} R_{ii}(\theta) R_{kk}(\theta)
\]
From (11), replacing \(N = (1 + \alpha)/(1 - \alpha + 2\beta\gamma)\), both pdf’s are asymptotically equivalent. For example, for a linear array of isotropic sensors;
\[
E[(\hat{\theta} - \theta)^2]_{\text{noise}} = \frac{1 - \alpha - 2\beta\gamma}{1 + \alpha} f(\theta)
\]
To obtain a stable system \(-1 < \alpha < 1\), from (21) \( \alpha + \beta + \gamma = 1\), and to have minimum error due to noise, which is mostly the dominant term, the reasonable choice is \(0 < \alpha, \beta, \gamma < 1\). Notice that the error power is reduced, because of the extra smoothing introduced by the advance sampling. The tracking lag error is evaluated as before by using
\[ \hat{R}(\theta(i)) = \alpha \hat{R}(\theta(i-1)) + \beta R(\theta(0) + i\Delta \theta) + \gamma R(\theta(0) + (i+1)\Delta \theta) \]  
\[ \text{(24)} \]

Finally, both errors are combined (18) to obtain the optimum \(\alpha, \beta, \text{ and } \gamma\) values.

Computer simulations have verified the theoretical results of both updating schemes for the noise effect. The theory can be extended to many-step ahead updating, thus providing smaller estimate error due to noise and the zero lagging error.

4. Subspace Tracking With The Jacobi Method

Another issue in DOA tracking is the adaptation of the subspaces of the correlation matrix with the new observations. The nonstationary sources are not expected to vary the DOA abruptly, and re-computation of the eigenstructure at each observation does not make use of the priori information. Some methods [2,7] obtain the subspace changes via an optimization problem, but they require extra computations to satisfy the orthonormality constraint of the eigenvectors. They also assume that the number of the sources are known. In [8], a two sided orthogonal decomposition (URV) is presented as an intermediary between the SVD and the QR decompositions.

The Jacobi method is a widely used technique to evaluate the eigenvectors and the eigenvalues of real symmetric matrices [9]. This method eliminates the off-diagonal entries with the basic orthonormal transformations. It can utilize the priori information, and therefore, it is very suitable to update the eigenstructure. A version of this method, the threshold cyclic Jacobi method annihilates the off-diagonals in turn until all are smaller than a predetermined threshold value, thus having an approximate diagonal matrix in the end. Since this method itself satisfy the orthonormality constraint, it does not need extra computations for that. Additionally, it also provides good estimates of the eigenvalues, and they can be used to determine the number of sources efficiently [10].

The estimated correlation matrix \( \hat{R}(\theta(i)) \) is Hermition and can be represented by a real symmetric matrix, \( \hat{R}_a(\theta(i)) \) as follows:
\[ \hat{R} = \hat{R}_{re} + j \hat{R}_{im} \quad \hat{R}_a = \begin{bmatrix} \hat{R}_{re} & -\hat{R}_{im} \\ \hat{R}_{im} & \hat{R}_{re} \end{bmatrix} \]  
\[ \text{(25)} \]

and the modal decomposition matrices are
\[ \hat{Q} = \hat{Q}_{re} + j \hat{Q}_{im} \quad \hat{Q}_a = \begin{bmatrix} \hat{Q}_{re} & -\hat{Q}_{im} \\ \hat{Q}_{im} & \hat{Q}_{re} \end{bmatrix} \]  
\[ \text{(26)} \]

\( \hat{R}_a \) has the same eigenvalues of \( \hat{R} \) with two repetitions. The eigenvectors to the second repetitions correspond to the "\( \bar{j} \)" multiplied versions of the others. Both of them can be used in the DOA estimation algorithm.

As the DOAs vary slowly, consequently, the eigenvectors do not change much with the new observation. The subspace obtained at the previous observation still contains information about the new subspaces. To use this priori knowledge, we first apply the similarity transformation
\[ \hat{Q}_a(\theta(i-1)) \hat{R}_a(\theta(i)) \hat{Q}_a(\theta(i)) = \hat{R}_a \]  
\[ \text{(27)} \]

This step makes the off-diagonals smaller, then we apply the threshold cyclic Jacobi method to get the updated subspaces
\[ V^T \hat{R}_a V = \Lambda(\theta(i)) \quad \hat{Q}_a(\theta(i)) = \hat{Q}_a(\theta(i-1)) V \]  
\[ \text{(28)} \]

Having small changes in the DOAs, the subspaces are not expected to change significantly in the consecutive observations. Considering this fact, by performing the subspace updating once every \( g \)-block of observations, the computations per observation can be reduced approximately by a factor \( g \). However, \( g \) should not be chose large, to keep the changes in the eigenstructure insignificant. With the paralel and block processing [9] and in view of VLSI, CORDIC (Coordinate Rotation Digital Corruper) based algorithms, the Jacobi method can be implemented in a faster fashion [11].

The threshold cyclic Jacobi method does not bring the off-diagonals completely to zero, thus causing an inaccuracy in the DOA estimate. A good indicator to this error can be the DOA error, when the exact correlation matrix is subject to Jacobi method. The perturbation analysis in mathematics is available for \( p = 1 \), to derive some simplified results [6]. The first order perturbation coefficients are dominant to yield
\[ |\tilde{\theta} - \theta|_{\text{inecc}} < \max \left| \text{Re} \right| \frac{\left( \frac{\partial \tilde{\theta}(\theta)}{\partial \theta} \right)^H \tilde{d} \tilde{\theta}^H \hat{R}(\theta) \left( \frac{\partial \tilde{\theta}(\theta)}{\partial \theta} \right)}{\left( \frac{\partial \tilde{\theta}(\theta)}{\partial \theta} \right)^H \left( I - \tilde{e}_1 \tilde{e}_1^H \right) \left( \frac{\partial \tilde{\theta}(\theta)}{\partial \theta} \right)} \]  
\[ \text{(29)} \]

\[ \tilde{d} = \mu \left( \frac{\pm 1 \pm j}{\sqrt{1 - \rho^2}} \right) \sum_{r=2}^{L} \tilde{e}_r \]  
\[ \text{(30)} \]

where \( \mu \) is the predetermined threshold. For a linear array of isotropic sensors, a strict upper bound is found and verified via computer simulations as
\[ |\tilde{\theta} - \theta|_{\text{inecc}} < \frac{2\pi f d}{c} \frac{24\mu}{c \cos \theta \sqrt{L^2 \rho^2 (L + 1)}} \times \left( \sum_{r=1}^{L-1} r^2 \left( \frac{2\pi f d}{c} \sin \theta \right) \right) \]  
\[ \text{(31)} \]

The chosen threshold naturally affects the number of rotations per update. On the other hand, this upper bound shows that, the desired DOA estimate accuracy can be obtained with much fewer rotations than usually used in the Jacobi method.

5. DOA Tracking With The Taylor Series Approach

As the subspaces are available in hand, re-evaluating \( D(\theta) \) for the moving sources at each observation over a grid is computationally inefficient. Taylor series is a
powerful tool in such cases. Defining the DOA change per step as $\Delta \theta_i (i = 1, \ldots, p)$ and using the Taylor series expansion of $D(\theta(i))$ around $\theta_i(i-1)$

$$\theta_i(i) = \theta_i(i-1) + \Delta \theta_i$$

$$D(\theta(i)) = \sum_{r=p+1}^{L} \alpha^{(r)}(\theta) \epsilon_r(i) \epsilon^{(r)}(i) \alpha(\theta)$$

The first and the second order approximations yield, respectively,

$$\Delta \theta_1(i) = -\frac{b_1}{b_2}$$

$$\Delta \theta_2(i) = \frac{-b_2 \pm \sqrt{b_2^2 - 2b_1b_3}}{b_3}$$

$$b_r = \frac{\partial^r D(\theta)}{\partial \theta^r} \bigg|_{\theta = \theta_i(i-1)}$$

In the second order approximation, there appear two solutions and the sign is chosen to have the smaller change. Once the initial DOAs are available, consecutive DOA changes are added to obtain the DOA estimates without needing a grid. The first order approximation is similar to the approximate Newton method, also including the neglected second order derivatives in [12]. Computer simulations for a linear array of isotropic sensors indicated that the second order approach produces more reliable results, especially when SNR is low.

References


Optimally Weighted ESPRIT for Direction Estimation

Anders ERIKSSON, Petre STOICA†, and Torsten SÖDERSTRÖM

Systems and Control Group, Uppsala University, P O Box 27, S-751 03 Uppsala, Sweden. E-Mail: ae@SysCon.uu.se

† On leave from the Department of Automatic Control, Polytechnic Institute of Bucharest, Romania.

Abstract. A weighted least-squares (WLS) ESPRIT estimate of the direction to narrowband signals impinging on an array of sensors is considered. The variances of the WLS ESPRIT direction estimates are derived, and a statistically optimal weighting matrix is proposed. The optimal WLS ESPRIT estimates of the unknown directions are shown to be significantly more accurate than the commonly-used least squares ESPRIT estimates.

1. Introduction

The statistical performance of the ESPRIT [3, 4] estimates of the directions-of-arrival (DOAs) of narrowband signals impinging on an array of sensors has been analysed in [8], where it was shown that these estimates may be statistically quite inefficient for arrays with many sensors. Additionally, the performance of ESPRIT is known to degrade significantly when the source signals are highly correlated. The accuracy in these scenarios can be improved by introducing weighted estimates and choosing the weights so as to minimize the estimation errors. In this paper we will do so by extending the ideas presented in [8]. First, the ESPRIT estimation problem is re-organized as a system of linear regression equations, which is solved using a weighted least-squares (WLS) procedure. An expression for the variance of the estimated directions is derived, and the weight in the WLS estimate is chosen so as to minimize the variances of the estimated directions. This idea leads to an optimally-weighted ESPRIT algorithm which significantly outperforms the commonly-used unweighted ESPRIT in difficult scenarios (such as cases with highly correlated sources).

2. Optimally Weighted ESPRIT

The problem of estimating the DOAs of n narrowband signals impinging on an array of m sensors (see, e.g., [3, 4, 7, 8]) can be formalized as determining the unknowns \( \{ \theta_i \}_{i=1}^n \) in the signal model

\[
y(t) = Ax(t) + e(t)
\]

where \( y(t) \in \mathbb{C}^{m \times 1} \) is the measured array output signal, \( x(t) \in \mathbb{C}^{n \times 1} \) is the source signal, \( e(t) \in \mathbb{C}^{m \times 1} \) is a noise vector, and

\[
A = (a(\theta_1) \cdots a(\theta_n))
\]

where \( a(\theta) \in \mathbb{C}^{m \times 1} \) is the array response to a unit-amplitude signal impinging from direction \( \theta \). It is assumed that \( \text{rank}(A) = n < m \), and that the signal and noise vectors are independent random variables with

\[
E(x(t)x^H(s)) = P \delta_{t,s}, \quad P > 0; \quad E(x(t)x^T(s)) = 0
\]

and

\[
E(e(t)e^H(s)) = \sigma_I \delta_{t,s}; \quad E(e(t)e^T(s)) = 0.
\]

Hence, the covariance matrix of the array output vector equals

\[
R \triangleq E(y(t)y^H(t)) = APA^H + \sigma^2 I.
\]

In the following, a brief review of the ESPRIT estimation procedure is given. For a more detailed derivation, we refer to the original works [3, 4]. Consider the eigendecomposition of \( R \),

\[
R = APA^H + \sigma^2 I = \sum_{k=1}^m \lambda_k s_k^H s_k \triangleq S A S^H + \sigma^2 G G^H
\]

where \( \lambda_k \) are the eigenvalues of \( R \) sorted in non-increasing order, \( s_k \) are the corresponding orthonormal eigenvectors, and

\[
S = (s_1 \cdots s_n), \quad G = (s_{n+1} \cdots s_m),
\]

\[
\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n).
\]

Since \( S^H G = 0 \) and \( S^H S = I \), it follows from (6) that \( APA^H S = S \Lambda \) where \( \Lambda = \Lambda - \sigma^2 I \). Hence, the range spaces of \( S \) and \( A \) coincide, which can be written as

\[
S = AC
\]

where

\[
C = PA^H S \Lambda^{-1}, \quad C^{-1} = S^H A.
\]

In the following, the superscripts "T", "H", and "*" denote transpose, complex conjugate transpose, and complex conjugate, respectively.
The ESPRIT method is based on the assumption that the matrix \( A \) possesses a certain displacement invariance (this means that the array should consist of two identical sub-arrays, separated by some known displacement vector), which mathematically can be written as

\[
J_1 A \Omega = J_2 A
\]

where \( J_1 \neq J_2 \) are two matrices of dimension \( \hat{m} \times m \) (\( \hat{m} < m \)) and

\[
\Omega = \text{diag}(e^{i\omega_1}, \ldots, e^{i\omega_n}).
\]

Hereafter, \( \{\omega_i\}_{i=1}^n \) are the so-called spatial frequencies, which relate to the signal directions in a simple way ([3, 7, 8]). For the case of a uniform linear array (ULA), the displacement matrices for maximum overlap sub-arrays are given by

\[
J_1 = (I_{m-1} \quad 0), \quad J_2 = (0 \quad I_{m-1})
\]

where \( I_{m-1} \) is the identity matrix of dimension \( m - 1 \). For this case, the spatial frequencies are related to the DOAs as \( \omega_k = \pi \sin(\theta_k) \) (\( \{\theta_k\} \) are measured with respect to the normal to the array).

From (7) and (8) it follows that \( J_1 SC^{-1} \Omega = J_2 SC^{-1} \) and thus

\[
S_1 \Theta = S_2
\]

where

\[
S_1 = J_1 S, \quad S_2 = J_2 S, \quad \Theta = C^{-1} \Omega C.
\]

Note that the eigenvalues of \( \Theta \) equal those of \( \Omega \), namely \( \{e^{i\omega_j}\} \). The ESPRIT estimates of the spatial frequencies \( \{\omega_k\} \) (and hence of the DOAs \( \{\theta_k\} \)) are obtained as the angular positions of the eigenvalues of a solution to the sample counterpart of (10),

\[
\hat{S}_1 \Theta \simeq \hat{S}_2
\]

where \( \hat{S}_1 = J_1 \hat{S} \) and \( \hat{S}_2 = J_2 \hat{S} \). \( \hat{S} \) is an estimate of the principal eigenvector matrix, obtained from an eigendecomposition of the sample covariance matrix

\[
\hat{R} = \frac{1}{N} \sum_{t=1}^{N} y(t) y^H(t).
\]

Several methods for solving the perturbed system of linear equations (11) have been proposed in the literature, including Least-Squares (LS) and Total-Least-Squares (TLS) procedures, see e.g. [3, 4, 8]. However, to solve (11) in a statistically sound manner, the statistical properties of the errors \( \hat{S} - S \) should be taken into account. This can be done by introducing a residual error matrix \( \Xi \),

\[
\hat{S}_1 \Theta = \hat{S}_2 + \Xi
\]

and solving (13) using a weighted least squares procedure. First, (13) is rewritten as a linear regression equation,\(^2\)

\[
(I_n \otimes \hat{S}_1) \text{vec}(\Theta) = \text{vec}(\hat{S}_2) + \text{vec}(\Xi)
\]

or, with obvious definitions of \( \hat{\Theta}, \hat{\theta}, \hat{Y} \) and \( \xi \),

\[
\hat{\Theta} = \hat{Y} + \xi.
\]

A general weighted least-squares (WLS) estimate of \( \hat{\theta} \) is obtained as

\[
\hat{\theta}_{WLS} = (\hat{\Phi}^H W \hat{\Phi})^{-1} \hat{\Phi}^H W \hat{Y} \triangleq \hat{\Phi}^H \hat{Y}.
\]

In the following, the covariances of the estimated spatial frequencies \( \{\omega_k\} \) obtained from the WLS estimate of \( \hat{\theta} \) are derived, and an optimal choice of the weighting matrix \( W \) in (15) is proposed.

Straightforward calculations, see Appendix A, show that the errors in the estimated spatial frequencies \( \{\hat{\omega}_k\} \) relate, within a first order approximation, to the errors in the estimated parameter vector as

\[
\hat{\omega}_k - \omega_k = \text{Im} \left\{ \beta_k^H (\hat{\theta} - \theta) \right\}
\]

where \( \beta_k \) is defined in (33) in Appendix A.

The statistical properties of the ESPRIT method are readily derived from (16). Since \( \text{Im}\{a\} - \text{Im}\{b\} = (\text{Re}\{ab^*\} - \text{Re}\{ab\})/2 \), it follows that

\[
E\left\{ (\hat{\omega}_j - \omega_j)(\hat{\omega}_k - \omega_k) \right\} = \frac{1}{2} \left( \text{Re} \left\{ \beta_j^H E\{ (\hat{\theta} - \theta)(\hat{\theta} - \theta)^H \} \beta_k \right\} \right. \\
- \left. \text{Re} \left\{ \beta_j^H E\{ (\hat{\theta} - \theta)(\hat{\theta} - \theta)^T \} \beta_k^T \right\} \right\}.
\]

Noting that \( \hat{S}_1 - \hat{S}_2 = 0 \), and using the following result (see, e.g., [5]),

\[
\text{vec}(ABC) = (C^T \otimes A) \text{vec}(B),
\]

the residual error vector in (14) can be written as

\[
\xi = \text{vec}(\Xi) = \text{vec}(J_1 (\hat{S} - S) \Theta) - \text{vec}(J_2 (\hat{S} - S)) = ((\Theta^T \otimes J_1) - (I_n \otimes J_2)) \text{vec}(\hat{S} - S)
\]

\[
\Delta \triangleq K \text{vec}(\hat{S} - S).
\]

Hence, the asymptotic covariance matrices of the parameter vector appearing in (17) can be found as (see [5, Complement: C4.4])

\[
E\{ (\hat{\theta} - \theta)(\hat{\theta} - \theta)^H \} = Z^H K E\{ \xi \xi^H \} Z
\]

\[
= Z^H K E\{ \text{vec}(\hat{S} - S)^T \text{vec}(\hat{S} - S) \} K^H Z
\]

\[
E\{ (\hat{\theta} - \theta)(\hat{\theta} - \theta)^T \} = Z^H K E\{ \xi \xi^T \} Z^*
\]

\[
= Z^H K E\{ \text{vec}(\hat{S} - S)^T \text{vec}(\hat{S} - S) \} K^T Z^*
\]

\(^2\)In what follows, \( \otimes \) stands for the Kronecker product and \( \text{vec}(X) \) denotes the column vector \((x_1^T \ldots x_n^T)^T\) where \( x_k \) is the kth column of the matrix \( X \).
where \( Z \) is defined similarly to \( \hat{Z} \) in (15), but with \( \hat{S} \) replaced by \( \hat{S} \).

Under the assumptions made, the covariance matrices of the sample eigenelements \( \{s_k\} \) are given by, see, e.g., [2],

\[
E\{\hat{e}_j - s_j\}(\hat{e}_k - s_k)^H = \frac{1}{N} \sum_{i j \neq j}^m \frac{\lambda_i \lambda_j}{(\lambda_j - \lambda_i)^2} s_i s_j^H \delta_{j,k} \\
= \frac{1}{N} \sum_{i j \neq j}^m \frac{\lambda_i \lambda_j}{(\lambda_j - \lambda_i)^2} s_i s_j^H \delta_{j,k}.
\]  

(22)

Exploiting the structure of the expressions (20) and (21), the expression (17) for the covariances of the estimated spatial frequencies can be significantly simplified. In [1], it is shown that

\[
\beta_j^H Z^H K (\gamma \otimes A) = 0
\]  

(24)

where \( \gamma \triangleq (\gamma_1 \ldots \gamma_n)^T \) and \( \{\gamma_k\} \) are arbitrary constants. Post-multiplying (24) by \( C \) yields

\[
\beta_j^H Z^H K (\gamma \otimes S) = 0.
\]  

(25)

From the expression (22), it is seen that the \( k \)th column block of \( E\{\text{vec}(\hat{S} - S)\text{vec}^H(\hat{S} - S)\} \) equals

\[
\begin{pmatrix}
\alpha_{1,k} s_k s_k^T \\
\vdots \\
\alpha_{n,k} s_k s_k^T
\end{pmatrix}
\]  

(26)

where \( \alpha_{j,k} = -(1/N)(\lambda_j \lambda_k)/(\lambda_j - \lambda_k)^2 (1 - \delta_{j,k}) \). Thus, the columns of (26) have the form \( (\gamma \otimes s_k) \), which, together with (25) shows that the second term in (17) vanishes. Similarly, from the expression (22), it is seen that the \( k \)th column block of \( E\{\text{vec}(\hat{S} - S)\text{vec}^H(\hat{S} - S)\} \) equals

\[
\begin{pmatrix}
\frac{1}{N} \sum_{i j \neq j}^m \frac{\lambda_i \lambda_j}{(\lambda_j - \lambda_i)^2} s_i s_j^H + \frac{\lambda_k}{\lambda_k - \lambda_j} s_j s_j^H \\
\frac{1}{N} \sum_{i j \neq j}^m \frac{\lambda_i \lambda_j}{(\lambda_j - \lambda_i)^2} s_i s_j^H + \frac{\lambda_k}{\lambda_k - \lambda_j} s_j s_j^H
\end{pmatrix}
\]  

A generic column of the subject column block equals

\[
\sum_{i=1}^n (\gamma_i \otimes s_i) + \frac{\sigma^2}{N} [\Lambda^{-2}]_{k,k} [GG^H]_{k,k}
\]  

(27)

for appropriately defined vectors \( \{\gamma_i\} \), and where \( [\cdot]_{k,k} \) denotes the \( k,k \)th element of the matrix in question, and \( [\cdot] \) denotes a generic column. In view of (25), the contribution of the first part of (27) to the overall expression vanishes. Using the identity \( GG^H = I - A(A^H A)^{-1} A \) in the second part of (27), and making use of (25) once more, yields the final expression for the covariances of the estimated spatial frequencies,

\[
E\{\hat{\omega}_j - \omega_j\} = \frac{\sigma^2}{2N} \beta_j^H Z^H K (\Lambda^{-2} \otimes I) K^H Z \beta_k.
\]  

(28)

Thus, the statistically optimal choice of weighting matrix \( W \) in (15) is given by (see, e.g., [5, Section 4.3]),

\[
W_{\text{opt}} \triangleq (K(\Lambda^{-2} \otimes I) K^H)^{-1}.
\]  

(29)

Since the optimal WLS estimate of \( \theta \) depends on unknown quantities, it cannot be implemented directly. However, the method can be implemented using the following two-step procedure which has the same asymptotic properties as the optimal estimator:

1. Compute the LS ESPRIT estimate of \( \Theta \) and form an estimate of the matrix \( K, \hat{K} \). Use the sample eigenvalues \( \{\lambda_i\} \) to form an estimate of \( \Lambda \). Combine \( \hat{K} \) and the estimates of the sample eigenvalues to obtain a consistent estimate \( \hat{W}_{\text{opt}} \) of \( W_{\text{opt}} \).

2. Use \( \hat{W}_{\text{opt}} \) in (15) to compute the statistically optimal estimate of \( \theta \). Determine estimates of the DOAs from the angular positions of the eigenvalues of the estimate of \( \Theta \) made from the WLS estimate of \( \theta \).

3. Examples and Concluding Remarks

In the following a numerical example is given which lends support to the theoretical expressions derived, and show that the improvement may be significant when using the optimally weighted ESPRIT estimate compared to using the standard least squares ESPRIT. The theoretical standard deviations for the optimally weighted ESPRIT, obtained from (28) with \( W \) given by (29), are compared to the empirical root-mean-square error values (RMSE) obtained from 100 Monte Carlo simulations, and also to the Cramér-Rao lower bound (CRB), [7]. For comparison purposes, also the empirical performance of the MUSIC algorithm on the same data is considered.

A uniform linear array (ULA) composed of \( m \) elements separated \( \lambda/2 \) apart (where \( \lambda \) is the wavelength of the incoming signals) is considered. The displacement matrices \( J_1 \) and \( J_2 \) are chosen as in (9).

Two equally strong signals, with an SNR of 10dB, are impinging on the array at angles \( -3^\circ \) and \( 5^\circ \), respectively. The directions \( \theta_k \) are estimated from 400 snapshots of the array output. In Figure 1, the influence of the signal correlation, \( \rho \), and the dimension of the array, \( m \), on the accuracy of the estimates is illustrated. The correlation factor is varied from 0 to 0.99, and \( m \) from 4 to 32. The 'surface' in the diagram represents the ratio between the theoretical standard deviation of
Figure 1: RMSE versus source correlation and array dimension for two source.

the optimally weighted ESPRIT and the CRB. From the diagram it is seen that for uncorrelated signals, the optimal WLS ESPRIT attains the CRB whereas the standard deviation of the LS ESPRIT is slightly larger. The difference in accuracy between the LS ESPRIT and the optimal WLS ESPRIT is more pronounced for highly correlated sources. For the case of a correlation coefficient equal to 0.99 and an array composed of 28 elements, the gain in accuracy of the optimal WLS ESPRIT is a factor 20. As might have been expected, MUSIC behaves similarly to the LS ESPRIT, and hence it is also significantly outperformed, in the above scenarios, by the optimally-weighted ESPRIT introduced herein.

Acknowledgements

This work has been supported by the Swedish Research Council for Engineering Sciences under contracts no 91-676 and 93-669.

References


Appendix A

Let $\rho_k = e^{i\omega_k}$ denote the $k$th eigenvalue of $\Theta$ and let $\hat{\rho}_k$ denote the corresponding eigenvalue of $\hat{\Theta}$, where $\hat{\Theta}$ is an estimate of $\Theta$ formed from the estimated parameter vector $\hat{\theta}$. For $\hat{\rho}_k$ close to $\rho_k$,

$$\hat{\omega}_k = \text{Im} \{ \ln \hat{\rho}_k \} = \text{Im} \{ \ln(\hat{\rho}_k - \rho_k + \rho_k) \}$$

$$= \text{Im} \left\{ \ln \rho_k + \ln \left( 1 + \frac{\hat{\rho}_k - \rho_k}{\rho_k} \right) \right\}$$

$$\simeq \omega_k + \text{Im} \{ e^{-i\omega_k} (\hat{\rho}_k - \rho_k) \}. \quad (30)$$

The perturbations in the eigenvalues are related, to within a first order approximation, to the perturbations in the matrix elements through (see, e.g., [6, Chap. 6, Th. 4.2])

$$\hat{\rho}_k \simeq \rho_k + e_k^T C \left( \hat{\Theta} - \Theta \right) C^{-1} e_k \quad (31)$$

where $e_k$ is the $k$th column of the identity matrix of dimension $n$. In (31), $e_k^T C$ and $C^{-1} e_k$ are left and right eigenvectors, respectively, of $\Theta$, associated with the eigenvalue $e^{i\omega_k}$. Using (18) gives

$$\hat{\rho}_k - \rho_k \simeq (e_k^T C - T \otimes e_k^T C)(\hat{\Theta} - \Theta). \quad (32)$$

For notational convenience, let

$$\beta_k^H \triangleq e^{-i\omega_k} (e_k^T C - T \otimes e_k^T C). \quad (33)$$

Then, from (30), we have asymptotically in $N$

$$\hat{\omega}_k - \omega_k = \text{Im} \left\{ \beta_k^H (\hat{\Theta} - \Theta) \right\}$$

which is the sought relation between the parameter vector and the spatial frequencies.

791
An Architecture of One-Class-One-Net Perceptrons Based Human Face Recognition System

Peir-Ren HONG†, Jyh-Bin REN †, Shuh-Chuan TSAI †∥, and Bin-chang CHIEU†

†Department of Electronic Engineering, National Taiwan Institute of Technology 43, Keelung Road, Section 4, Taipei, Taiwan, Republic of China
∥Department of Electrical Engineering, National Yulin Institute of Technology, Touliu, Yulin 640, Taiwan, Republic of China

Abstract. We apply the clustering and parallel distributed property of neural network to the recognition of different facial expressions from 300 persons. A modified architecture of perceptron “one-class one-net” is adopted as training and recalling machine and it is proved very powerful for the identification of human face task.

1. Introduction

Face recognition has become an important issue in many applications such as security system, credit-card verification and criminal identification. Though it is clear that people are good at face identification, it is not at all obvious how faces are encoded or stored, much less how familiarity affects these complex process.

The conventional techniques used in face identification have two categories: one is based on human profiles [1,2] and the other is based on face features [3,4,5].

In the former category, Kaufman and Breeding [2] performs human face identification on the basis of feature vectors extracted from profile silhouettes. Each silhouette is preprocessed to remove noise, smooth the edges, and extract the front edges. The processed silhouettes were represented by a 12 dimensional feature vectors, the components of which were obtained by a circular autocorrelation function. Using a weighted k-nearest neighbor decision rule, this method obtained a 90% recognition accuracy in ten-class problem. Harmon and Hunt [1] proposed a recognition system based on an 11-dimensional vectors, the components of which included the distance between feature points, areas, angles and curvatures. In this system, however, each component is calculated by using feature points (nose tip, chin, forehead, bridge, nose bottom, throat, upper lip, mouth, lower lip) extracted from manually entered profile traces.

The second class of face identification system is based on feature points extracted from a front view of the face. The feature points include the corner points of the eyes, eyebrows, mouth, tip of the nose, etc. Sakai et al [3] developed a powerful face identification system which automatically selected and measured the appropriate features by using context-controlled picture processing technique and obtained a 91% recognition accuracy. Turk and Pentland [6] used “Eigenfaces” as the features for human face recognition. The eigenfaces are obtained by principle-component analysis. Hong [7] proposed a human face recognition method in which the singular value vectors are used as the features and he proved that the singular value vectors have some good properties.

The recognition of human face needs to apply the clustering method which essentially have two different ways: unsupervised and supervised. For the former approach, there are Batchelor & Wilkin’s method, H-means method, neighborhood function, covering method and Kohonen’s self-organization feature maps, etc. On the other hand, for the later approach, there are, multi-layer perceptron and Hidden Markov Model, etc.

Back propagation algorithm [8] is one of the most popular training algorithms implemented on perceptrons of neural networks. This algorithm has been used by Kusugi in face recognition system [9]. However, the back-propagation algorithm is handicapped by many implementational difficulties including slow convergence, existence of local minima and has to completely retrained whenever a new class is added that is without extensible property. One possible reason for these difficulties is that the perceptron is inherently a massively interconnected network with a large number of connection weights that need to be determined.

In our study of human face recognition, the OCON (one class one net neural network) [10,11], is used as a classifier to recognize the face features which are extracted from the front view of human faces. The OCON is proposed because it has many advantages, for example, faster convergence and much stronger discrimination ability than conventional multi-layer perceptron [10,11].

In the remainder of this paper, we will first describe the procedure of feature extraction followed by discussion of the OCON system and its corresponding learning algorithm. The experimental results and the conclusions will be presented finally.

2. Feature Extraction

Images of human faces are acquired by CCD camera and may be transformed to get normalized contrast and average grey level by histogram processing. To avoid the influence of hair styles, the hair’s portion is cut from the original image. The center of the face is fixed, and the size of the face is kept as constant as possible. After that, the face is partitioned into several non-overlapping rectangular blocks.
as shown in Figure 1. At each image block, the block mean and block variance are calculated. The means and variances of these image blocks are used as features for that specified human face.

3. One Class One Neural Network
The architecture of the OCON is shown in Figure 2. The main feature of the architecture is that the entire network is composed of subnetworks of single output perceptrons and each of the subnetworks represents one class. The major advantage of OCON’s architecture for information processing is that the information is processed at a greater extent of parallelism and distribution. In pattern classification and recognition applications, each class is realized by one subnetwork which is independent of all other subnetworks. Training and retrieving of the connection weights of these subnetworks are thus naturally parallel processing.

4. OCON Training and Retrieving Algorithm
For notational purposes, the k-th subnetwork is denoted by $\beta^{[k]} = (W^{[k]}, \theta^{[k]})$, where $W^{[k]}$ are the connection weights and $\theta^{[k]}$ are the thresholds. Given the input feature vector $I = (I_1, \ldots, I_m)$, we have the following definition.

$h_i^{[k]}$: i-th hidden node output value in the subnetwork $\beta^{[k]}$.

$w_{ij}^{[k]}$: the weights (reciprocal of covariances, i.e., $1/\sigma_{ij}$) from j-th input node to i-th hidden nodes in the subnetwork $\beta^{[k]}$.

$\overline{w}_i^{[k]}$: the weights from the i-th hidden node to the output node in the subnetwork $\beta^{[k]}$.

$\gamma^{[k]}$: the center coordinate with respect to $w_{ij}^{[k]}$ in the subnetwork $\beta^{[k]}$.

$\overline{\gamma}^{[k]}$: the bias (threshold) of the output node in the subnetwork $\beta^{[k]}$.

$\eta^{[k]}$: the learning rate in the subnetwork $\beta^{[k]}$.

$\varepsilon$: the given error tolerance.

In our application, the radial based function is used to characterize the hidden neurons

$$h_i = f_h \left( \sum_j \left[ (I_j - \alpha) / \sigma_{ij} \right]^2 \right)$$

(1)

where $f_h$ is

$$f_h(x) = \exp(-x^2/\alpha^2)$$

In this case, $\alpha$ is the center of the hyperellipsoid $f_h$ while $\sigma_{ij}$ is the length of the j-th semiaxis. Output values of hidden nodes determined by Eq. (2) are

$$h_i^{[k]} = f_h \left( \sum_j \left[ (w_{ij}^{[k]} I_j - \gamma_i^{[k]})^2 \right] \right)$$

$$= \exp \left( -\frac{1}{2} \sum_j \left[ (w_{ij}^{[k]} I_j - \gamma_i^{[k]})^2 \right] \right)$$

and output value of the k-th network

$$O^{[k]}(I) = \sum_i \left( \overline{w}_i^{[k]} h_i^{[k]} + \overline{\gamma}^{[k]} \right)$$

$$= 1 / \left[ 1 + \exp \left( -\sum_i \left( \overline{w}_i^{[k]} h_i^{[k]} + \overline{\gamma}^{[k]} \right) \right) \right]$$

In the consideration of training and retrieving, the values of the output neurons are assigned as follows: when the input pattern belongs to the k-th class, the desired output of the k-th subnetwork is one while all of other subnetworks are zero. However, in application of pattern classification, it is not necessary to strictly confine the desired output value to be 1 or 0. We can relaxedly assign the actual output values $O^{[k]}(I)$ to be $\sigma = \theta + \varepsilon$ or $\gamma = 1 - \varepsilon$ as long as $0 \leq \sigma, \gamma \leq 1$. Given the input feature vector I, how can we evaluate the value $O^{[k]}(I)$ and seek the best feature match among the subnetworks? The answer is that we are seeking the one which has the highest output value for all $1 \leq k \leq M$ in the
M class classification problem. In practice this can be implemented with a maxnetwork as shown in Figure 3. The maxnetwork simply chooses the subnetwork of highest output value or equivalently the best matching subnetwork to represent the input pattern.

For each pattern \( I \) in the \( k \)-th subnetwork, the error between the desired output and actual output is
\[
E^{[k]}(I) = 1 - O^{[k]}(I) \quad \text{if pattern } I \text{ belongs to } k\text{-th class,}
\]
and
\[
E^{[k]}(I) = 0 - O^{[k]}(I) \quad \text{otherwise.}
\]
Thus, the error signal, \( \delta^{[k]} \), for the output of \( k \)-th subnetwork is given by
\[
\delta^{[k]} = (1 - O^{[k]}(I)) O^{[k]}(I) (1 - O^{[k]}(I)),
\]
if patterns belong to the \( k \)-th class.
\[
\delta^{[k]} = (0 - O^{[k]}(I)) O^{[k]}(I) (1 - O^{[k]}(I)), \quad \text{otherwise.}
\]
The error signal, \( \delta_i^{[k]} \), for hidden node \( h_i^{[k]} \) is given by
\[
\delta_i^{[k]} = - \delta^{[k]} w_i^{[k]} h_i^{[k]}.
\]
The update of weights and thresholds at time \( t \) are
\[
\Delta w_{ij}^{[k]} = w_{ij}^{[k](t+1)} - w_{ij}^{[k](t)}
\]
\[
= \eta^{[k]} \delta_i^{[k]} w_{ij}^{[k]} (I_j - c_{ij}^{[k]}) (I_j - c_{ij}^{[k]})
\]
\[
\Delta c_{ij}^{[k]} = [c_{ij}^{[k](t+1)} - c_{ij}^{[k](t)}]
\]
\[
= - \eta^{[k]} \delta_i^{[k]} w_{ij}^{[k]} h_i^{[k]} (I_j - c_{ij}^{[k]})
\]
\[
\Delta w_i^{[k]} = w_i^{[k](t+1)} - w_i^{[k](t)}
\]
\[
= \eta^{[k]} \delta_i^{[k]} h_i^{[k]}
\]
\[
\Delta \theta_j^{[k]} = \theta_j^{[k](t+1)} - \theta_j^{[k](t)}
\]
\[
= \eta^{[k]} \delta_j^{[k]}.
\]
This training procedure is repeated until \( E^{[k]}(I) < \epsilon \) for every training pattern.

5. Simulation and Conclusion
In this experiment, we have presented 5 different facial expression images and one defocused image for each person to the OCON system. Figure 4 shows the typical defocus image and Figure 5 shows the typical different facial expression images. The database we have created contains 300 persons.

Among the five different facial images we randomly pick 4 images to be our training patterns, the last one image and the defocused image are used as the testing patterns. Because the possible combinations for the training and testing patterns may be infinitive, it is impossible to implement in this way. We therefore reasonably decide to repeat this experiment for 5 random sampling cases. We need to claim here that we take the average test accuracy as our final accuracy.

Because each subnetwork represents a specified person's face and there are no interconnections among the subnetworks, thus the input data of 300 specified persons'
faces are to be classified into 300 distinct classes. The network will contain exact 300 subnetworks. Note further that all subnetworks have the same number of input neurons as this number is determined by the number of features of input patterns. Each specified person's face has 336 features, so that the number of input neurons is 336. Also, for the sake of uniformity and regularity, all subnetworks may have the same number of hidden neurons. Here we picked 45 hidden neurons.

From the simulation, the recognition rate for the inside test is 100%, while that for the outside test is 95%. The primary recognition rate is satisfactory and this performance is good enough for an assurance of further modification or real world industrial applications. Practically, the most important merits of this scheme are that the training of each subnet converges very easy and very fast if compared to the conventional multi-layer perceptron (MLP) which has also been tested under our human face database. From the simulation, it is proved that MLP never converge. As for the future work, the fuzzy set extension of OCON is our interest.

References


Artificial Neural Networks for Attitude Determination from Visual Image Processing

Juan SEIJAS\(^4\), José L. SANZ-GONZALEZ\(^3\)

\(^3\) SENER Ingeniería y Sistemas, Parque Tecnológico de Madrid, 28760 Tres cantos.
SPAIN, Tel/Fax [+34] 807 7300 / 807 7201

\(^1\) Departamento de Ingeniería de Circuitos y Sistemas, E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid, Ciudad Universitaria, 28040 Madrid SPAIN Tel/Fax [+34] 549 5700 (ext 384) / 543 9652, E.Mail:sanz@ics.upm.es

Abstract. This paper presents an optimised implementation of a Neural Network Architecture used for spacecraft attitude and position determination. The proposed solution consists of a Neural Network structure configuration and a training algorithm selection. A new algorithm (Basic-Evolutive) is presented, this training algorithm is capable of setting the appropriate dimension of the neural network and the adequate weights interconnecting the neurons. The recommended solution is derived from a very wide set of different manoeuvres simulations, these simulations provide the training and testing sets for Basic-Evolutive algorithm performance measurement. The algorithm performances are also contrasted with backpropagation training algorithm performances.

1. Introduction
The application works determining the relative position and attitude of a passive spacecraft (called target) that it is seen by a camera placed in another spacecraft (called chaser) during the approaching manoeuvre of both spacecrafts. The relative movement between the object and the camera has six degrees of freedom: three for determining the Xc, Yc, Zc, Zc, \(\phi\), \(\theta\), \(\gamma\). The target has a Coupled Charged Device (CCD) matrix of 512x512 pixels and a lens that focuses the target. Because of the relative movement of the object in some cases part of the object image exceeds the visual field of the camera and the target temporarily loses some of its contour lines and vertices. This "visual saturation effect" causes an important non-linearity in the input of the system solved by the Neural Network implementation.

2. Structure and Initial Values Selection
When a Neural Network (NN) has more than one output, and one or more hidden layers, the weight modification carried out in order to reduce the error in one output, may cause an increment in the error of another output. This error-cross-coupling among outputs requires more iterations in the training algorithm. Moreover, the preprocessing done for accommodating the inputs in the way of improving the error rate in one output, could not be the most adequate for the same purpose in other output. Because of these reasons, we have decided to implement a separate neural network for each output and, thus, for obtaining the six outputs (Xc, Yc, Zc, \(\phi\), \(\theta\), \(\gamma\)), six different neural networks, grouped in a Neural Network Lattice (NNL), are implemented. In our initial works, implementing only one NN with 6 outputs, backpropagation required 10 times more iterations to obtain 5% mean square (m.s.) error.

In the beginning of the training, each NN of the NNL has two hidden layers implemented (more hidden layers are considered in theory [1], but do not perform better); however, after the pruning process, one hidden layer may disappear.

The target information is obtained by the camera, and an initial preprocessing is carried out in order to extract its border lines and vertices. In our simulations, the target is delimited by 5 vertices. Simulations with more vertices, up to 20 where carried out, but our results demonstrate that more inputs do not contribute to improve the performance and increase quadratically the simulation time. Four vertices could be enough but, if one of them exceeds the CCD surface, the problem could admit more than one solution. So, the input layer has 10 neurons (there are 5 vertices and 2 CCD coordinates per vertex) and the output layer has 1 neuron.

Every NN is initialised with 50 neurons in each of the two hidden layers. In the former simulations more neurons were used, but the simulation showed that pruning process never stopped with more than 30 neurons (excluding those cases where the pruning process was blocked due to a bad threshold selection). The interconnecting weights among neurons are randomly initialised with triangular probabilistic density function between \(\pm \text{Max weight}\). Where \(\text{Max weight} = (\text{Number of neurons in layer} \div 2)^2\). In the beginning of each simulation, the slope of the sigmoidal function at the origin (0) is chosen between 0.2 and 10.0. Each neuron works in a continuous space of real numbers from -1.0 to +1.0.

3. Training Algorithm
Training algorithms for NN may be grouped in two main categories: Deterministic Algorithms or Random Search algorithms. The most representative element of the first category is backpropagation [2]. Backpropagation is a
standard algorithm used as a benchmark to compare other algorithms. However backpropagation, and most of its variations, have problems like divergence, oscillations near the solution, local minima blocking, pseudo-paralysis in early stages of the training and chaotic response [3], [4] and [5]. Random Search algorithms can not be pure randomness, they require some intelligence in order to search the solution where it can be found, so it is sensible to copy some random search strategies from Nature, that uses them in Evolution [6].

Basic-Evolutative algorithm is our contribution to the genetic algorithm's panoply compiled by Goldberg [7]. In our algorithm, we operate with real numbers instead of binary numbers (somebody could then classify Basic-Evolutative Algorithm in the border line between Genetic Algorithms and Evolution Strategies [8]). The "genome" of a given NN is sequentially coded in a string of real-valued words with variable length (see fix length difficulties [9] and the Genetic Programming Paradigm [10]): the first word has the number of hidden layers, the second word has the number of neurons per hidden layer, the third word has the first interconnecting weight, the fourth word has the next interconnecting weight, and so forth. The information of the structure (first and second word of the string) is redundantly coded at the two edges of the sequence.

The population has a fix number of individuals (50) for all the training process. Mutation consists of a random addition (Pm=0.02) of a real number to a random selected word of the genome. The added number is uniformly distributed between $\pm \text{Max Weight}$/3. Crossover is the standard genetic operator (Pc=0.25). The first offspring is always the best one of the former generation, the other are selected on a roulette wheel & fitness basis.

At the moment of doing crossover or mutations, the redundancy is used for avoiding the creation absurd offsprings; for example, an offspring having more weights than neurons or having a neuron without some interconnecting weights is not allowed to be born. Redundancy acts in the same way than the redundant bits of a binary transmission correct some errors (absurd offsprings) but are unable to correct other errors (mutations and crossover).

The pruning algorithm, based on [11] and [12], is as follows: If the pruning option is active, every time the training set is learnt, the neurons in the hidden layers with "almost constant output" ($\pm 5\%$) for all the samples of the training set or "similar output" ($\pm 3\%$) to another neuron in the same hidden layer, are randomly selected (one neuron per iteration) to be removed (pruned).

The reduced NN is re-trained until no pruning can be done, and then it may also be trained again with the pruning option inactive. These thresholds ($\pm 5\%$ and $\pm 3\%$) have been found experimentally. Slightly smaller values do not change the results substantially, but may block the training in a stage where the error is bigger and pruning is un-affordable. Threshold values above those mentioned may block the pruning process or cause unstable situations.

4. Preprocessing

Preprocessing of the samples in the training and test set consists of two steps: For coordinates $Xc$, $Yc$, $\phi$, $\theta$ and $\gamma$ it is normalisation and quadtratic transformation, for coordinate $Zc$ (the distance between target and chaser) it is inversion and normalisation. We normalise between $-0.9$ for the most negative input value in the training (or testing set), and $+0.9$ for the most positive input value. This preprocess reduces the probability of having the NN trapped in the saturation area of the sigmoidal function (in this area backpropagation is very slow [13]). A quadratic transformation ($y=ax^2+b$) is done in order to place most of the input signals where maximum accuracy is required: between the bounds of $-0.1$ and $+0.1$. Doing this, each NN performs better for all the training algorithms implemented, because of the high sensitivity of the first layer neurons when the inputs are in the $-0.1 +0.1$ margin. For $Zc$ coordinate the actual output is its inverse ($1/Zc$) in order to increase the accuracy at close distance (small $Zc$) instead of long distance (greater $Zc$), and thus, to reduce the relative error.

5. Simulation Process

We call "lesson" a set of five specific simulated displacements of the target in the vision field of the camera. A displacement is composed of 128 temporal samples, each sample consists of the 5x2 coordinates of the target in the CCD surface and the six attitude variables: $Xc$, $Yc$, $Zc$, $\phi$, $\theta$ and $\gamma$.

The displacement of the target along the CCD surface may be linear, triangular or sigmoidal from a pseudo-random start point to a pseudo-random finish point. The camera and associated lens are capable of focusing two different vertices of the target in two different pixels when the target is as far as 100 meters; on the other hand results with 1 meter resolution require more than 100 samples in a single lesson comprising a displacement from one end to another. The 640 samples in a lesson are randomly ordered at the beginning, in order to decouple the effect of the sequential translation of the target along the vision field.

No guess about the relationship between number of neurons (or number of individuals) and iteration time consumption has been done because the simulation software has a lot of graphical display subroutines and they take almost the same time independently of the number of neurons to be shown.

6. Results

The NNL is trained with all samples of the lesson $i$, and then, the errors at each NNL output are calculated with all the samples of another lesson $j$, where $i$, $j$ are different random elements in the library that we have built, it has 225 lessons. For each output, the m.s. error is obtained and then, a global m.s. error of all the outputs is calculated. The training process can be stopped under both of the following conditions: To stop after a predefined number of iterations or to stop when the error is less than a predefined magnitude.
First stop condition: the NNL was trained 250 times with different lessons \( l_j \), and the NNL errors were tested with always the same lesson \( l_j \) (where "i" is randomly selected (250 times) and "j" is randomly selected (once) at the beginning), the results, presented in Table 1, were obtained.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Neurons in hidden layers</th>
<th>Global output error m.s.</th>
<th>Worst NN output error m.s.</th>
<th>Best NN output error m.s.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bck. Prop.</td>
<td>50 + 50</td>
<td>1.58x10^{-2}</td>
<td>1.90x10^{-2}</td>
<td>1.40x10^{-2}</td>
</tr>
<tr>
<td>Basic Evol.</td>
<td>50 + 50</td>
<td>1.60x10^{-2}</td>
<td>1.98x10^{-2}</td>
<td>1.42x10^{-2}</td>
</tr>
<tr>
<td>B.P. + prun.</td>
<td>2 + 0</td>
<td>0.86x10^{-3}</td>
<td>1.59x10^{-3}</td>
<td>0.33x10^{-4}</td>
</tr>
<tr>
<td>B.E. + prun.</td>
<td>2 + 0</td>
<td>0.88x10^{-3}</td>
<td>1.63x10^{-3}</td>
<td>0.56x10^{-2}</td>
</tr>
</tbody>
</table>

**Table 1: NNL trained 250 times**

The first 2 rows of Table 1 were obtained without pruning the NNL. Under this condition backpropagation was the best algorithm in the 250 first lessons, at a first look, it could be interpreted like backpropagation "learns“ better. The last 2 rows of the table were obtained with the pruning option set to "yes". Under this condition backpropagation and evolutive algorithms reduced the structure of the best NN in the NNL to only one hidden layer with only two neurons, the worse output of the NNL has similar error, but Basic-Evolutive is unable to converge to a lower error whereas backpropagation is capable of improving its best output considerably.

Second stop condition: the NNL was trained with different lessons \( l_j \), the training was stopped when the m.s. error (tested with lesson \( l_j \)) is constant was less than \( 1.5x10^{-2} \); the results, presented in Table 2, were obtained.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Neurons in hidden layers</th>
<th>Iterations for 0.015 m.s err at best output</th>
<th>Iterations for 0.015 m.s err at worst output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bck. Prep.</td>
<td>50 + 50</td>
<td>281</td>
<td>369</td>
</tr>
<tr>
<td>Basic Evol.</td>
<td>50 + 50</td>
<td>301</td>
<td>535</td>
</tr>
<tr>
<td>Bc. Prop. + prun</td>
<td>2 + 0</td>
<td>189</td>
<td>291</td>
</tr>
<tr>
<td>B. Evol. + prun</td>
<td>2 + 0</td>
<td>167</td>
<td>315</td>
</tr>
</tbody>
</table>

**Table 2: NNL trained down to 1.5*10^{-2} m.s. error**

The last row in Table 2 shows that the best output may reach the \( 1.5x10^{-2} \) m.s. error faster than backpropagation but the worst output needs more iterations. This easily understood because Basic-Evolutive has a random basis that can improve one output, whereas it can make worse another output; it is statistically impossible for a random-based algorithm to improve all the outputs at every iteration, the algorithm "has" to fail in some input, at some iteration, it is not deterministic.

For both tables the slope of the sigmoidal function \( f(01) \) has been set to 0.2. This low-r value produces the most accurate results if no visual saturation effect appears, the higher value (10.0) tolerates the lost of one vertex for 10% of the samples in the testing set with only 6% of error degradation at the worst output.

Noise injection is simulated multiplying every sample of the training set by \( (1+\Delta) \), where \( \Delta \) is uniformly distributed between \( -\Delta_{\text{max}} \) and \( +\Delta_{\text{max}} \). If \( \Delta_{\text{max}} \) is less than \( 10^{-3} \) no performance degradation occurs. If \( \Delta_{\text{max}} \) is less than \( 10^{-3} \) the accuracy does not get worse but the structure requires about 15% more neurons. With values of \( \Delta_{\text{max}} \) greater than 0.01 the pruning algorithm is incapable of working.

Error of the six output v.s. iteration number is depicted in Figures 1 & 2. It can be seen that backpropagation produces six similar error curves while Basic-Evolutive produces more differences among the six outputs and it seems to stagnate for several iterations more frequently than backpropagation.

![Figure 1: NNL m.s. error for the 6 outputs](image1)

*Training: Backpropagation without pruning*

![Figure 2: NNL m.s. error for the six outputs](image2)

*Training: Basic-Evolutive with pruning*
7. Conclusions

C1 The NNL requires less number of neurons than six NN for equivalent performances.
C2 If pruning is required, Basic-Evolutitive algorithm is the fastest.
C3 If the objective is to reduce the m.s. error among the six outputs, Basic-Evolutitive algorithm is the fastest until certain margin where randomness stagnates, that margin is appreciated in Figure 2.
C4 If the objective is to stop training at 1.5x10^-2 m.s. error in the worst output, backpropagation is the best option at the end of the training of that output.
C5 For Xc and Yc coordinates, pruning algorithm stops with only two neurons in one single hidden layer, this seems reasonable because those coordinates are the "easy ones".
C6 In most simulations, Zc is the output with higher error and the one that requires more neurons. Perhaps no suitable pre-processing has been found, yet.
C7 Pre-processing is a crucial step. An adequate pre-processing produces significant results while bad or null pre-processing makes the NNL training unapproachable.

In summary, our best conclusion is that Basic-Evolutitive algorithm has to be used first because it is faster and is capable of pruning the NNL and after using it, if more accuracy is required, no random-based methods have to be used and backpropagation is recommended. These results are in accordance with [14] and [15].

In the future the work will be oriented in finding the robustness of the NNL in the case of individual neuron failure and pruning algorithm translation to a genetic approach like [16].

References


A Neural Approach to Lips Movements Modeling

Enrico CASELLA, Fabio LAVAGETTO and Roberto MIANI

DIST - Department of Communication, Computer and System Science
University of Genova, Via Opera Pia 13, 16145, Genova, Italy
Tel. (+39)-10-3532208, Fax. (+39)-10-3532948, email fabio@dist.dist.unige.it

Abstract. Progress in research and technology is opening new scenarios of a future technological society based on integrated multimedia services. An increasing share of consumers will explicitly demand for special aids because of age, handicaps, temporary or permanent diseases, and multimedia approaches will represent a means of formidable strength to meet their needs. In this paper a neural approach is described, oriented to the acoustic-to-visual transduction of speech into graphical animation. By means of a multilayer neural network, continuous speech is processed and converted into visual output suitable to lipreading. Preliminary results are reported describing the on-going experimentation of a "visual telephone" capable to convert speech into graphical lips animation for training deaf children to learn lipreading.

1 Introduction

Verbal face to face communication represents an evident example of bimodal human interaction where both the acoustic and the visual channel are used simultaneously to convey the message. The visible effects produced on the external articulatory organs during the speech production are in fact strictly correlated to the underlying acoustic information and can be translated into meaningful linguistic information by a skilled observer. The task of understanding the cues of speech associated to lips movements, commonly named "lipreading", can be performed successfully through automatic procedures implemented on a computer emulating the "human lipreader". The addressed fields of application are numerous, ranging from visual-auditory integrated systems for phonetic recognition to training systems for teaching lipreading to hearing impaired people. The interdependencies between video and audio information can be exploited further in videoconference/videotelephony for data compression or in advanced robotics for "antropomorphic" man-machine interfaces. The integrated analysis of the bimodal speech source, in fact, provides powerful means in rehabilitation for overcoming possible perception impairments affecting a specific perception channel through the alternative (or simultaneous) use of a more suitable one. Different approaches have been followed [1,3,4,5], based on statistical clustering, Hidden Markov Models and neural networks, oriented to the conversion of speech into "readable" visual outputs. An example is the selection, driven by speech parameters, of suitable mouth images which are then concatenated for rendering the "film effect" or, as an alternative, the use of synthetic models animated directly by the speech parameters.

In Section 2 we briefly outline the scheme of the implemented system, based on the neural architecture discussed in Section 3. The animation system employed for the visualization of the speech cues is described in Section 4, while the experimental results and final conclusions are reported in Sections 5 and 6, respectively.

2 Converting speech into lips movements

Speech conversion into lipreadable visual output can be performed through two different approaches, depending on the direct or indirect estimation of the articulatory parameters from the speech signal. In the former approaches speech is processed and the explicit phoneme recognition is performed followed by a strict association phoneme-visual information. In some proposed algorithms the elementary acoustic units are not necessarily phonemes but also diphones or even more complex structures, linked "as an ensemble" to the corresponding visual units called visemes. In case of misclassification, the non linearity of the phoneme recognition block usually introduces unacceptable distortion in the visualized output, i.e. inconsistent lip movements are synthesized. In contrast, the latter approaches by-pass the intermediate stage of the acoustic classification thus implementing a smooth and robust procedure for mapping the acoustic input domain directly onto the output space of the visual parameters. This association works either through a correspondence table obtained by means of clustering procedures, like vector quantization techniques, or through a trained neural network used as classifier. The most critical impairment present in both approaches, is the one to one acoustic to visual correspondence which does not take into account the temporal coarticulation dynamics. As a consequence of the articulatory phenomenon, the configuration of the whole vocal tract and of the mouth in particular does not depend only on one phoneme but rather on a longer phonetic unit like the sequence of two or even more phonemes. Beside this first dependency, usually referred to as "posticitation", the mouth configuration depends also on the future phonetic information. Since the mouth modifies its shape an-
ticipating the acoustic production, this second dependency is consequently called "anticipation".

Within rehabilitation applications oriented to teach lipreading to hard of hearing people, these two relevant dependencies must be necessarily taken into account and must be suitably modeled. The visual parameters exchanged of driving the synthesis of the labial movements must therefore depend on a non-causal observation of the speech signal. As a direct consequence of the different time constants characterizing the acoustic and the articulatory dynamics, the transition between two successive stable acoustic frames (phonemes) corresponds to a "significantly long" (up to some hundreds milliseconds) articulatory transition called "coarticulation". No reliable model has been yet proposed for describing the acoustic-articulatory dependencies during coarticulation and trivial interpolative solutions provide too poor quality, often insufficient for allowing lipreading. It is therefore necessary to implement a system capable to learn implicitly the articulatory rules by means of suitable training and perform the acoustic to visual conversion through the analysis of a suitably sized speech segment.

The Time-Delay neural network (TDNN), proposed ad used by Waibel [6,7,8] within applications of phonetic recognition, provides exactly the features which are required for performing the correct conversion. As explained in the next Section and confirmed by the achieved experimental results, this kind of neural architecture is in fact capable of transferring effectively the cues of speech to a limited number of parameters suitable for driving a "lipreadable" visualization.

3 Time-Delay Neural Network

Differently from conventional neural networks based on feedforward Multi-Layer Perceptron, each neuron of the TDNN performs the weighted sum including, beside the current acoustic frame, also the coefficients of a predefined number of previous frames stored by means of suitable delay lines. The time-integrated weighted sum is then used as input to the activation function, to generate the neuron response (see Figure 1):

$$ w_{i,j} = f(s_{i,j}) = f \left( \sum_{d=0}^{D(t)} \sum_{t=0}^{N(t)} w_{i,t,j,i,j} \right) $$

(1)

Using this kind of perceptron within a layered structure where each layer contains a variable number of parallel neurons each of them fed with all the outputs coming from the previous layer, it is possible to implement a neural mechanism capable to associate an output to a fixed number of input patterns.

In Figure 2 the time evolution of the input pattern coefficients is shown together with the output of a TDNN with one single hidden layer. The lines which link two successive layers indicate the the number of input vectors used by each neuron to generate a valid output. Let us notice how the information associated to the input pattern is integrated layer by layer: neurons in the hidden layer sum data from series of three successive strongly correlated input patterns and arrange a "preprocessed" information to supply to the subsequent layer. Thanks to this time-pyramidal structure, the network is induced to handle in a very natural fashion certain types of information like that associated to transients between stationary configurations or to time varying signals. During the learning phase performed according to the backpropagation algorithm, instead of comparing the TDNN output to the current target value, a past target reference can be employed thus implementing easily the non causal conversion filter necessary for coarticulation modeling.

3.1 Construction of the database

The neural network has been trained by means of pairs pattern-target related to the speech production process. The visual component of the database consists of images representing the central region of the speaker face, while he is pronouncing the speech corpus, including the nose, mouth and chin. The speech corpus, that is the acoustic component of the database, is composed of three sections: the first of them (DB1) contains only vocalic information including vowels with and without accent, the second section (DB2) is formed of all possible syllabic combination of the type /V_iCV_i/ while the third (DB3) consists of words containing all possible phonetic combination of 3rd order. Obviously, since a single speaker database is used, the system is
3.1.1 Analysis of the speech signal

The speech representation supplied to the TDNN network is composed of a set of coefficient vectors, computed through linear predictive analysis of the original signal samples. Speech processing operations concern:

1. preemphasis filtering;
2. signal segmentation into non-overlapping frames of 20 msec each;
3. Hamming windowing of each speech segment;
4. 10-th order linear predictive analysis and computation of the first 12 cepstrum coefficients.

Since the cepstrum coefficients exhibit a very wide dynamics, before feeding the network they must be normalized to guarantee homogeneity between training-set and testing-set data. Moreover, the normalization procedure must be applied locally to each frame of coefficients for providing real-time performances while preserving the information on the instantaneous power, necessary for detecting silence intervals. Because of these requirements, the following procedure has been followed:

- coefficients are normalized locally frame by frame within the interval [-1:1] in order to force null average;
- in correspondence of the same frame, the instantaneous power is normalized within the interval [0:1], based on the maximum expected power;
- normalized coefficients are finally multiplied by the normalized power.

3.2 TDNN optimization

Before applying the final training to the TDNNs, exhaustive simulations have been carried out and analyzed for finding out their optimal configuration. The network has been trained using different values for each configuration parameter (that is the number of neurons for each layer, the memory size of each neuron and the time offset between the network output and the target) and its performances have evaluated and compared. The network has been trained using database DB1 and then database DB2 so that it starts learning simple vocalic sounds and only after this first stage it passes to more complex syllabic sounds. By doing so, the occurrence of possible impairments in learning convergence is minimized. As stop condition, a constraint on the maximum number of allowed iterations is used. The learning performances corresponding to each different configuration are evaluated in terms of the Mean Square Error (MSE) and of the Maximum Error (ME). Once the optimal values for each parameter have been found out, the cross-validation method has been used for the final training of the TDNN in order to maximize the generalization performances.

4 Visual synthesis

The articulatory parameters which have been estimated through the TDNN are then used to drive the animation of a three-dimensional facial model.

The facial parametric model "fascia", implemented by A. Marriott and V. Hall from the Curtin University of Technology of Western Australia, has been employed. In order to obtain a readable output some suitable changes have been introduced: the lips geometry has been significantly improved and new articulatory organs like the palate and the tongue have been introduced (see Figure 4). The synthesis system has been implemented on a SGI Indigo 4000 XZ workstation, 100 MHz, whose limited power has prevented from real-time operations. A few seconds delay occurs before an input speech segment is converted into its corresponding visual representation. Speech is directly acquired from the workstation through a microphone, sampled at 8 kHz, linearly quantized with 16 bits and eventually processed, as previously described, to obtain the 12 cepstrum coefficients corresponding to non-overlapping 20 msec segments. Cepstrum coefficients are then normalized frame by frame and supplied to a bank of TDNNs, each of them trained to extract a specific articulatory parameter. It has been experimentally
found out that, to guarantee a lipreadable synthesis, at least five uncorrelated parameters must be used, i.e. the mouth width and height, the lips thickness, the jaw aperture and the tongue position. Before being applied directly to the model, the parameter estimates undergo suitable "smoothing" through a low-pass filter and are subsampled lowering the time resolution to 80 msec. The synthesis procedure uses, at each iteration, the current parameter estimates to modify the polygon structure (wire-frame) of the facial model by means of pseudo-muscle primitives. In the meanwhile, the corresponding speech samples are loaded into the audio buffer to assure acoustic-visual synchronization. Because of the heavy computational overhead required by the frame generation procedures, we are presently capable of displaying a maximum of 13 frames per seconds.

5 Experimental results

In Figure 5 we report the cost curves, for each "by epoch" iteration, obtained during the TDNN training for different values of pattern-target time offset $\tau$.

It is apparent that the estimation improves as long as the pattern-target time offset is increased, until a stable value of $\tau = 8$. This result confirms and provides a measure of the "anticipation" coarticulatory phenomenon. The optimal TDNN architecture, obtained after exhaustive simulation and comparison, contains two hidden layers composed of 8 and 3 neurons, respectively, while the width of the time windows "observed" by each layer are 3, 5, 7 starting from the first hidden layer.

In Figure 6 the actual value of the jaw aperture is compared with the estimate provided by the TDNN with reference to a portion of DB3 previously used to train the network. It is apparent that, even if some error occur in amplitude and phase, the main modes of the parameter trajectory are tracked successfully.

6 Conclusions

In this paper we describe a system capable to animate a wire-frame facial model by means of few articulatory parameters estimated by means of TDNN networks. The TDNN structure has provided very promising results as far its capability of converting speech parameters into articulatory parameters is concerned. Further satisfactory results concern the capability of the TDNN to model the coarticulation phenomenon thus allowing high quality synthesis, suitable to lipreading applications for rehabilitating hard of hearing people. Among the many application scenarios addressed by such conversion system we can cite:

- compression of facial image sequences in videotelephony;
- interpersonal telecommunication systems for disabled and elderly people;
- rehabilitation of deaf and hearing impaired people;
- multimodal man-machine interfaces.

References


A Neural Network Approach for Modeling the Vocal Tract

K.DEERGHA RAO † and G.SRIDHAR ‡

† Research and Training Unit for Navigational Electronics, Osmania University
Hyderabad - 500 007, India. Tel.No.662006 Fax.No.01 0842-860020
‡‡ Dept. of Electrical Engg., Indian Institute of Technology
Bombay - 400 076, India.

Abstract. A neural network approach for modeling the vocal tract has been proposed. The Extended Kalman Filtering (EKF) based learning algorithm which has much faster convergence speed than the conventional Backpropagation algorithm has been used for training the neural network. Simulation results that confirm the efficacy of the approach are given. Further, a comparative study of the conventional Backpropagation and the EKF based learning algorithms in terms of convergence behaviour has been reported.

1. Introduction

It is well known that the vocal tract (VT) can be reasonably approximated in terms of a rational transfer function represented by poles and zeros, the poles corresponding to vocal tract resonances and the zeros introduced due to the coupling between the vocal tract and nasal cavity. A variety of all pole and pole-zero modelling techniques [1-3] of speech have been considered based on linear prediction, homomorphic deconvolution and cepstral estimation. However, these methods involving large matrix operations and iterative optimizing techniques are generally undesirable in the environment of speech analysis-synthesis, in which memory and speed requirements are important and real time processing is required.

Shynk [4] described a frequency domain adaptive filter (FDAF) for the realization of pole-zero filtering using the parallel form realisation and Gauss-Newton algorithm and proved that it has general applications. This method can be used for modeling the vocal tract. But it is also computationally very intensive as it involves large matrix operations and complex arithmetic.

A novel scheme based on neural network approach that yields response that is identical to that of FDAF but with much reduced computational complexity has been proposed. The conventional Backpropagation algorithm is well known in the literature to train the neural network. However, the convergence speed is inherently slow because the learning rate is fixed. Further, it is necessary to tune the learning rate and the momentum term in a heuristic manner so that a quick convergence is obtained. An improper choice of these parameters may cause problems of stability or slow convergence [5].

A multilayered neural network (MNN) is a multi-input multi-output nonlinear system having a layered structure, and its learning algorithm can be regarded as parameter estimation problem. A number of estimation methods of nonlinear system have been reported in literature. Among them the Extended Kalman Filter (EKF) is well known one, which is derived by linearizing the system equations at each time step and applying the Kalman filter technique for the linear system [6]. Since the EKF based learning algorithm approximately gives the minimum variance estimate of the linkweights, it is expected that it converges in fewer iterations than the conventional Backpropagation algorithm.

A better algorithm for training multilayered neural networks based on the Extended Kalman Filter (EKF), which was shown to be much faster than the Backpropagation algorithm was presented and used for XOR and parity problems [7].

In this paper, the EKF based learning algorithm with modifications has been extended for training a three layered neural network used in modeling the vocal tract. A comparative study of the conventional Backpropagation and the EKF based learning algorithms in terms of convergence behaviour has been reported.

2. Multilayered Neural Network for Modeling the Vocal Tract

It is assumed that the VT behaves like a pole-zero system. The main motive behind the neural structure is to produce an output which is identical to that of the actual VT. If the response of the MNN matches with that of the VT, it may be interpreted that the MNN has the same transfer function as that of the VT.

The proposed Neural Network model of the VT is shown in fig.1. It is a three layered feed-forward feed back type network. There is a single node in the third layer which produces the vocal tract response. The past 'P' samples of the output signal along with M-1 past samples and the current sample of the input signal x(t) was used as inputs to the input layer.

3. The EKF Based Learning Algorithm

In multilayered neural network, each node which is connected by the links with all nodes in the adjacent layer computes a weighted sum of inputs, and then add an offset to the sum. The computed result is output through a nonlinear function. The 0th node in the nth layer is denoted by node(n,i).
The structure of the network considered here is illustrated in fig.1. In this network, the node in the input layer is assumed to do no operation, that is,

\[ \begin{align*}
    x_1^i(t) &= x(t - i + 1) & 1 \leq i \leq M \\
    x_1^i(t) &= x_1^i(t - i + M) & (M + 1) \leq i \leq P
\end{align*} \]  

where \( M + P = N_1 - 1 \)

Further, the offset is treated as the link weight by setting

\[ a^i_{n1} = \theta^i_{n1}, 1 \leq i \leq N_{n+1} - 1, \text{ when } n = 1, \quad i=1 \text{ when } n=2 \]

The operation of the node \((n+1,i)\) is then characterized by

\[ x_1^{n+1}(t) = f(\sum_{j=1}^{N_n} a^i_{nj} x_1^n(t) + \theta^i_{n1}) \quad \text{when } n = 1 \]

\[ x_1^{n+1}(t) = f(\sum_{j=1}^{N_n} a^i_{nj} x_1^n(t)) \quad \text{when } n = 2 \]

The function \( f(z) \) is chosen by the following sigmoid function:

\[ f(z) = \frac{1 - e^{-z}}{1 + e^{-z}} \]

The derivative of \( f(z) \) is given by

\[ f'(z) = (1 - f^2(z))/2 \]

where \( N_n \) is the total number of nodes in the \( n \)th layer \( x(t) \) is the input \( x_1^n(n) \) is the output of the node \((n,i)\) \( a^i_{nj} \) is the link weight from the node \((n,j)\) to the node \((n+1,i)\) \( \theta^i_{n1} \) is the offset of the node \((n,i)\)

It may be noted here that the sigmoid function was included only in the third layer (output layer)

Since the EKF is a method of estimating the state vector, the unknown link weights can be put as the state vector

\[ a = [(a_1^1)^T, (a_2^1)^T]^T, \quad (L \times 1) \]

where,

\[ a^1 = [a_1^1, a_2^1, \ldots, a_{N_n}^1]^T, \quad (N_n \times 1) \]

\[ a^n = [(a_1^n)^T, (a_2^n)^T, \ldots, (a_{N_{n+1}}^n)^T]^T, \quad (N_n(N_{n+1} - 1) \times 1) \]

where \( L \) indicates the total number of link weights as defined by:

\[ L = \sum_{n=1}^{2} N_n(N_{n+1} - 1) \]

Let the output vector of the nodes in the \( n \)th layer be

\[ x_1^n(t) = [x_1^n(t), x_2^n(t), \ldots, x_{N_n}^n(t)]^T, \quad (N_n \times 1) \]

and the desired output be \( d(t) \).

The multilayered neural network is then expressed by the following nonlinear system equations:

\[ \begin{align*}
    a(t + 1) &= a(t) \\
    d(t) &= x_1^n(t) + v(t)
\end{align*} \]

where \( x_1^n(t) \) is the output of the node in the output layer. The observation is represented by the desired output \( d(t) \) and \( v(t) \) is assumed to be a white noise with covariance \( E(t) \). Applying the Extended Kalman filtering to the eqs. 12 and 13 and following the procedure adopted in [7], leads to the following EKF based learning algorithm as shown in Table 1.

4. Simulation Results

The desired (vocal tract) response \( d(t) \) is assumed to be generated by:

\[ d(t) = G(q^{-1})x(t) \]

where \( x(t) \) is a random process with uniform distribution and a variance of 0.0801 and \( G(z) \) is the system to be identified and given by the following transfer function

\[ G(z) = \frac{b(1 - 0.9z^{-1})(1 + 0.81z^{-2})}{(1 - 0.71z^{-1} + 0.25z^{-2})(1 + 0.75z^{-1} + 0.56z^{-2})} \]

which has poles at \( p_{1,2} = 0.5 \pm 45^\circ, p_{3,4} = 0.75 \pm 120^\circ \)

and zeros at \( z_{1,2} = 0.9 \pm 90^\circ, z_3 = 0.9, z_4 = 0 \). The gain \( k \) was chosen such that the variance of \( d(t) \) is 0.0244.

The multilayered neural network (fig.1) with \( N_1 = 9, N_2 = 4, N_3 = 1 \) was used to model the pole-zero system (vocal tract). The weights associated with various layers of the neural network are initialized to random numbers in the range from -0.5 to +0.5. Both the backpropagation and the EKF based learning algorithms are used to update the weights. The actual and estimated VT responses are shown in fig. 2. It may be observed that they are in close proximity. The error convergence trajectories of both the algorithms are shown in fig. 3 and 4. From the convergence trajectories, it may be observed that the EKF based learning algorithm shows better convergence behaviour than the Backpropagation algorithm.

5. Conclusions

In this paper, a neural network approach for modeling the vocal tract has been proposed. The EKF based learning algorithm has been extended for training the neural network. From the results of simulation studies, it has been observed that the EKF based learning algorithm shows drastically improved convergence behaviour
than the conventional Backpropagation algorithm. Because of the improved convergence behaviour and less computation time, the new learning algorithm has good potential for neural network based real time processing applications such as speech processing.

References


Table 1: The EKF based learning algorithm

\[
\begin{align*}
\hat{z}_{n+1}(t) &= \left( \hat{a}_n(t) - \hat{\lambda}(t-1) \right)^T \hat{z}_n(t) \\
L_i &= 1 i \leq N_{n+1} - 1, \text{ when } n=1 \\
L_i &= 0 i \leq N_{n+1} - 1, \text{ when } n=2 \\
\hat{\lambda}(t) &= \hat{\lambda}(t-1) + \mu(t-1)(d(t) - \hat{z}_n(t))^2 - \hat{\lambda}(t-1) \\
\hat{y}_n(t) &= \hat{z}_n(t) \\
\text{for } n = 2 & \text{ to } 1 \text{ step } -1 \\
\hat{\Lambda}_n &= \begin{cases} 
1 - (\hat{z}_n(t))^2/2, & \text{for } n = 2 \\
(1 - (\hat{z}_n(t))^2/2) \sum_{i=1}^{N_{n+1}-1} \hat{a}_i(t) \hat{\Lambda}_n(t), & n \neq 2 
\end{cases} \\
\hat{a}_n(t) &= \hat{a}_n(t-1) + \Delta \hat{\lambda}(t) \hat{z}_n(t) \\
\hat{\psi}_n(t) &= \hat{\psi}_n(t-1) + \Delta \hat{\psi}_n(t) \\
\beta_n(t) &= \Delta \hat{\psi}_n(t)^T \hat{z}_n(t) \\
\hat{p}_n(t) &= \hat{p}_n(t-1) - \hat{\lambda}(t)^2 \hat{z}_n(t) \hat{z}_n(t)^T \\
\hat{y}_{n+1}(t) &= \hat{y}_n(t-1) + \Delta \hat{y}_n(t) \hat{z}_n(t) - \hat{y}_n(t-1) \\
\text{with } \hat{y}_1 &= \hat{y}_{N_{n+1}}
\end{align*}
\]

Fig.1: Structure of the Neural Network Model of Vocal Tract
Fig. 2 Comparison of actual & estimated VT

- actual VT
- estimated VT from NN

Fig. 3 Error with EKF

Fig. 4 Error with BP
Using the Gamma Net in Speaker Recognition

António TEIXEIRA and Francisco VAZ

Dep. Eletrónica e Telecomunicações - INESC, Universidade de Aveiro, 3800 AVEIRO, PORTUGAL, Tel.: +351-34-370500, Fax: +351-34-370545, E-mail: ajst@zappa.inesc.pt

Abstract. This paper presents an experience in using a dynamic neural network, the focused gamma net, in the area of speaker recognition, in the two variants of verification and identification. The approach used is based on the classical model for speech recognition. First we extract relevant features with a filter bank. Features are compared with previously constructed references using a neural network. Decision is done using neural network outputs. The dynamic neural network, the gamma net, was used because the patterns are time varying.

We tried the usage of a network per speaker and several speakers per network. The first approach, being modular, gave better results making possible the construction of large systems.

The results obtained are promising and show that the method is robust.

1. Introduction

Speaker recognition [1,2] is very important in speech processing. It can be applied to security to improve the performance of speech recognition systems and forensic applications.

The term speaker recognition is used to refer to two related functions: speaker verification (SV) and speaker identification (SI). In SV, an identity is claimed by the user, and a binary decision is made whether to accept or reject this claim. In SI, the machine is required to make a positive identification from a population of speakers. In this work we are interested in both SI and SV.

Speaker recognition methods can be divided into two types: text-dependent and text-independent. In the first case the speech used is fixed, and in the second there is no restriction to the speech. Our work used a fixed phrase.

Speakers recognized by the system are called Clients, the others Impostors. Impostors that disseminate their voices to imitate one client are called Mimics.

A few experiments have been reported on the literature on the use of connectionist model for speaker recognition. The earliest references seem to be [3] using Multilayer Perceptrons with one neural network per speaker, and [4,5] where speakers were recognized by a codebook in LVQ. Both approaches are text-dependent.

More recent work in text-independent systems used Time Delay Neural Networks (TDNN) [6] and Predictive Neural Networks [7]. The last one uses both static and dynamic features, that are required for higher performance.

In our work we used a different way of using the dynamic of the speech spectral information called focused gamma net [8].

2. System

![System Model](image)

Figure 1. System Model

The developed system has a classical architecture in the area. It uses blocks for feature extraction, comparison, references, and decision. Due to the nature of the neural network the comparison and references blocks are both implemented by the neural network. The adaptable network parameters constitute the references and are adapted in the training phase. In the test phase the network acts as a comparator.

2.1. Feature Extraction

![Feature Extraction](image)

Figure 2. Feature Extraction
Speech features are extracted using a filter bank based on a model that aims to incorporate knowledge about the human ear.

The filter bank covers a bandwidth from 200 to 3300 Hz and is composed by 18 gammatone [9] bandpass filters. Center frequencies and bandwidths were calculated using human ear information, according to the work of Moore and Glasberg [10]. Pass band output is passed by a non-linearity (squared module) and low pass filtered to obtain the energy of the signal in each band.

The output is sampled at 30 Hz and normalized between 0 and 1.

2.2. Comparison and References

Due to the time varying nature of the extracted patterns the neural network used, the focused gamma net [8], has a built-in dynamic input layer. This network is a simplification of the general convolution model, where the dynamic is introduced using delays. The gamma model is also a simplification of the general auto regressive moving average (ARMA) model. The simplifications gave us a neural additive model with trivial stability. Being additive we could use the back-propagation-thru-time (BPTT) or real-time-recurrent-learning (RTRL), but the first presents the unfolding problem and the second can only be applied to small networks. The previoud focused architecture [11], with only memory in the input layer, makes possible to use a simple learning rule derived using the back propagation technique [12].

![Gamma Memory](image)

**Figure 3.** The focused gamma net

2.2.1. The Focused Gamma Net

The Focused Gamma Net is represented in Figure 3. Assume a signal $I(t)$ with $M$ dimensions. The past of this signal is stored in a cascade of first order infinite response (IIR) filters, described by the relations,

$$x_i(t) = I_i(t)$$

$$x_{ik}(t) = (1 - \mu_i) x_{ik}(t-1) + \mu_i x_{ik}(t-1), \quad (1)$$

with $t = 0, 1, 2, \ldots, t_i$; $i = 1, \ldots, M$ and $k = 1, \ldots, K$. This MEMORY structure is represented in Figure 4. Memory properties, depth and resolution, are controled adapting the $\mu_i$ parameter. TDNNs are a special case of this structure with parameter $\mu_i$ fixed and equal to 1.

The activation of the entry level are mapped into a set of units in the output through a static feed-forward net. The network units are indexed from $M+1$ to $N$.

![Static Feed-forward Net](image)

**Figure 4.** Gamma Memory

Activations, in the static net, are calculated by,

$$x_i(t) = \sigma(\sum_{j=1}^{M} \sum_{k=1}^{K} w_{jk} x_{jk}(t)),$$  

(2)

where, for simplification, we used the notation $x_{jk}(t) \equiv x_{jk}(t) e_{k} w_{jk} \equiv w_{jk}$.

2.2.2. Learning

Assuming a quadratic performance index, and using the back propagation technique [13], we reach the following usual equation for the static net,

$$\frac{\partial E}{\partial w_{jk}} = \delta(t)x_{jk}(t)$$

(3)

The parameter $\mu_i$, which modulates the memory depth ($K/\mu_i$), can also be adapted using this ideas. Using the chain-rule

$$\frac{\partial E}{\partial \mu_i} = \frac{\partial E}{\partial x_i} \frac{\partial x_i}{\partial \mu_i},$$

we get

$$\Delta \mu_i(n) = \eta \sum_{k=0}^{K} \frac{\partial x_i(n)}{\partial \mu_i} p_k(n),$$

(4)

where $p_k(n)$ is the back propagated error to the input. The other derivative can be obtained extending the back propagation of the error to the input units. All contributions are added, because we have only one $\mu_i$ per memory structure.

Applying the results of Oliveira e Silva [12], we reach the very simple formula

$$\frac{\partial x_i(n)}{\partial \mu_i(n)} = \frac{k}{\mu_i}[x_i(n) - x_i + (n)].$$

(5)

The implementation was made using the Stuttgart Neural Networks Simulator (SNNS) [14]; which was modified to
support the gamma net, extending the application of the package to new areas [15].

2.3. Decision

![Decision Diagram]

**Figure 5. Decision Process**

Neural Network output values are used by the decision block to reject or accept the verification demand, or to determine the identity of the speaker, in the identification case. The process, for identification, consists in determining the output unit with the maximum value and using a threshold. In verification only the output unit representing the claimed identity is used. It is only applied a threshold.

3. Data Base

The phrase ("Universidade de Aveiro") was chosen aiming at to represent the phonetic variety of the Portuguese language [16], beginning and ending with vowels to facilitate segmentation [17].

Data was collected in a Laboratory with moderate noise, and the phrase was roughly segmented using a signal editor. We collect data from 15 males (5 clients and 10 impostors) in several sessions of 10 pronunciations. Impostors have recorded 2 sessions, two of the clients 10 sessions, and the 3 others 7 sessions. The average duration of the phrases is approximately 1 second.

4. Results

4.1. System Parameters Definition

We performed several preliminary tests to define the network topology. We used only one hidden layer with a variable number of units and the best results were obtained with 14 and 16 units. For the output layer we used the number of output units equal to the number of clients making a direct correspondence between units and clients, without codification. The activation of one unit indicates a client, and all units not active represent impostors.

The tests showed also that it is necessary to stop the training with a reasonably high error to avoid overtraining and improve generalization. We choose to stop training for an equal false positive and false negative rate and used the resulting error value on subsequent tests.

It is important that the network generalizes well the client(s), but also that generalizes well what an impostor is. Tests showed that the net needed more training examples of impostors than of clients.

4.2. Verification

![Verification Diagram]

**Figure 6. Fully Connected Network with 1 output unit, for Verification**

For verification the utilization of a network for each one of the clients was tried. The network had only one output unit. Due to the training the network is specialized in recognizing one particular person. To add another client to the system there is no need for retraining the existent networks, i.e., the system is modular. This type of approach gives also smaller networks that can be trained faster and generalize better.

The best result obtained was 8% of false rejection of the client and 6.15% of wrong acceptance of impostors. The worst result (using the voice of the first author) gave 26% of false rejection and 22.31% of false acceptance.

Some of the results are presented in Table 1 (in %).

<table>
<thead>
<tr>
<th>Speaker</th>
<th>False Rejection</th>
<th>False Acceptance</th>
<th>(FRxFA) (1/2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIG</td>
<td>8%</td>
<td>6.15%</td>
<td>7%</td>
</tr>
<tr>
<td>AJST</td>
<td>26%</td>
<td>22.31%</td>
<td>24%</td>
</tr>
<tr>
<td>JLVR</td>
<td>0</td>
<td>10%</td>
<td>----</td>
</tr>
</tbody>
</table>

**Table 1. Some Verification Results**
4.3. Identification

We have also done a Identification test, using one network with five output units, corresponding to the same number of clients.
The average of correct decisions for the 5 speakers was 72%. The worst speaker, the same that caused problems in the verification, only achieved 40% and the other 4 achieved 60, 85, 90 and 95%. Impostors were rejected correctly in 62% of the cases. The impostors with training patterns were rejected correctly in 82% of the tests, but the other impostors, with no training, have been rejected correctly in only 52% of the cases.

5. Conclusions

Despite the reduced number of tests performed and the small size of the database, recorded in a noisy environment, and segmented very primitively, the results can be considered promising and point to a certain robustness of the method. The worst results were obtained with a very difficult speaker.
The major practical difficulties are the training time, the best topology being different from speaker to speaker, and the lack of an adaptation to voice variation over time. There is also no mechanism for network adaptation to the changes in the speaker voice over time.
Our results also point to the advantage of using a network for each speaker, reducing the needed training data and time to train.
The identification results are not so conclusive due to the small database.

6. Future Work

Currently we are interested in not completely connected topologies. We will try the specialization of hidden neural cells, motivated by similar neural spatio-frequency mapping in biological auditory systems, as activations originating at different displacements along the basilar membrane are transmitted via dense neural pathways to different regions of the temporal lobe [18]. We are also interested in investigating alternative dispersive kernels for employment in memory layers, for example the Laguerre Filters [19].

7. Acknowledgements

The first author would like to thank JNICT for the MsC scholarship. We also thank the SNNS group at University of Stuttgart for making the simulator available to others, and the Computational Neuroengineering Laboratory in University of Florida for the support on the gamma model.

References

Speaker identification using networks of radial basis functions

Steven E. Fredrickson, Stephen Roberts, Neil W. Townsend, & Lionel Tarassenko

Neural Networks Research Group, Department of Engineering Science
University of Oxford, Oxford OX1 3PJ, UK.
Tel/Fax: +44 (0)865 273126/273908, E-mail: sef@robots.oxford.ac.uk

Abstract: The hidden layer of a radial basis function (RBF) neural network can be exploited to validate RBF network outputs by checking for test data that is 'novel' with respect to the training data. In this paper we evaluate three methods of novelty assessment (two hidden-layer methods and a Parzen windows approach) on a speaker identification task. A population of 50 speakers is partitioned into subsets that are assigned individual RBF networks, and novelty assessment is used to select the correct subnetwork for a given test utterance. Compared to using a large single network, subnetworks are shown to achieve an equal or better error rate with significantly less training time.

1 Introduction

Recognising the voice of a familiar speaker is relatively easy for human listeners, but speaker identification by machine remains a challenging speech processing problem. One promising approach to speaker identification involves the use of artificial neural networks. Previously we compared radial basis function (RBF) neural networks [3] [7] to multilayer perceptron (MLP) networks [7] and found RBF networks to provide superior performance on a text-dependent isolated word speaker identification task [5]. We now consider segmenting large speaker sets into subsets to improve classification accuracy and reduce training time.

2 Radial basis functions

An RBF neural network has two layers. The first layer maps an \( N_I \) dimensional input vector to \( N_C \) hidden-layer nodes or 'centres,' \( (R^{N_I} \rightarrow R^{NC}) \), and the second layer maps the activations of these centres to \( N_K \) output nodes, \( (R^{NC} \rightarrow R^{NK}) \). For speaker identification, we use an input feature vector consisting of cepstral coefficients [6]. On the output we employ one-out-of-\( N_K \) encoding in which each speaker to be recognised is assigned an output class.

The activation of an RBF centre is usually a Gaussian function of its proximity to the current input pattern:

\[
\phi_j(\overline{z}) = \exp \left\{ -\frac{1}{2} (\overline{z} - \overline{c}_j)^T \Sigma_j^{-1} (\overline{z} - \overline{c}_j) \right\}
\]

where \( \phi_j(\overline{z}) \) is the response of the \( j \)th RBF centre, \( \overline{z} \) is the \( N_I \) dimensional input pattern, \( \overline{c}_j \) is the \( j \)th \( N_I \) dimensional centre, and \( \Sigma_j \) is its covariance matrix. RBF centre positions and covariance matrices can be determined by unsupervised learning techniques, such as K-means clustering followed by a width heuristic [9], or the Expectation-Maximisation (EM) algorithm [1].

The second layer of the network is a linear mapping from the RBF activations to the output nodes:

\[
y_k = w_{k0} + \sum_{j=1}^{NC} w_{kj} \phi_j(\overline{z})
\]

where \( y_k \) is the \( k \)th output node, \( w_{kj} \) is the weight from the \( j \)th RBF centre to the \( k \)th output, and \( w_{k0} \) is the bias of the \( k \)th output. Output weights can be computed via supervised learning techniques, such as least mean-square (LMS) gradient descent or matrix pseudo-inversion.

Because we employ one-out-of-\( N_K \) encoding and a squared error cost function, network outputs are estimates of bayesian a posteriori class probabilities [10]. Therefore, network outputs can be combined probabilistically. For example, \( R \) speech vectors from the same speaker can be used collectively to make a speaker classification decision. Assuming the vectors form an independent sample set, the combined output probabilities can be computed by multiplying individual outputs:

\[
[P_k]_{comb} = \prod_{m=1}^{R} [y_k]_m
\]

where \([P_k]_{comb}\) is the \( k \)th combined output probability, and \([y_k]_m\) is the \( k \)th output from the \( m \)th observation.\(^1\)

\(^1\)Network outputs near zero should be treated as having a small positive value to ensure nonnegative probabilities.
The index \( k \in \{1, 2, \ldots, N_X \} \) of the maximum combined probability, \( \max \{P_k|\text{comb}\} \), is the index of the speaker.

3 Connex speech data

Simulations described in this paper were conducted using the British Telecom Connex alphabet database. The Connex database consists of 3 utterances (2 for training and 1 for testing) of each of the 26 letters of the alphabet from 104 British speakers. We confine our testing to subsets of the 50 speakers with complete data. Utterances were recorded in a silence cabinet at a sampling rate of 20 kHz using a 16 bit A/D converter. After endpointing, speech samples were segmented into 512 sample (25.6 msec), overlapping, pre-emphasised, Hamming-windowed frames, and the FFT was used to derive 8 mel-frequency cepstral coefficients (MFCCs) for each frame. We apply linear time normalisation to produce 8 frames of 8 MFCCs for all utterances and then normalise the 64 components of the input feature vectors to have zero mean and unit variance.

4 Speaker set segmentation

4.1 Motivation

In contrast to speaker verification, speaker identification performance degrades significantly as the size of the speaker set increases. Figure 1 shows speaker identification test error rate versus speaker set size obtained from simulations on three utterances (two for training, one for testing) of each of the E-set letters B, D, E, P and V from Connex database speakers. The number of RBF centres \( N_C \) for each simulation was \( N_C = 0.3 \times N_P \), where \( N_P \) is the number of training patterns.

A common approach to handling large numbers of speakers is to “divide and conquer” [2] [4]. In our work, we assign small subgroups of speakers to individual RBF networks. Compared to large single networks, ‘subnetworks’ offer the potential for better classification performance and reduced training time, but the key is to determine the correct subnetwork reliably.

4.2 Subnetwork selection using novelty

Partitioning the speaker set into subsets involves a tradeoff between facilitating subgroup selection and facilitating speaker discrimination within a subgroup. It may be possible to optimise this tradeoff with clustering and/or nearest neighbour techniques, but for the purposes of this paper, we partition the speaker set into arbitrary subsets of equal size using speaker class labels.

After training \( N \) RBF subnetworks (one subnetwork for each subgroup of speakers), we select the subnetwork associated with a given input based on novelty. Any valid test speaker should be deemed novel by \( N-1 \) of the subnetworks and identified by the \( N \)th subnetwork. If there is a risk of impostors, an explicit threshold is needed to exclude impostors from all subnetworks, but for closed-set speaker identification, it is sufficient to assign the current input pattern to the subnetwork with the lowest novelty value.

4.3 Novelty assessment techniques

We investigated three novelty assessment techniques for selecting the trained subnetwork associated with the current input: Minimum Mahalanobis distance, Projective geometry, and Farzen windows. Each of these techniques is applied to the subset of the training dataset assigned to each subnetwork.

4.3.1 Minimum Mahalanobis distance method

An appealing feature of RBF networks trained using a maximum-likelihood approach is that the hidden layer forms a Gaussian mixture model of the training data probability density function (PDF). The likelihood or evidence associated with a test input is given by

\[
p(\mathbf{z}) = \sum_{j=1}^{N_C} P(j) p(\mathbf{z}|j) \tag{4}
\]

where \( P(j) \) is the prior probability of the \( j \)th centre, and \( p(\mathbf{z}|j) \) is the likelihood of the data \( \mathbf{z} \) conditioned on the \( j \)th centre. The prior probabilities are typically assumed to be equal, \( P(j) = \frac{1}{N_C} \), and the conditional likelihood

\[3\text{K-means clustering can provide an adequate PDF approximation at much less computational cost.}\]
terms are computed using normalised RBF responses:

\[ p(\tilde{z}|j) = \frac{1}{\sqrt{(2\pi)^{N_f} |\tilde{\Phi}_j(\tilde{z})|}} \]

(5)

where \( \tilde{\Phi}_j(\tilde{z}) \) is the unnormalised response from equation 1.

Since novelty is related to data density, likelihood values can be used to assess whether test inputs are novel with respect to the training database \([11]\). A precise PDF estimate is not necessarily the best novelty detector, however. We have found that it is generally better to use unnormalised RBF responses for novelty assessment. Moreover, for simulations presented in this paper, we consider only the maximum kernel response, \( \max\{\tilde{\Phi}_j(\tilde{z})\} \), which is equivalent to identifying the centre with the minimum Mahalanobis distance (MMD) to the input vector \([11]\). Thus, after applying a test pattern to all subnetworks, we choose the subnetwork with the smallest MMD (lowest novelty) value.

4.3.2 Projection method

Another novelty estimate that can be derived from the RBF hidden layer is based on projective distance \([8]\). The novelty component \( \tilde{z}_n \) of an input \( \tilde{z} \) is given by

\[ \tilde{z}_n = (I - CC^+ )\tilde{z} \]

(6)

where \( I \) is the \( N_f \) by \( N_f \) identity matrix, \( C \) is the \( N_f \) by \( N_C \) matrix of Gaussian kernels, and \( C^+ \) is its pseudo-inverse. Kohonen describes \( CC^+ \) as a novelty filter. After computing \( \tilde{z}_n \) for all subnetworks, we choose the subnetwork with the lowest novelty vector norm \( \| \tilde{z}_n \| \).

4.3.3 Parzen window method

Parzen estimation is a nonparametric method of PDF estimation. Since it is separate from the network, it serves as a benchmark for comparison with the RBF methods. A Gaussian Parzen window \( N(\tilde{z}_p, \sigma) \) is assigned to every training pattern \( \tilde{z}_p \), and the global width \( \sigma \) is determined heuristically or by cross validation. Novelty assessment and subnetwork selection for test patterns is based on likelihood values. Likelihood computation for a test pattern \( \tilde{z} \) follows the description in subsection 4.3.1, except that training patterns are used instead of RBF centres:

\[ p(\tilde{z}) = \frac{1}{N_P(2\pi)^{N_f} \sigma^{N_f}} \sum_{p=1}^{N_P} \exp \left\{ -\frac{|\tilde{z} - \tilde{z}_p|^2}{2\sigma^2} \right\} \]

(7)

where \( N_P \) is the number of training patterns, \( \tilde{z}_p \) is the \( p \)th training pattern, \( \tilde{z} \) is the test pattern, and \( N_f \) is the dimensionality of the input.

<table>
<thead>
<tr>
<th>Speakers per net</th>
<th>Total nets</th>
<th>Subnet select method</th>
<th>Subnet select error</th>
<th>Overall test error</th>
<th>Train time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>27.0%</td>
<td>18.0</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>MMD</td>
<td>19.6%</td>
<td>25.8%</td>
<td>1.7</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>Proj.</td>
<td>40.2%</td>
<td>44.8%</td>
<td>1.7</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>Parzen</td>
<td>16.8%</td>
<td>23.7%</td>
<td>1.7</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>MMD</td>
<td>28.2%</td>
<td>29.8%</td>
<td>2.0</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>Proj.</td>
<td>37.4%</td>
<td>38.9%</td>
<td>2.0</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>Parzen</td>
<td>18.8%</td>
<td>22.2%</td>
<td>2.0</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>MMD</td>
<td>50.7%</td>
<td>50.7%</td>
<td>6.3</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>Proj.</td>
<td>27.1%</td>
<td>27.1%</td>
<td>6.3</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>Parzen</td>
<td>16.0%</td>
<td>16.0%</td>
<td>6.3</td>
</tr>
</tbody>
</table>

**TABLE 1: RBF results using subnets**

5 Results

Table 1 presents results of RBF simulations on Connex data from 50 speakers of the E-set letters B, D, E, P, and V. We use two utterances of each letter for training and one for testing. To enable comparisons between different partition sizes (speakers per subnetwork), the number of RBF centres per subnetwork was fixed at \( N_C = 0.3 \times N_P \), where \( N_P \) is the number of training patterns assigned to each subnetwork. Centre positions were optimized using \( K \)-means clustering and widths were set using a Global First Nearest Neighbour (GFPNN) heuristic. Results for each line of table 1 are the average of 10 simulations, and reported training time includes training all subnetworks.

6 Conclusions

Table 1 shows clearly that subnetworks can improve classification performance. On all three sizes of subnetworks tested, the overall error rate (subnetwork selection and speaker classification) using Parzen subnetwork selection was lower than for the single network. The best results were achieved using single speaker networks which, in fact, do not require RBF classification layers. Overall error rates obtained using the hidden-layer methods were higher, but the minimum Mahalanobis distance method applied to small numbers of larger subnetworks performed as well as the single network. (The poor MMD results on single-speaker networks is understandable given that there are only 3 centres per subnetwork.) Projective geometry, by contrast, worked best on the 50 single-speaker networks, achieving an error rate equivalent to the single network.

The difference in training time between the single network and the subnetworks is significant. For example, training 10 networks of 5 speakers on a Sparc 2 requires \( 10 \times 0.20 = 2.0 \) minutes of CPU time compared with 18.0 minutes for one network of 50 speakers. Subnetworks are slower than single networks for processing test data, but identification of a test speaker is performed
so quickly (within a fraction of a second), that differences
in testing time are not of practical concern.

7 Discussion

Since the Parzen window method uses all training patterns
to form a PDF estimate, it is not surprising that it assesses
novelty better than the MMD and projection methods. At
the cost of large storage requirements and full likelihood
computation, Parzen windows offer a very localized PDF
which works well for novelty ‘gatekeeping.’ It would
however, be nice to eliminate the overhead of Parzen
windows.

We are most concerned about the overall error rate.
A 16% error rate is clearly unacceptable for practical
applications. To reduce the error rate to below, say 1%, it
will be necessary to combine multiple utterances using
posterior probability estimates as outlined at the end of
section 2. (For simulation results on single networks, see
also [5].) Unfortunately, the single-speaker networks do
not utilise a classification layer, so no posterior probabili-
ties are available. The 5 and 10 speaker subnetworks pose
a similar problem. Intra-subnetwork posterior probabili-
ties are available, but it is not clear how to use multiple
utterances optimally to reduce errors in subnetwork selec-
tion.

As noted in subsection 4.2, it may be possible to
partition speakers intelligently a priori. On a per utter-
ance basis, we would not expect large performance gains
from grouping speakers: By grouping similar speakers,
it becomes easier to select the correct subnetwork, but
intra-subnetwork classification becomes more difficult.
Conversely, by grouping unlike speakers, classification
performance improves at the expense of subnetwork selec-
tion. However, the use of multiple utterances may
provide a partial solution to this tradeoff. As explained
above, multiple utterances can be combined probabilis-
tically for classification, but not for novelty assessment.
Therefore, it might make sense to favour novelty assess-
ment (by grouping like speakers) and rely on multiple
utterances for classification.

Acknowledgements

The authors wish to thank British Telecom for making
the Connex speech database available.

REFERENCES

1. N. M. Laird A. P. Dempster and D. B. Rubin. Max-
imum likelihood from incomplete data via the EM
algorithm. Journal of the Royal Statistical Society,

2. Y. Bennani. Speaker identification through a modular
connectionist architecture: Evaluation on the TIMIT

interpolation and adaptive networks. Complex Sys-

Speaker recognition using neural networks and con-
ventional classifiers. IEEE Transactions on Acous-
tics, Speech, and Signal Processing, 2(1):194–205,

5. S. E. Fredrickson and L. Tarassenko. Radial basis
functions for speaker identification. In Proceedings
of the ESRA Workshop on Automatic Speaker
Recognition, Identification, and Verification, pages

6. S. Furui. Cepstral analysis technique for auto-
matic speaker verification. IEEE Transactions on
Acoustics, Speech, and Signal Processing, ASSP-

7. D. R. Hush and B. G. Horne. Progress in super-
vised neural networks: What’s new since Lippmann?
IEEE Signal Processing Magazine, 10(1):8–39, Au-
gust 1993.

8. T. Kohonen. Self-Organization and Associative
Sciences. Springer-Verlag, Berlin-Heidelberg-New

of locally-tuned processing units. Neural Compu-

classifiers estimate bayesian a posteriori probabili-

11. S. J. Roberts and L. Tarassenko. A probabilistic
resource allocating network for novelty detection.
Dimensionality and Dynamic Temporal Structure of a Time-Delay Neural Network for Speech Processing

Victoria RODELLAR¹, Pedro GOMEZ¹, Mercedes PEREZ¹, Victor NIETO¹, Rosario ROMERA², and José M. MURUZABAL²

¹Dept. de Arquitectura y Tecnología de Sistemas Informáticos, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, Urbanización Montepríncipe, Boadilla del Monte - 28660 Madrid - SPAIN
Tel/Fax: +34.1.336.73.84/336.74.12 E-mail: pedro@pino.datst.fi.upm.es

²Departamento de Estadística y Econometría, Universidad Carlos III, C/Madrid 126, Getafe - 28903 Madrid - SPAIN,
Tel/Fax: +34.1.624.98.64/624.98.49 E-mail: mruonera@eco.uc3m.es

Abstract. The present work studies the optimal dimensioning of Time-Delay Neural Networks (TDNN’s) for Speech Processing according with the order of non-stationarity of the Speech Traces. The relationship between TDNN’s and Adaptive Filters is explored and it may be shown that under certain conditions the structure of a TDNN’s perceptron may be modeled by a Multiple Regression Lattice Joint Estimator. From this fact, optimal estimations of the autoregressive order of Speech Spectral Vectors may be inferred. It may be shown that for practical cases of study (glides in Spanish), this order may be kept equal to one, thus allowing to reduce the delay levels in an equivalent TDNN to two, although preserving their capability to trace the non-stationary nature of Speech quite efficiently.

1. Introduction
Since the outburst of Neural Networks, many efforts have been devoted to their application in Speech Processing and Recognition problems [1]. One of the most useful structures developed for such purposes is the Time-Delay Neural Network (TDNN) [2]. This network has been successfully applied to Time Alignment and Pattern Matching problems, where good results have been reported. Nevertheless, TDNN’s have been criticized because of their relatively high computational cost. Based on these criticisms, hybrid HMM-NN models have been proposed [3]. Nevertheless, TDNN’s present a hybrid nature between Adaptive Transversal Filters, and Neural Networks, which renders them especially suited to characterize the intrinsic non-stationary nature of Speech. This aspect is especially relevant, since a powerful theory has been developed during the last three decades for Adaptive Filters [4], which could be used to dive in the capability of TDNN’s to represent non-stationary processes, depending on the number of delay stages used. It does not seem that this problem has deserved much attention up to now, as is generally assumed [1] that the bigger the number of delay stages used, the better the tracing of the dynamic nature of Speech. This rule, being basically true, prevents the optimization of computational complexity, as it creates oversized structures. This research is intended to open such a study, considering that what is really meaningful is the order of autoregression of the input data used to train the network.

2. TDNN’s and Digital Filters
To explore in more detail this relationship, a simplification of a generic TDNN will be carried out, and the relation between this structure and that of the equivalent Transversal Adaptive Joint Estimator, will be established. We will consider the case of a TDNN with three layers (input, hidden, output), as the one presented in Fig. 1. The activation vector to the input layer \(x_{1n}\) is composed of \(k+1\) delayed versions of \(p\)-dim. PARCOR vectors, each one being obtained from the output of a Gradient Adaptive Lattice [4], and a thresholding element preset to 1, thus configuring a vector of dimension \(p(k+1)+1\), with \(p=16\). This is projected on the \(q(qm+1)+1\) weight matrix \(W_{1n}\), producing a \(q\)-dimensional vector \(s_{2n}\). This is passed through a bipolar sigmoid function \(F\), and used to compose the activation vector of the hidden layer \(x_{2n}\), with dimension \(q(m+1)+1\).

![Figure 1. General structure of the TDNN under study.](image)

Similarly, this is projected on the \(r(qm+1)+1\) weight matrix \(W_{2n}\), to yield \(s_{3n}\), and renormalized by the nonlinear function \(F\), to produce the output vector \(x_{3n}\), which is compared with the code vector \(e_{3n}\) resulting the r-dim. error vector \(e_{3n}\). A BackPropagation algorithm is used to adjust the weight matrices through the adaptation matrices \(\delta W_1\) and \(\delta W_2\). A reduced version of the TDNN in Fig. 1 may be seen in Fig.2, in which the hidden layer has been eliminated, and the number of hyperplanes has been reduced to one (\(q=1\)).
The elimination of the hidden layer seems to be rather artificial, as there is not an objective associated with the input layer. Nevertheless, a virtual objective could be defined:

\[ z_n' = e_{2n} - x_{2n} \quad (1) \]

\( e_{2n} \) being the error back-propagated from the output layer. Besides, provided that \( \beta < 1 \), which is the case for robust convergence, the nonlinear function \( F \) may be approximated by:

\[ x_{2n} = \frac{1 - e^{\beta s_{2n}}}{1 + e^{\beta s_{2n}}} \equiv \frac{1}{2} \beta s_{2n} \quad (2) \]

As usual, the Energy of the Error is defined as:

\[ L = E(||e_n||^2) \quad (3) \]

where \( E(\cdot) \) is the Expectation Operator. The gradient of the Error Energy is:

\[ \frac{\partial L}{\partial w_i} = -E[e_{2n} x_{1n-i}] = -e_{2n} x_{1n-i} \quad 0 \leq i \leq k \quad (4) \]

\[ \frac{\partial L}{\partial w_k} = -E(e_{2n}) = -e_{2n} \quad (5) \]

where \( w_k \) is the threshold weight. In the above expressions an "stochastic estimator" has been used for the Expectation Operator, as it is customarily done in Neural Network Theory. Thus, the adaptation of the weights using the gradient, becomes:

\[ w_{ln} = w_{ln-1} + 1/2 \mu \beta e_{2n} x_{1n-i} \quad 0 \leq i \leq k \quad (6) \]

\[ w_{un} = w_{un-1} + 1/2 \mu \beta e_{2n} \quad (7) \]

where \( \mu \) is the step factor. From (7) \( w_{un-1} \) may be re-written as:

\[ w_{un-1} = w_0 + 1/2 \sum_{j=0}^{n-1} \beta \mu e_{2j} \quad (8) \]

As it is well known, a transversal prediction-error filter produces a "whitening" effect on the error series \([4, \text{ pg. } 215]\), which would force the threshold weight toward a stable value:

\[ w_0 = w_{un-1} + 1/2 \beta \mu \lim_{n \to \infty} \sum_{j=0}^{n-1} e_{2j} \quad (9) \]

Thence, the evaluation of the filtering error would result in:

\[ e_{2n} = z_{2n} - x_{2n} = z_{2n} - 1/2 \beta w_0 - 1/2 \sum_{i=0}^{k} \beta w_i x_{1n-i} = \]

\[ = z_{2n} - z_{2n}' = z_{2n} - 1/2 \sum_{i=0}^{k} \beta w_i x_{1n-i} \quad (10) \]

These expressions are related to the Adaptive Transversal Filter given in Fig. 3.

\[ L = E(||e_n||^2) \quad (3) \]

This conclusion establishes the conditions under which a hyperplane of the TDNN may be compared with an Adaptive Transversal Joint Estimator. We are interested in evaluating the optimal order \( k \) of this estimator given \( x_{1n} \) and \( z_{2n} \). This problem has found an optimal solution in the mean-square sense \([4, \text{ pg. } 231]\) using a Lattice Predictor instead of a Transversal Adaptive Predictor. This structure may be seen in Fig. 4.

\[ w_{un} = w_{un-1} + 1/2 \mu \beta e_{2n} \quad (7) \]

\[ w_{un-1} = w_0 + 1/2 \sum_{j=0}^{n-1} \beta \mu e_{2j} \quad (8) \]

As it is well known, a transversal prediction-error filter produces a "whitening" effect on the error series \([4, \text{ pg. } 215]\), which would force the threshold weight toward a stable value:

\[ w_0 = w_{un-1} + 1/2 \beta \mu \lim_{n \to \infty} \sum_{j=0}^{n-1} e_{2j} \quad (9) \]

\[ L = E(||e_n||^2) \quad (3) \]

This conclusion establishes the conditions under which a hyperplane of the TDNN may be compared with an Adaptive Transversal Joint Estimator. We are interested in evaluating the optimal order \( k \) of this estimator given \( x_{1n} \) and \( z_{2n} \). This problem has found an optimal solution in the mean-square sense \([4, \text{ pg. } 231]\) using a Lattice Predictor instead of a Transversal Adaptive Predictor. This structure may be seen in Fig. 4.

\[ w_{un} = w_{un-1} + 1/2 \mu \beta e_{2n} \quad (7) \]

\[ w_{un-1} = w_0 + 1/2 \sum_{j=0}^{n-1} \beta \mu e_{2j} \quad (8) \]

As it is well known, a transversal prediction-error filter produces a "whitening" effect on the error series \([4, \text{ pg. } 215]\), which would force the threshold weight toward a stable value:

\[ w_0 = w_{un-1} + 1/2 \beta \mu \lim_{n \to \infty} \sum_{j=0}^{n-1} e_{2j} \quad (9) \]
3. Delay Order Determination

Through this section a practical case to determine the order of a process will be presented. Assuming that \( u(t) \) is a Speech Trace, the PARCOR coefficients associated with it will be extracted using a Gradient Adaptive Lattice Algorithm of order \( p \), thus producing a set of traces \( x_n \), with \( i \leq p \), where \( i \) is the Sampling Time Index. These traces are re-sampled every 5 msecs., thus resulting in \( x_{jn} \), where \( n \) is the Decimation Index, and will be processed by a Lattice Predictor, to determine its Autoregressive Order as shown in Figure 5.

![Figure 5](image)

**Figure 5. Estimating the Autoregressive Order of trace \( x_{jn} \)**

\( L_{ek} \) of this trace obtained with a Lattice Predictor of order \( k=15 \) using (11). \( i_{ek} \) decreases strongly for \( i=1 \) (order 1), and keeps stable up to \( i=8 \), where a new decrement is observed. Finally, there are three additional minor decrements for \( i=10, i=11 \) and \( i=12 \). From (11) the relative descent in the Error Variance in each step may be quantified by \( h_j^2 \). We will compare this amount for step \( i \) against that for step 1, and detect the cases where this ratio is above the 10%:

\[
\frac{h_j^2}{h_1^2} \geq 0.1
\]

(12)

**Table 1. Lattice parameters for trace \( x_{9n} \)**

<table>
<thead>
<tr>
<th>( h_j )</th>
<th>-48</th>
<th>-02</th>
<th>-12</th>
<th>-01</th>
<th>-06</th>
<th>-01</th>
<th>-01</th>
<th>-17</th>
<th>-09</th>
<th>-15</th>
<th>-22</th>
<th>-17</th>
<th>-10</th>
<th>-02</th>
<th>-05</th>
<th>-09</th>
</tr>
</thead>
<tbody>
<tr>
<td>( e_{9n} )</td>
<td>0.06</td>
<td>0.01</td>
<td>0</td>
<td>0.13</td>
<td>0</td>
<td>0.10</td>
<td>0.21</td>
<td>0.12</td>
<td>0.04</td>
<td>0</td>
<td>0.01</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Table 1, the values of the Lattice parameters \( h_{9j} \) from trace \( x_{9n} \), and the selection criterion given by eq. (12) are shown.

**Table 2. Highest indices of the target Speech Traces**

<table>
<thead>
<tr>
<th>( x_{jn} )</th>
<th>/a/ ( a / )</th>
<th>/a/ ( a / )</th>
<th>/a/ ( a / )</th>
<th>/a/ ( a / )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_{1n} )</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>( x_{2n} )</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>( x_{3n} )</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( x_{4n} )</td>
<td>4</td>
<td>14</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>( x_{5n} )</td>
<td>1</td>
<td>2</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>( x_{6n} )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>( x_{7n} )</td>
<td>1</td>
<td>12</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>( x_{8n} )</td>
<td>16</td>
<td>2</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>( x_{9n} )</td>
<td>12</td>
<td>1</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>( x_{10n} )</td>
<td>1</td>
<td>6</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>( x_{11n} )</td>
<td>7</td>
<td>12</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>( x_{12n} )</td>
<td>15</td>
<td>14</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>( x_{13n} )</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( x_{14n} )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( x_{15n} )</td>
<td>2</td>
<td>15</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>( x_{16n} )</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>

4. Dimensioning TDNN's

If we examine the case exposed in Table 1, which was selected for being one of the worst-behaving cases, it may be checked that the action of the 15 stages from \( i=2 \) to \( i=16 \) produce a reduction in the Error Variance of only 71% of that one induced by the first coefficient. In \( x_{1n} \) corresponding to \( /a/ \), the first stage is responsible of a 97.8% reduction of the total Error Variance. This is the case with most of the traces treated, which could be represented by first-order autoregressive structures. To check this fact, a set of experiments has been carried out. The first one, used the traces \( x_{jn} \) corresponding to \( /a/ \) to train a TDNN in an experiment of Phonetic Decoding [5]. The TDNN had an input
layer of 16x(3+1)+1 nodes, a hidden layer of 8x(3+1)+1 nodes, and an output layer of 8 nodes (Network #1). The complexity of this network is 77.16% that of Waibel’s Network [2] for the same experiment (designated as Waibel’s Equivalent, or WE).

**Figure 7.a. Results after 96 training epochs for the Speech Trace corresponding to /aβa/ with a 65:33:8 TDNN (#1)**

**Figure 7.b. Results after 93 training epochs for the Speech Trace corresponding to /aβa/ with a 33:17:8 TDNN (#2)**

During each training epoch, the network received at its input 64 templates of 16-element PARCOR vectors, together with 64 8-element code words at its output. Fig. 7.a shows the weight matrices $W_{1H}$ and $W_{2H}$ organized by perceptrons after 96 training epochs using network #1. The Relative Error Energy (REE) after the training was reduced to $3.24 \times 10^{-3}$, from its initial value of 1. Perceptrons 1 and 6 of $W_{1H}$ show equivalent inverted versions. The same happens with perceptrons 2 and 3 with respect to 4, 5 and 8. Perceptron 7 remained idle. The most significant weights are concentrated in the 0-delay columns. In $W_{2H}$ perceptrons 1, 2 and 8 remained constant along the whole training epochs. Perceptrons 3 and 7 result in order 0 and order 1 autoregressive systems. Perceptrons 4, 5 and 6, reflect a similar inverted activity. These facts, justify a reduction of delay orders in the input and hidden layers to $k=1$ (Network #2)(39.37% of WE), the results being presented in Fig. 7.b. The number of training epochs to attain almost the same final REE ($3.24 \times 10^{-3}$) was 93, which is even lower than in the first case, with half the computational complexity, re-enforcing the point of view that first order systems may cope with most of the dynamic nature of the traces. Table 3 show the results for other speech traces processed.

**Table 3. Results for the different Speech Traces**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>/aβa/</td>
<td>95</td>
<td>93</td>
</tr>
<tr>
<td>/aβa/</td>
<td>99</td>
<td>122</td>
</tr>
<tr>
<td>/aβa/</td>
<td>93</td>
<td>107</td>
</tr>
<tr>
<td>/aβa/</td>
<td>83</td>
<td>98</td>
</tr>
</tbody>
</table>

**5. Discussion and Conclusions**

An important aspect to be highlighted from the results shown in Figs. 7.a and b is that the hidden layer for both experiments was oversized from the degree of redundancy observed in the perceptrons of $W_{1H}$ is high. Thence, the number of nodes in the hidden layer could be reduced to 2 for both cases reflected in Figs. 7.a and b. An experiment was set up to fix /aβa/, /aβa/, /aβa/ and /aβa/. 256 PARCOR vectors were used on each training epoch. The necessary number of training epochs for a network of type (#1) was 77. When a network of type (#2) was used, this number was 89. Reducing the number of perceptrons in the hidden layer to 3, 448 training epochs were required to attain similar results. This was twice the training time, although the resulting structure is the 15.25% of WE. As a conclusion, in most cases it does not seem justified to introduce more than one hidden layer, as the autoregressiveness of input data is traced by the input matrix, and that of output data is traced by the output matrix. Besides, the dimensionality of the hidden layer may be greatly reduced. To trace the non-stationarity of the Speech Trace the autoregressiveness of the parameters used must be studied. For most of the cases, this order may be reduced to one (two delay stages per layer).

**Acknowledgements**

This research is being funded by Comisión Interministerial de Ciencia y Tecnología, Grants Nos. TIC 93-0702-C02-01, and Grants Nos. CO72-90, B0792 and CO59392 from Plan Regional de Investigación de la Comunidad Autónoma de Madrid.

**References**


A Neural Network Based Solution for Multiple-Target Radar Tracking*

Valérie SCHMIDLIN(1)(2), Gérard FAVIER(1) and Bernard TOMASINI(2)

(1) IGS-CNRS/University of Nice-Sophia Antipolis, Bât SPI n°4 - 250, rue A. Einstein, 06560 Valbonne, FRANCE
phone: (33) 92.94.27.36, fax: (33) 92.94.28.98, e-mail: schmidl@imosa.unice.fr, e-mail: favier@imosa.unice.fr
(2) SYSTELIA Technologies, 6, Bd Maréchal Leclerc, 83320 Carqueiranne, FRANCE
phone: (33) 94.58.76.36, fax: (33) 94.58.76.48

Abstract: In this paper, we present a new neural network based solution for multiple target radar tracking problems, composed of two neural networks. The first one is an incremental network, and is dedicated to trajectory recognition: from all the plots received during one antenna turn, it recognizes which of them can be kinematically associated to already detected targets, and initializes new tracks with the remaining plots. The second one is a recurrent network, used to maintain the detected tracks, by performing data association: a Kalman filter delivers track predictions, from which coefficients of association between the previously detected targets and the received plots are calculated; then, the network takes the decision concerning the plot/track associations to perform. Simulation results for monosensor tracking are given to illustrate the behavior of the proposed neural network.

1- Introduction

Nowadays, different ways of research are under investigation for solving the multitarget radar tracking problem. At each antenna turn, a radar receives a certain number of plots (echoes originating from two kinds of sources: targets of interest and false alarms) to be correlated successively in time and space, in order to recover target trajectories (also named tracks). Generally, the tracking problem is decomposed into three steps: the first one, called track initiation, consists in finding the true tracks, as soon as possible, while minimizing the creation of false tracks. Usually, any measurement that is not used for track maintenance is considered as a track initiation, in other words as a potential track. If an initiated track is assigned to no plot, it is eliminated. The second step is the track maintenance, or track update, i.e. the data association process which consists in associating and correlating the data coming from a sensor. A lot of methods are available for solving this data association problem: nearest-neighbor correlation methods [1], Bayesian probabilistic methods such as the multiple hypothesis tracking (MHT) method [2], and the joint probabilistic data association (JPDA) method [3], and many variants of these methods. Other methods involving artificial intelligence techniques, fuzzy sets theory, evidence theory or neural networks, were recently developed. The last step is the track deletion, which occurs when the track quality is degraded, either because there is a lot of missing echoes in the tracking, either because this track was assigned with false-echoes. We can illustrate these different track life stages with figure 1.

![Figure 1. Track life stages](image)

This paper presents a new neural network based solution for performing the two first steps. The interest in using neural networks results from their properties: ability to adapt or to learn, to generalize, or to cluster or to organize data; capacity for fault tolerance with respect to noisy, distorted or incomplete data, and with respect to the physical degradation of the network itself, owing to their information storage in a distributed (or redundant) manner; fast processing speed due to their parallel processing capability; easy implementation and integration in almost any type of system. The paper is organized as follows. In section 2, we show how the tracks can be initiated. For this, we use an incremental network, which means that it chooses its size automatically, during the learning phase. After a brief overview of the existing neural networks with

---

*This work is supported by the Centre Technique des Systèmes Navals (CTSN-LSA), Toulon, France.
2 Track initiation with an incremental network

2.1 Existing networks with evolving structure

Usually, the choice of the network structure (number of layers and number of units in each layer for a feedforward network) is totally heuristic. Nevertheless, the structure directly influences the network performance, since a too big network will easily learn the learning collection, but will generalize very poorly (the network learns the examples "by heart", without looking for any relationship between them). Inversely, a too small sized network will not be able to remember all the examples, but will generalize very well from the learnt examples. In general, the structure is determined in trying many different structures, and keeping the one that gives the best results. Moreover, the computation of the synaptic weights is simultaneous with the construction of the neural architecture, the aim being to find the optimal neural structure, i.e. to get the best performance possible with a network of minimal size. Three principal ways can be investigated, that are briefly described in the following. Incremental networks start with only one hidden cell, and make the network grow by adding hidden units one by one, until convergence is reached. The network growth can be controlled by using different criteria: classification error minimization (Tiling algorithm [4], Upstart algorithm [5], Splitnet algorithm [6], [7], Cascade-correlation [8]), or geometrical consideration [9]. Decremental networks start with an over-dimensioned network, and eliminate useless cells one-by-one, either by evaluating the importance of the cell/connection (Skeletonization [10], [11], [12]) or by modifying the error-functional to be minimized [13]. Some solutions mix the two previous techniques: OFFSET [14] and GAL [15].

2.2 Track initiation with incremental neural network

We decided to develop a neural network, inspired from the GAL network. It is an incremental neural network composed of three layers of sub-networks, able to make trajectory recognition, building a tree-like distributed representation of the dynamic model of the track. Each neuron is a hypersphere, characterized by its center and its ray, and linked with few sub-networks in the following layer. These links are constructed during the learning phase. Each of these sub-networks corresponds to all the gyrations that can do the targets, starting from the neuron center.

![Figure 2. The trajectory recognition network](image)

This network registers all the possible spatio/temporal successions of four plots. Its architecture is determined from the knowledge of the possible target kinematics. We could have made a progressive learning from a collection of real tracks. The problem here is that we would need a complete library of prototype-tracks, which we can never, in practice, be certain to have. We designed a tiling algorithm that guarantees a convenient coverage of the observation space (target trajectories composed of four plots).

![Figure 3. The different search zones](image)

With this algorithm, we generate synthetic tracks, used to define the net. This net chooses its size, by stocking a library of prototype-tracks. Let's now consider the initiation of one particular track. First, we receive plot1, which is assigned to no pre-existing track. It is chosen as the reference of a new track, i.e. plot2, plot3 and plot4 are considered relative to plot1. At the next antenna turn, if we receive a second plot, say plot2, on the first layer of the network, then we associate plot2 to plot1, which association activates a search zone in the second layer. Since we know nothing about the speed and direction of the corresponding target, the first layer covers an entire ring around the first plot (determined by the minimum and maximum target speeds). Now, the track is composed of two plots: plot1 and plot2, and we have an idea about the
speed and the direction of the corresponding tracked target. This is illustrated in the network by the fact that not all the prototypes can correspond to this track, in other words that the search zone for the third plot is smaller than for the second plot. During the next antenna turn, if a plot is received inside this area, it is in turn associated to plot 2, extending the track with plot 3, and activating a search zone on the last layer. The same process is repeated for plot 4 in the layer 3. In this way, we obtain a track consisting of four kinematically possible plots. This kind of track is then considered as a confirmed track, that we need to maintain through data association, which is tackled in the following paragraph. Plots that do not associate with existing tracks are taken as the first plots of new tracks.

3- Track maintenance with a recurrent network

In the literature, plot/track association methods using neural nets have been developed, all of them based on the Hopfield network [16], [17], [18], [19], [20]. These methods have major drawbacks: (i) the network generally converges towards local minima of the energy function, (ii) the convergence is very time consuming, (iii) the Hopfield coefficients directly affect the neural network’s performance, and their choice is difficult to do, and at last, (iv) the proposed neural solutions generally do not take explicitly into account the cases of no plot-to-track association, that result from detection failures, track terminations, false alarms or new targets. The solution we propose for solving the data association problem is illustrated by means of fig. 4. Association coefficients are transmitted at the input of the network, which decides the plot/track associations to perform.

The optimal solution, which improves the convergence.

\[ E_{\text{obj}} = \frac{1}{M} \sum_{x \in I} \prod_{i=1}^{N} (1 - \alpha_{xi} V_{xi}) \]  

(2)

For more details, see [21], [22].

4- Simulations

By means of simulations, we have tested our neural solution in a monosensor tracking case. Six target trajectories have been synthesized by using a trajectory generator which simulates the kinematics of each target, as illustrated by figure 5.

![Figure 5. The real trajectories](image)

Figure 5. The real trajectories

![Figure 6. The plots](image)

Figure 6. The plots

![Figure 7. Tracking results](image)

Figure 7. Tracking results

The plots treated are the result of a plot generation program allowing to choose (among others): the false-echoes
density, the detection probability of the sensor, and the standard deviation of the noises in distance and azimuth. The used sensor is a two-dimensional radar sensor. Here, the characteristics of the simulation are: a detection probability of 1, no noise, and 80 false-echoes per antenna turn. Figure 6 shows the complete set of plots received during the simulation, and Figure 7 the tracking obtained with our neural solution. We see that all the targets are quickly detected, with very few false-initiation of tracks.

5. Conclusion and perspectives

The approach that we have developed and described in this paper, seems encouraging. It has been tested on real data, and at this time, we are comparing it with classical Hopfield techniques, and with the JPDAF. We first initiate tracks by a trajectory recognition technique based on an incremental neural network. The advantages of this method is a fast learning with a precise control on the internal representation of the network, and with an optimal structure. This network provides confirmed tracks, that are afterwards maintained, thanks to (i) a Kalman filter delivering the predicted position of the detected targets and (ii) a recurrent network taking the plot/track association decisions through neural optimization, in order to solve eventual conflicts. Several simulations validated our method, showing that the trajectory recognition method is able to refuse plot/track associations when they are kinematically impossible. The recurrent network, for its part, quickly and surely converges, through a global minimization. Future work will address the extension of our neural solution to the multiple sensor case, and the possibility of integrating other numeric (plot amplitude, as well as position) and symbolic data.

References

Application Studies of Co-occurrence Features in Neural Network Based Texture Classification

Changjing SHANG and Keith BROWN

Dept. of Computing and Electrical Eng., Heriot-Watt University, Edinburgh EH14 4AS, U.K.
Tel/Fax: +44 (0) 449 5111 1451 3327, E-Mail: {sj, keb}@uk.ac.hw.cee

Abstract. This paper presents a systematic investigation of the application of co-occurrence features in the texture classification systems where the classification process is implemented by a trained multi-layer feedforward neural network. It reveals the implications of the variation of the dimensionality of such feature patterns on both the training process and the classification performance of the network, thereby providing an important guideline for the utilisation of the co-occurrence features based texture classifiers in actual applications.

1. Introduction

Texture classification concerns the determination of different regions within a given texture image. It is particularly useful in the analysis of natural environments, such as classification of terrain and the side-scan sonar surveys of sea-floor for the exploration of mineral resources and environmental monitoring [3, 6, 7]. However, classifying such (large-scale) textures based on their potential features usually involves extensive computations. To avoid this problem more efficient computational methods are therefore necessary.

Recently, there has been much interest in utilising artificial neural networks (ANNs) as an architectural alternative to some of the statistically based pattern classification techniques. This is due to the nature of their parallel structure of simple computational elements and good classification performance that ANNs exhibit [2, 4, 5, 6]. Essentially, a neural network works after it has been trained to gain a knowledge of desired outputs from given inputs. Many approaches to developing ANNs exist, among them, the development of multi-layer feedforward neural networks (MFNNs) is one of the most successful [5]. In particular, an MFNN is capable of classifying texture images once representative features have been correctly presented to the network for it to learn. Viewing this, we use a trained MFNN to classify texture images represented by the co-occurrence features (extracted by the use of the co-occurrence matrices of the target texture images) [7]. Clearly, the successful training of the MFNN, termed the Pattern Classification Network (PCN) ensures obtaining a good classification performance of the entire classifier when being applied to an actual classification task. In particular, a PCN trained via the use of a fixed dimensionality of co-occurrence features classifies a target image by requesting every feature pattern to have that dimensionality. Nevertheless, given the structure of the PCN, the effectiveness and efficiency of a training process is almost solely determined by the co-occurrence feature patterns employed as the training input. Although this is apparent, it presents an important requirement for the system developer to attain knowledge of the effects that the use of those features with different dimensionalities in training may have on the classification performance of the classifier.

This paper presents a systematic investigation of such effects. To be self-contained, it first provides an overview of the feature extraction technique based on the co-occurrence matrix that is used to generate the input feature patterns for the PCN. Next, the co-occurrence feature based texture classifier is described, basically, consisting of the co-occurrence feature generator and a PCN. A substantial part of this paper is then dedicated to the presentation of the application studies of this classifier with considerable experimental results concerning the effects of the variation of the dimensionality of feature-patterns both on the training and the running of the classifier. Such an approach provides an important guideline for the actual application of the present texture classification system in dealing with real classification tasks.

2. Co-occurrence Features

The technique to generate co-occurrence features is based on an estimation of the second-order joint conditional probability density function \( f(i, j, d, a) \) of a given texture image [7]. Herein, the probability distribution \( f(i, j, d, a) \) reflects the frequency at which, within a given (sub-) image and along a particular direction of measurement \( a \), a pair of pixels appears in such a way that the distance between them takes a fixed value of \( d \) and their grey levels are \( i \) and \( j \) respectively. In general, if an image has \( G \) grey levels in representation the density function can then be written as a \( G \times G \) matrix and called the co-occurrence matrix. The most important and commonly used five types of feature measurements obtainable from such a matrix are given in the following:

1) Energy Type features:

\[
E_NG = \sum_{i=1}^{G} \sum_{j=1}^{G} (f(i, j, d, a))^2
\]

2) Contrast Type features:

\[
C_MT = \sum_{i=1}^{G} \sum_{j=1}^{G} [(i - j)^2 f(i, j, d, a)]
\]
3) Local Homogeneity Type features:

\[ LH = \sum_{i=1}^{k} \sum_{j=1}^{k} f(i, j, d, \alpha) \left( 1 + (i - j)^2 \right) \]

4) Maximum Probability Type features:

\[ MP = \max_{i, j, d, \alpha} \{ f(i, j, d, \alpha) \} \]

5) Entropy Type features:

\[ EN = - f(i, j, d, \alpha) \log(f(i, j, d, \alpha)) \]

In actual realisation of such a technique, as usual, we shall use a distance of one pixel and angles quantised to 45° intervals. Therefore, four matrices (respectively generated along the direction of 0°, 45°, 90°, and 135°, or horizontal, first diagonal, vertical, and second diagonal) are exploited. Following the definition of each of the five types of features, a total of 5x4=20 co-occurrence feature images are then produced. This is implemented by scanning a given image from the top to the bottom in a row by row manner using a size-fixed window. In so doing, a pixel at any position within a target image is represented by a vector of (co-occurrence) features or a feature pattern.

3. MFNN Based Texture Classifier

This section presents the texture classification system that consists of a feature generator and a PCN. As indicated earlier, we use the co-occurrence matrix technique to implement the feature generator. In particular, the implemented generator produces 20 co-occurrence features (see the preceding section).

Clearly, the task of the PCN within the classifier is to accomplish the feature pattern classification by mapping the co-occurrence features onto their corresponding underlying texture classes. The structure of the PCN is thereby specified as follows: the number of nodes in its input layer is designed to be \( i \), equal to the dimensionality of a feature pattern generated by the feature generator, and the number of nodes within its output layer will, of course, depend on the number of texture classes of interest in an application domain. The network is designed to contain two hidden layers [6].

The training of a MFNN for use as the PCN within the classifier is essential to its success in performing actual texture classification. For this purpose, a certain number of co-occurrence feature patterns that represent different regions of typically known texture images in the application domain, coupled with their corresponding underlying class indices, are selected as the training data. The back-propagation learning algorithm [5] is utilised to complete such training. Once the weights have been learnt, the PCN is constructed.

4. Application Studies

To enable an effective use of the neural network based classifier for a given texture classification task, this section provides a detailed study of two potentially very important effects resulting from the use of co-occurrence features, one upon the training process of the PCN and the other upon the execution of the trained PCN. We discuss both issues based on various experimental results obtained by using a varying dimensionality of the feature patterns. In the following, to have a common ground for comparison, we focus on the use of networks with each of their two hidden layers having a fixed number of hidden nodes to act as various PCN under the investigation. Also, a fixed number of training data are presented in the training of each PCN. Further, a composed texture image with a size of 384x384 in the range of grey levels varying from 0 to 255, as shown in figure 1, is used as the common training image within all the experiments. It consists of nine different textures, including plastic bubbles, fur, pigskin, calf leather, grass lawn, raffia, cork, hand made paper, and seaweed (normally coded as D111, D93, D92, D24, D9, D84, D4, D57, D87) [8].

Figure 1. Composed Texture Image

4.1 Impact of Co-occurrence Features on Training

As pointed out earlier, the success of the classifier in correctly classifying actual textures relies upon its successful training. In order to visualise the implication of the use of the co-occurrence feature patterns with different dimensionalsities in training a systematic, simulations have been carried out so that the PCNs, and hence the entire classifiers are trained by using \( i \) features with \( i \) ranging from 1 to 20. In terms of the structural specification of the PCN, the number of nodes within the input layer of each network differs accordingly, whilst the number of nodes within the output layer is designed to be a fixed value of 9 (indicating the number of different texture classes of interest) and each of the two hidden layers composes 18 hidden nodes. 5184 patterns are chosen from the co-occurrence feature images returned by the feature generator for use in training. As examples of the simulation outcome, learning curves for a PCN trained with a pattern dimensionality of \( i \), \( i=2, 4, 6, 8, 12, 16, 20 \), are plotted in figure 2.
These learning curves clearly illustrate that, in general, the normalised system error on the training set reduces when increasing the dimensionality of the feature patterns used in training. In particular, for the networks trained with feature patterns using a small dimensionality, (see those less than 8), although the training process tends to convergence faster than the others, the remaining system error on the training set is comparatively much larger than that obtained by using larger sized feature patterns. This indicates that if the dimensionality of the input feature patterns is too low it is rather difficult to accomplish the training process with a satisfactory system performance. The next sub-section will show that, as can be expected, using a trained PCN with a large remaining system error will not provide an acceptable classification performance when the PCN is in actual use. Thus, when considering designing a co-occurrence features based classifier, a trade-off is required between the speed of the training (which demands the use of feature patterns consisting of few features) and the accuracy of the training (which, in general, requires the use of a larger number of features in each feature pattern).

4.2 Impact of Co-occurrence Features on Classification Performance
Having accomplished the training of a network acting as the PCN within the classification system, the design process of the classifier is then completed. To examine the performance of a classifier trained with feature patterns certain dimensionality, the same composed texture image, from which only 5184 out of a total of 384x384 feature patterns were used as training input, is utilised as the target to run each of the implemented systems. The classification results are plotted in figure 3. To be clearer, furthermore, a set of classified images are shown in figure 4 with the boundaries between different classified texture regions being highlighted.
number of nodes at the input layer must then be the same as that minimum number.

<table>
<thead>
<tr>
<th>Threshold of Classification Rate</th>
<th>Number of Features Used</th>
<th>Actual Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 75%</td>
<td>4</td>
<td>22.53%</td>
</tr>
<tr>
<td>≥ 80%</td>
<td>5</td>
<td>19.73%</td>
</tr>
<tr>
<td>≥ 85%</td>
<td>7</td>
<td>14.42%</td>
</tr>
<tr>
<td>≥ 90%</td>
<td>9</td>
<td>7.75%</td>
</tr>
<tr>
<td>≥ 95%</td>
<td>10</td>
<td>4.29%</td>
</tr>
<tr>
<td>≥ 96%</td>
<td>12</td>
<td>3.62%</td>
</tr>
<tr>
<td>≥ 97%</td>
<td>16</td>
<td>2.89%</td>
</tr>
</tbody>
</table>

Table 1. Threshold of Classification Rate vs Pattern Dimensionality

5. Conclusion

This paper presents a systematic investigation of the application of co-occurrence features in the texture classification system which are implemented by feedforward multi-layer neural networks. The method to realise such classifiers is described. The paper reports on the implication of the variation of the dimensionality of feature patterns upon both the training process and the classification performance of the classifiers, demonstrated by detailed experimental results. Importantly, the presented experimental studies provide a useful insight into the determination of the number of co-occurrence features and hence the structure of the classification network (PCN) employed in a classifier for given application tasks.

Acknowledgement

This work was supported by UK-SERC grant GR/US5199.

References

A New Structure for Noise Filtering

Santiago MUÑOZ-MUÑOZ and José L. SANZ-GONZALEZ

†Centro Tecnológico de Madrid (CETEMA), Ctra. Valencia Km. 7 Centro Tecnológico "La Arboleda" 28031 Madrid, SPAIN, Tel/Fax: [+34 1] 331 2460 / 331 5472
††Departamento de Ingeniería de Circuitos y Sistemas, E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid, Ciudad Universitaria, 28040 Madrid, SPAIN, Tel/Fax: [+34 1] 549 5700 (ext. 384) / 543 9652, E-Mail: sanz@ics.upm.es

Abstract. The scope of this paper is to present a new network structure for additive and impulsive noise filtering using neural filters, as well as its similarity with other structures. The method employed here is based on the threshold decomposition of the signal, which is used as input to a group of neural network whose outputs are followed by a samples combiner. The network will be trained with the first samples of the signal, once the network is trained, this remains configured in order to filter the remainder of the signal. Results reveal that this new structure is more adequate than others when the noise is not additive or when there is no knowledge about the noise statistics in the signal.

1. Introduction

Most filters are linear, because they are easy to implement and analyze. They are optimal in the class of all filtering operations when the noise is additive and Gaussian; the linear filter performance deteriorates severely if those conditions are not assumed. In order to solve this problem, several new classes of nonlinear filters have been developed and sometimes neural networks are a good choice.

The behavior of neural networks in noisy environments was discussed by Klimasauskas [1], Anderson and Montgomery [2] who proposed the use of layered feed-forward networks for noise filtering (employing back-propagation learning rule). The signals used by them were a mixture of two chaotic signals or two random frequency sine waves, all corrupted with additive white noise.

In [3], we presented a similar neural network architecture for additive and multiplicative noise filtering, considering sine waves with noisy-amplitudes and noisy-frequencies. An averaging method was proposed in order to minimize the effect of the disturbance at the output.

Based on the threshold decomposition architecture and neural networks with sigmoidal function, a new class of nonlinear filters appear: the neural filters [4,5]. Neural filters can approximate both linear FIR (Finite Impulsive Response) filters and WOS (Weighted Order Statistic) filters, since sigmoidal functions can approximate linear functions and unit step functions. It is obvious that any neural filter with a linear neuron function reduces to a FIR filter.

Two classes of neural filters are defined, hard neural filters and soft neural filters. The hard neural filters, which perform Boolean operations, are defined by a set of neural networks whose activation functions are unit-step functions.

Soft neural filters are defined by neural networks whose activation functions are sigmoidal. The architecture of neural networks suggest that soft neural filters can approximate all filters defined by linear and nonlinear continuous functions such as linear FIR filters. Furthermore, soft neural filters can also approximate the hard ones because the sigmoidal function tends to be hard at high gains, that is, for large input weights.

The objective of this paper is to study a new structure that joints the features of the above described structures, performing a threshold decomposition of the signal to feed a group of typical neural networks used for data compression. In noise filtering there is usually a high degree of correlation between adjacent time samples. The contribution of the noise to the signal will tend to be random with respect to other strong features; then, the noise will most likely be one of the features eliminated in the compression process.

In most experiments the target vectors were samples of the pure signals and the input vectors were samples of the corrupted one, getting good results when we apply the signal with additive and multiplicative noise.

We point out that in some experiments, the input and target vectors were identical, and both of them consisted of sequential samples from the corrupted signal. Note that this training process does not use the pure signal, but does use the corrupted signal as pattern. Here, the most important attribute of the network architecture, is the very little knowledge about the signal and noise needed. Of course, knowledge about the signal may be useful for deciding a good network configuration.

The learning rule used was the back-propagation algorithm that is a gradient descent-based procedure to adjust thresholds and weights with the objective to minimize the error function. One way used to increase the learning rate without leading to oscillation was modifying the generalized delta rule to include a momentum term. This provides a kind of momentum in weight space that effectively filters out high-frequency variations of the error surface in the weight space [6].
2. Method

2.1 Neural Filters

The definition of neural filters is based on the threshold decomposition architecture [4]. Let \( R(t) \) be an arbitrary time continuous bounded signal. Let \( R(n) \) denote the sampled and quantized (\( M \) levels) version of \( R(t) \). This signal can be decomposed into \( M-1 \) binary sequences \( \{ r^m(n) \}, \) \( m = 1, 2, ..., M-1 \), by applying the threshold operation called \( T^m \), defined as follow.

\[
r^m(n) = T^m(R(n)) = \begin{cases} 1 & \text{if } R(n) \geq m \\ 0 & \text{else} \end{cases}
\]

Note that

\[
R(n) = \sum_{m=1}^{M-1} T^m(R(n)) = \sum_{m=1}^{M-1} r^m(n)
\]

Let a window of width \( L \), where \( L \) is an odd integer, slide across the input process \( R(n) \), and let \( R(n) \) be the vector containing the \( L \) samples in the window at time \( n \), i.e.

\[
R(n) = [R(n-(L-1)/2), \ldots , R(n), \ldots , R(n+(L-1)/2)]
\]

In similar way, by applying the threshold operation \( T^m \) to \( R(n) \), we obtain:

\[
x^m(n) = [x^m(n-(L-1)/2), \ldots , x^m(n), \ldots , x^m(n+(L-1)/2)]
\]

Where \( x^m(n) = T^m(R(n)) \).

Each of the \( M-1 \) binary vectors \( x^m(n) \) will be introduced into a neural network called level neural network, as is described below.

The structure of the neural filter is shown in Figure 1. There are \( M-1 \) level neural networks with \( L \) inputs. The vectors presented in each, will be composed by \( 0 \)'s and/or \( 1 \)'s corresponding to the threshold decomposition level of the input vector.

If we decide to use \( P \) number of patterns of the signal for the training, the structure has to train \( P(M-1) \) binary patterns. We proceed to train each neural network using the backpropagation learning rule presenting the patterns several times. Once all the level neural networks are trained, the structure will be able to filter the remainder of the signal.

2.2 Level Neural Network

The level neural network has three layers: input, hidden and output layer. The number of neurons in the input and output layer was identical and the middle layer always had fewer neurons. Outputs of each level neural network correspond with the filtered vector of the input binary vector. We must sum the response of those neurons with the same order of each level neural network, to obtain a filtered signal. So we have \( L \) versions of the filtered signal. In Figure 1, \( S(n) \) represents the vector with the different version of the filtered signal, which has the following aspect:

\[
S(n) = [S(n-(L-1)/2), \ldots , S(n), \ldots , S(n+(L-1)/2)]
\]

To obtain only one response we introduce the different versions of the filtered signal in a samples combiner.

The level neural network configuration and the samples combiner used is displayed in Figure 2. Where \( L \) and \( L \) are the number of units in each level.

In the samples combiner each version of the filtered signal is delayed and additive combined to obtain \( S(n) \), as is displayed in Figure 2, where \( x^m \) is a delay of \( k \) samples \( k=1,2,3, \ldots , L-1 \). The idea is to join the different responsesses presented in the output neurons into one.

![Figure 2. Architecture of Neural Network and Samples Combiner.](image)

Because, in some experiments, we use as the target vectors continuous samples of the pure signal (two continuous samples in the output signal do not have a large difference in their magnitude), sequential patterns used for training level neural networks, corresponding to the first levels, are very similar. Lower similarity appear in the next levels, so different momentums were used to improve the learning process and reduce the computing time of the learning in each network. For high similarity of continuous patterns, a large value of the momentum was used.

This architecture is also used for data compression, the approach followed here is to find a way of encoding or compressing the input data and reencoding it afterwards. Similar architectures have been used in other works (e.g. [7]).
The compression process eliminates portions of the input data, which represent small or nonrecursive features. The use of a limited number of neurons in the hidden layer ensures that some information (the noise) will be eliminated in the reconstruction. In which case, the internal representation of the patterns on the hidden units must be more efficient than the representation on the input layer. We attempt to map \( P \) input patterns for each level neural network onto \( P \) output patterns, suppose further that the number of hidden units \( N \) is provided by the equation \( \log P \), as is described in [6]. Because the network maps an input from a dimension \( L \) down to a lower dimension and back to dimension \( L \), any signal present in the input cannot be recovered exactly at the output. Thus, the optimization is necessary in order to the network be able to recover the pure signal from the corrupted one.

3. Results

To visualize differences in the performance of the new structure and some classical filters which are widely used in practice, we have simulated several experiments modifying the parameters of the structure: size of the window, number of levels in the threshold decomposition, number of inputs, hidden and output neurons, etc. In all the experiments, the level neural network used had three layers with the same number of neurons in the input and output layer, and fewer middle layer neurons.

In the first experiment, a combination of sine waves was corrupted with additive Gaussian noise. The second experiment consists of a combination of several random frequency sine waves, whose amplitudes and frequencies are corrupted by noise.

A particular case corresponding to the second experiment is presented below. The pure signal and the corrupted one used can be seen in Figure 3, where we present the two signals in the time domain. Figure 4 shows the power spectral density of both signals. The pure signal is presented by the continuous line and the corrupted one by the broken line.

![Figure 3. Pure and corrupted signal in the time domain.](image)

![Figure 4. Pure and corrupted signal Power Spectral Density.](image)

![Figure 5. Filtered and corrupted signal in the time domain.](image)

In Figure 6, 7 we present a comparison in the performance between a Median Filter, Wiener filter, Neural Network described in [3] and the new structure presented in this paper. In the first experiment, the performance of the structure is very close to that of the linear filter, as it is shown in Figure 6. Generally the Wiener filter gives the best result except for some sizes of the window, in which way the new structure manages a lower MSE.

In the second experiment, the superiority of the new structure to other filters is obtained for anyone size of the window, as we can see in Figure 7.

There are several important features of the architecture shown here. The first one is related to the number of neurons in the hidden layer and the other one to the number of neurons in the input/output layer.

By varying the number of neurons in the hidden layer, increasing from the initial value of \( \log P \), the amount of details retained by the filtering process decreases. This feature can be seen in [3], by different configurations of the network. On the other hand, by increasing the number of input and output neurons more filter selectivity is carried out.

In other cases, by varying the number of levels or the number of patterns used in the training process, generally we can achieve better filters, but the computing time spent in the learning process is increased.
The principal problem of the architecture is to decide the ideal dimensionality of the network: number of levels of the threshold decomposition, number of input and hidden neurons, patterns, etc. Furthermore, poorly constructed networks will do some filtering; however knowledge about the signal to be filtered, is very useful.

4. Conclusions

In this paper, we have proposed a new approach to noise filtering by a new structure based on threshold decomposition. This structure has a nonlinear filtering ability, because it is a multi-layer neural network whose middle layer has fewer units than those of the input or output layer. This structure is trained to map an identity using a window samples of the corrupted signal. Our goal here was to present the new structure filter not only for the additive noise but also for multiplicative and impulsive noise.

The experiment results showed that this structure performance was better for noise filtering than that of classical filters and conventional neural network. The performance of the structure can be further improved by allowing different level neural networks and learning rules to operate at each one.

In future works we will present other results of this structure involved with the handling of the different parameters of the structure. Also it will be present the application of these structures to real signals, such as voice or image signals, as well as their parameters more adequate.

Signals with other kinds of noises will be presented to prove structure robustness.

References

Optimum Genetic Algorithms for the Design of Stack-Filters

Neal R. Harvey and Stephen Marshall

Department of Electronic and Electrical Engineering, University of Strathclyde, 204 George Street, Glasgow G1 1XW, U.K., Tel/Fax: +44 (41) 552 4400 / 552 2487, E-mail: harve@spd.eee.strathclyde.ac.uk

Abstract. Genetic Algorithms are a relatively new paradigm for search, based on the principles of natural selection. They have been shown to be capable of use in the design of Stack-Filters. This paper presents a method of employing Genetic Algorithms with optimal performance in the design of stack-filters. Investigations into determining the set of parameters for optimum algorithm performance are described, together with the results and conclusions of these investigations. Results of applying the genetic algorithms, set up for optimal performance, to noise reduction tasks are shown. These results show that genetic algorithms are capable of efficiently searching for the optimum stack-filter in noise-reduction tasks.

1. Introduction

In recent years there has been a growing interest in the area of non-linear signal/image processing techniques. Blurring and other problems associated with linear filters lead to unacceptable performance in many applications. Stack-filters are a particular class of non-linear filters which possess a limited superposition property, making them attractive from the standpoint of ease of design. Stack-filters form a very large class of easily implemented non-linear filters having robust performance in applications where it is difficult to characterize the noise source. The class of stack-filters includes median filters, some morphological filters and rank-order filters. The task of stack-filter design necessitates the choice of a particular stack-filter from the huge number of stack-filter configurations available.

Genetic algorithms are probabilistic algorithms which provide a tool for investigating phenomena generated by complex systems. Problem areas which may benefit from such a tool are optimisation problems which are made difficult by substantial complexity and uncertainty. Genetic algorithms have been established as a valid approach to problems requiring efficient, effective and robust search in complex spaces.

By describing the abilities of genetic algorithms in search problems and the particular search problem of choosing a stack-filter from the large number of possible stack-filter configurations, it can be seen that genetic algorithms can provide a useful method for stack-filter design.

A method for the design of stack-filters using genetic algorithms has been described by Chu [1]. This describes a method of using a simple genetic algorithm to design stack-filters, together with some preliminary results. Detailed here are investigations into improving the genetic algorithm in order to optimise its performance, and the results of applying this optimum algorithm to noise reduction tasks are presented.

2. Stack-Filters

Stack-filters are non-linear digital filters, being a subset of all sliding window operators of a fixed window width. They have two defining properties: the weak superposition property known as threshold decomposition and an ordering characteristic known as the stacking property. A stack-filter is completely defined by the Boolean operation performed at each threshold level. Any minimum sum-of-products expression of a Boolean logic function which is free of uncomplemented literals obeys the stacking property, and is known as a positive Boolean function. There are exactly as many stack-filters as there are positive Boolean functions. Thus, there are 20 stack-filters of window width three, 7581 of window width five, and more than 238 of window width seven. The exact number of positive Boolean functions for window widths, , where > 6 is in excess of 278.

The structure of a stack-filter is simple and is known as the threshold decomposition architecture. An -level input signal is thresholded to levels giving separate binary signals. Each of these binary signals is operated upon by the defining positive Boolean function. At each time instant the binary outputs of the positive Boolean function at each threshold level appear as a column of zeroes on top of a column of ones. The multi-valued output of the stack-filter is obtained by summing the binary outputs at each threshold level or, alternatively, tracking the transition between the ones and zeroes.

3. Genetic Algorithms

A genetic algorithm is a formal implementation of an adaptive plan, based on the mechanics of natural selection, employing specific structural modifiers (operators) in the search for optimally performing structures in a particular environment. A population of individuals is successively modified from one generation to another. Genetic algorithms have been both theoretically and empirically proven to provide robust search in complex spaces and, are therefore finding widespread application in business, science and engineering. Genetic algorithms are computationally simple, yet they are powerful in their search for improvement.

Individuals in the population are known as chromosomes and are made up of strings of genes. Each gene can have several different forms, know as alleles, producing differences in the set of characteristics associated with that gene.

The metaphor underlying genetic algorithms is that of evolution. In evolution, the problem each species faces is one of searching for beneficial adaptations to a complicated and changing environment. The knowledge that each species has gained is embodied in the makeup of the chromosomes of its members. Operators that alter this chromosomal makeup are applied when parents reproduce. Among these genetic operators are mutation and crossover, the most commonly used operators in genetic algorithms.

Crossover is basically the exchange of genetic material.
between two parents' chromosomes, exchanging corresponding genetic material from the two parent chromosomes, allowing beneficial chromosomes on different genes to be combined in their offspring. Mutation provides background variation and occasionally introduces beneficial material into a species' chromosomes.

For a genetic algorithm to perform its adaptive task, there has to be some means of assigning an observable performance to each individual. A fitness function is employed which attributes some objective measure of how well an individual performs in an environment—the fitness of an individual. At the beginning of each time period (generation) the accumulated information about the environment resides in the finite population. The population serves not only as the GA's store of accumulated information, but also as the source of new variants which will form the next generation. Formation of the next generation proceeds in two phases: During phase one, the population is modified to produce an intermediate population, by copying each individual in the population, the number of times being dependent upon the individual's observed performance. The number of copies made is determined stochastically, so that the expected number of copies made increases in proportion to its observed performance. During phase two, genetic operators are applied to the intermediate population, interchanging and modifying sets of alleles, to produce the new generation. Figure 2 shows an overview of the components of a simple GA.

4. Stack-Filter Design using Genetic Algorithms

Space prohibits a detailed description of the scheme for the representation of stack-filters as chromosomes. For a detailed description the reader should refer to Chu's paper [1]. It is sufficient to state that the chromosome is a binary string, where each binary gene in the chromosome can represent a product term of the minimum sum-of-products expression of a positive Boolean function. A description of the simple genetic algorithm used by Chu can also be found in [1]. In this paper we describe modifications made to this genetic algorithm in order to improve its performance. The principal difference was that an additional function, fitness scaling, was performed between the calculation of the fitness values and the selection process.

4.1 Fitness Scaling

Regulation of the number of copies is important in GAs with relatively small populations. At the start of GA runs it is quite common to have a few extraordinary individuals in a population of mediocre colleagues. If left to a standard selection rule, the extraordinary individuals would take over a significant proportion of the finite population in a single generation, which is undesirable, and a leading cause of premature convergence. Later in a run, there may still be significant diversity within the population; however, the population average fitness may be close to the population best fitness. Left alone, this situation causes average members and best members to get nearly the same number of copies in future generations, and the survival of the fittest, necessary for improvement becomes a random walk among the mediocre. Fitness scaling can help in both situations: at the beginning of the run and as the run matures, helping to prevent the early domination of extraordinary individuals while later on encouraging a healthy competition among near equals. A detailed description of fitness scaling can be found in [3, pp 76-79].

4.2 Genetic Algorithm Performance

To quantify the effectiveness of different genetic algorithms, performance measures are used. Two such measures are those devised by De Jong [4]: gauge convergence and ongoing performance, called off-line performance (convergence) and on-line performance (ongoing). The on-line performance can be defined as an average of all function evaluations up to and including the current trial and off-line performance can be defined as a running average of the best performance values to a particular time.

4.3 Genetic Algorithm Parameters

The parameters affecting genetic algorithm performance, such as probabilities of crossover and mutation, population size and number of generations over which to run the GA were investigated to determine the set of parameter values giving optimum performance.

5. Results and Discussion

5.1 Probabilities of Crossover and Mutation

In order to determine the best combination of crossover and mutation probabilities, the crossover and mutation probabilities were systematically varied and for each pair of crossover and mutation probabilities, the GA was run over a fixed number of generations using the same data set. The on-line and off-line performance of the GA at the end of each run was then determined. The results of this are shown in Figure 3. It can be seen that there is a marked decrease in both off-line and on-line performance with an increase in the probability of mutation, whereas the performance of the GA, generally, increases slightly with increasing probability of crossover, though the change is not as dramatic as that for the probability of mutation. We can conclude from these results that a very low probability of mutation and high probability of crossover will give the best GA performance. A probability of crossover of 0.9 and a probability of mutation of 0.001 were therefore adopted for the subsequent optimal GA.
5.2 Population Size

The population serves as the plan’s store of accumulated information. The greater the population size, the greater the amount of information processed about the search space in each generation. The greater the amount of information processed, the higher the convergence rate should be toward the optimum. Unfortunately there is a practical consideration. There is a trade-off between population size and computation cost: the larger the population size, the greater the computation cost. It would seem logical to ascertain the best compromise between the two. In order to determine the population size making the best compromise, the population size was varied and for each population size, the GA was run over a fixed number of generations, with the same data set. The on-line and off-line performances of the GA, at the end of the run, were calculated. Figure 3 shows these results. It can be seen from these results that on-line performance decreases with increasing population size, whereas off-line performance increases with increasing population size up to a population size of about 80, at which point there is no appreciable increase in off-line performance with any further increase in population size.

The reason for the decrease in on-line performance with increasing population size can best be attributed to the inertia of a larger population. Smaller populations are able to change more rapidly and may thus have better initial performance. As the applications of the GA we were interested in were off-line, the results for the off-line performance were more relevant. Therefore, the best compromise between population and computation cost would appear to be a population size of 80.

![Figure 3. Variation of GA performance with population size](image)

6. Optimum Genetic Algorithm

The parameter values, determined as described above, were incorporated into a genetic algorithm and this algorithm was run using portions of two images as the input data set. These two images consisted of an uncorrupted image, shown in Figure 4a, and the same image corrupted with positive and negative impulsive (salt and pepper) noise, shown in Figure 4b.

![Figure 4. (a) Uncorrupted test image (b) Test image corrupted with positive and negative impulsive (salt and pepper) noise.](image)
algorithm appears to be converging towards the optimum, and for every additional generation there is little improvement in algorithm performance. A possible stopping criterion could be based on the variance of the performance values of a number of the preceding generations, e.g. if the variance of both the on-line and off-line performance over the last N generations is less than a threshold value, then the algorithm could be assumed to have converged and could be stopped. The best performing stack-filter configuration to that time could then be considered as the final outcome of the algorithm. Investigations are ongoing into the determination of a stopping criterion suitable for noise reduction tasks such as those described here.

![Figure 5. Performance of optimal genetic algorithms with input data set shown in Figure 4.](image)

$$f(x_0, x_1, x_2, x_3, x_4) = x_0x_5 + x_0x_1x_2 + x_1x_2x_3 + x_0x_1x_4 + x_0x_2x_4 + x_1x_2x_4 + x_1x_3x_4$$

![Figure 6. Stack-filter configuration with best performance found after running optimum genetic algorithm](image)

7. Conclusions

The parameters affecting the performance of a genetic algorithm for the design of stack-filters were analysed and the set of those parameter values were found which optimised the GA's performance, whilst keeping computation cost as low as possible. This genetic algorithm performed well, converging rapidly toward the optimum stack-filter configuration for noise-reduction applications. The use of GAs in the design of the related field of mathematical morphology has led on from the work described here. GAs have been used in the selection of filter types and structuring elements, for optimum performance in noise-reduction tasks. These results are presented in [8]

![Figure 7. Result of filtering corrupted image with stack-filter, shown in Figure 6, found using optimal GA.](image)

<table>
<thead>
<tr>
<th>MAE</th>
<th>MSE</th>
<th>Corrupted Image Filtered with Stack-Filter found by optimal GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.35</td>
<td>22.44</td>
<td>634.52</td>
</tr>
</tbody>
</table>

Table 1. Comparison of Original, Corrupted and Filtered versions of the images, with respect to mean absolute error and mean squared error.

8. References

Genetic Optimization of Piecewise Self-Affine Fractal Interpolation with Application to Speech Modeling

Paolo AGATI and Enzo MUMOLO

DEEI, Universita' di Trieste, Via Valerio 10, 34127 Trieste, Italy
Tell/Fax: +39.40.6763414/3460, E-mail: mumolo@uits340.univ.trieste.it

Abstract. It is known that speech signals possess some fractal properties. A novel approach for speech modeling is therefore by means of fractal models. However, when speech is neither self-affine nor self-similar, suitable fractal models, such as the piecewise self-affine ones, should be used. In this paper we describe an optimal determination of the parameters of the piecewise model using Genetic Algorithms. A proper tuning of the algorithm was first performed. Then, different types of constraints on the search space were analyzed and the best trade-off between SNR, compression ratio and computational complexity was found. Finally, the fractal models estimated with GA was applied to speech signals. Some results, namely original and synthetic signals and segmental SNR, will be reported.

1. Introduction.

The fractal geometry extends the euclidean geometry and describes the objects characterized by a non-integer topological dimension. An important property of the fractal sets, called self-similarity, implies that by increasing the degree of magnification more and more details are revealed. This property holds in general for real-world images and also for many signal and speech waveforms [1]. The Iterated Function System (IFS) [3], is a powerful mathematic tool for the representation and the generation of fractal sets. An IFS is a set of contractive affine transformations. Since in this paper we want to represent a function, the affine transformations suitable for us are the so called shear transformations:

\[
\begin{pmatrix}
\quad \!
\end{pmatrix} \begin{pmatrix}
\quad \!
\end{pmatrix} = \begin{pmatrix}
0 & c_y \\
0 & 1
\end{pmatrix} \begin{pmatrix}
\quad \!
\end{pmatrix} + \begin{pmatrix}
\quad \!
\end{pmatrix}
\]

where the \( d_i \) parameters are called contraction factors. A shear transformation maps vertical lines into vertical lines, and therefore it can represent a single-valued function. The types of data that can be represented well with a deterministic fractal model, such as the IFS model, is limited to the self-similar or approximatively self-similar data. As a matter of fact, however, many non stationary signals like speech, often show a periodic behaviour. For such signals a piecewise self-affine fractal model [2] is more appropriate. Given a data sequence to be modeled, the piecewise self-affine model requires the determination of a number of interpolation points, a number of sections associated to each interpolation point and a number of contraction factors such that the function generated by the model is close to the given data sequence. Thus, the parameters determination leads to an optimization problem. However, this optimization problem is very complex, because it means nonlinear minimization with respect to many variable. Since a complete analytic solution to this problem cannot be envisaged, algorithms which make an exhaustive search of the solution over the search space have to be used. Efficient, though sub-optimal, solutions can be obviously devised if constraints on the search space are imposed. The algorithm described in [2], for example, explores only 32 out of \( 1.32\times 10^{36} \) possible solutions. The main motivation of this work was therefore to use an efficient and robust stochastic algorithm for solving the described optimization problem. We have chosen Genetic Algorithms (GAs) mainly because they achieve better convergence to the global minimum than other stochastic algorithms. Indeed, in this work we have experienced that very good solutions to the described optimization problem can be obtained with GAs. This way, moreover, it was possible to test several different versions of the piecewise fractal interpolation model and to come out with the best trade-off. Finally, we used the piecewise fractal interpolation model to model speech and electro-glottographic signals.

2. Piecewise Fractal Interpolation.

Let us assume that a number of data points \((u_n, v_n)\), \(n=0..N\) and \(u = u_M\) be given. A set of interpolation points \((x_i, y_i)\), \(i=0..M\) and \(M+N\) be also given. The interpolation points are a subset of the data points with \((x_0, y_0)=(u_0, v_0)\), \((x_M, y_M)=(u_N, v_N)\). At each interpolation point \((x_i, y_i), i=0..M\) an affine map \( w_i \) and two points, called addresses in [2], described as \((x_i^1, y_i^1)\) and \((x_i^2, y_i^2)\) are associated. The affine maps \( w_i \) map the function between the addresses and the interpolation points. Therefore the following conditions can be written:

\[
\begin{pmatrix}
\quad \!
\end{pmatrix} \begin{pmatrix}
\quad \!
\end{pmatrix} = \begin{pmatrix}
\quad \!
\end{pmatrix} \begin{pmatrix}
\quad \!
\end{pmatrix} \begin{pmatrix}
\quad \!
\end{pmatrix} ; i=0..M
\]
In other words, each section of the function between the interpolation points, \( F_i \), is the piece of the function between the related address, \( F_a \), mapped to that section by map \( w_i \):

\[
F_i = w_i(F_a)
\]  

(3)

From (2) it comes out that, if the interpolation points and the addresses are known, a set of 4 equations into 5 unknowns (the IFS's parameters) can be written \([2], [3]\). Of course, this means that one parameter, usually the \( d_i \)'s, should be computed in some way. Throughout this paper, we used the optimal determination of the \( d_i \)'s factors devised in [2], which is based on a mean-squared error determination. Moreover, in [2] some constraints on the interpolation point and addresses were imposed. Because of these constraints, the algorithm proposed in [2] is suboptimal. The initial motivation of our work was to get rid of these constraints in order to look for optimal solutions. The problem was formulated within the context of Genetic Algorithms (GA) optimization.


GAs are optimization and search algorithms \([4]\). In a genetic algorithm the solutions of a problem are coded using strings built with elements of a particular alphabet, in our case \([0,1,\ldots,32767]\). These strings are called 'chromosomes'. The process performed by GA is iterative, and at each iteration a small set of individuals is chosen by a selection procedure according to their fitness value, which is a measure of how well they satisfy the optimization function. To interpolate between states with high fitness values, the crossover operators genetically recombines individual states. New individuals are created by encoding their values as bit strings of a given length. Two parents states from the new population are chosen randomly, a crossover position within the string is chosen at random and the two parents exchange portions of their binary representations. This process continues until enough children are produced to replace the parent population, although some parents may kept in the new population. Occasionally, i.e. with a small probability, individual states are altered and this process is called Mutation. The results is that the population samples states away from its mean, preventing the population from converging at any minima. The whole process (i.e. selection, crossover and mutation) continue until an acceptable solution is reached.

In our implementation each element of the population has only one chromosome. As described above, each iteration of the process, called 'generation', new solutions are created to replace the old population, applying the crossover and mutation procedures to the chromosomes. The way one choses the parents of a new solution is of most importance. The parents are chosen between the best scoring elements of the population, i.e. the ones with the best values of the objective function. In the simplest procedure, Roulette Wheel Selection, (RWS), parents are chosen according to a probability distribution based on the fitness of the elements of the population. We have found that another procedure, the so-called Local Selection, (LS) gives the best results: in this procedure the elements are arranged on a bidimensional grid and two parents are chosen through a walk over the grid, starting from the element to be replaced, selecting the element with the best fitness. This kind of selection, while giving the chance to mix the genetic patrimony, maintains a good genetic diversity and thus preserves the process from premature convergence.

GAs allow to solve the described optimization problem by getting rid of the above mentioned constraints imposed on the addresses and interpolation points. For instance, if we want to find the optimum \( d_i \)'s and addresses positions, a chromosome must be set as follows:

\[
\{d_1, x_1^1, x_1^2, d_2, x_2^1, x_2^2, \ldots, d_n, x_n^1, x_n^2\}
\]  

(4)

Thus, depending on what and how many parameters are included in the optimization process, a number of optimal solutions can be obtained. Each solution is a trade-off between computational power and performances, in terms of compression ratios and SNR.

4. Tuning of the Genetic Algorithm

Looking for the best GA configuration, in terms of selection, crossover and mutation, several experiments were carried out. The results reported in this section have been obtained using a 16 ms frame of voiced speech sampled at 8 KHz.

In Fig. 1a, 1b the convergence behaviour is shown. In Fig.1a the minimum value at each generation and in Fig.1b the average value of the population are shown. The three selection procedures, namely RWS, RWM and LS have been tested, and it turns out that the best procedure is LS, since it gives a better convergence with a smaller difference between minimum and average values as shown in Fig.1a.

In the context of GAs, crossover is the operator which realized the genetic recombination, which is fundamental because it allows for an efficient sampling of the solutions space. If crossover is too low, its effect is negligible and a premature convergence is eventually obtained. On the other hand, if it is too high, it eventually rejects the high fitness solution. In the classical 1-point operator, the crossover can happen anywhere inside the genetic string, while in the so called 1-point modified operator the crossover can happen only at boundary between groups of genes which represent an IFS map. Experimental result indicate that the best crossover method is the 1-point modified.

As regard mutation, three approaches have been considered, namely the soft, hard and dynamic mutations. Soft mutation is performed by increasing the integer value of one parameter, chosen with a given probability. Hard mutation is performed by changing the value of a parameter with a random value, and dynamic mutation is
performed increasing the value of a given parameter by a value which decreases, according to a given function, towards the end of the process. We found that the best mutation operator for our optimization problem is the dynamic mutation.

![Figure 1. Minimum (a) and average (b) values versus the number of generation.](image)

We have compared the following versions of the fractal models:

a) self-affine fractal interpolation with genetic optimization of the interpolation nodes;
b) piecewise self-affine with uniformly distributed interpolation nodes and genetic optimization of the addresses;
c) piecewise self-affine with genetic optimization of interpolation nodes and addresses;
d) piecewise self-affine with uniformly distributed addresses and genetic optimization of the interpolation nodes;
e) piecewise self-affine with uniformly distributed addresses and interpolation nodes, and genetic optimization of the indexing between interpolation sections and addresses;
f) piecewise self-affine with addresses equal to the interpolation nodes and genetic optimization of the nodes and the indexing between interpolation sections and addresses.

The test signal has been modeled with a piecewise self-affine interpolation model and the parameters optimized with a GA with the following characteristics: Local Selection, Modified 1-point Crossover with a 90% probability and dynamic mutation with a rate of 1/100. The SNR results related to the versions a)-f) are reported in Fig. 2 where the results obtained with the efficient sub-optimal algorithm described in [2] are also reported. With 16 IFS maps, the efficient algorithm gives an SNR equal to 4.5 dB, a value which is largely outperformed by all the other versions. For example, with method f) more than 12 dB are obtained with 16 IFS.

![Figure 2. SNR results (dB) for the different models versions versus the number of IFS maps.](image)

The computational requirements, we measured in terms of execution times, of the described methods are shown in Fig. 3. These data have been normalized with respect to the efficient algorithm. It is shown that, while for 16 IFS maps the difference between the efficient method and the method f) is of 1 order of magnitude, for higher number of IFS the difference is smaller. For 24 transformations the execution time is almost the same.

![Figure 3. Normalized execution time for the different versions](image)

5. Experimental Results
The piecewise fractal interpolation model, version f), estimated with GA has been used to model some speech signals, and the results will be reported in this section. In Fig. 4 a vowel onset is shown, original overlapped to synthetic. With 20 IFS maps the SNR was 11.4 dB.

![Figure 4. Fractal modeling of a vowel onset, original and synthetic.](image)
In Fig. 5 the fractal modeling of the italian word /volano/ is shown. The signal has been divided into 128 samples frames and analyzed. In the first and second panels the original and modeled (with 24 IFS) signals are shown, and in the third panel the segmental SNR is reported.

![Waveform](image)

**Figure 5.** Fractal model of the italian word /volano/. Original (a), synthetic (b) and segmental SNR (dB) (c).

Finally, the piecewise model was applied to a segment of Electroglottographic (EGG) waveform [5] and the results are shown in Fig.6. The EGG signal, which describes the opening between the vocal folds and is therefore linked to the volume velocity, is obtained with electrical impedance measurements. EGG has many applications, such as detection of larynx pathologies, speech analysis (for the V/UV classification, pitch and formant detection, estimation of glottal volume velocity), coding (for the glottal vocoder), recognition and synthesis. In any application area, its modeling with the described technique is of interest. We used method 5 with 16 IFS transformations per frame and 256 samples per frame. With parameters coded at 7 bits and contraction factors coded at 4 bit, we achieved a compression ratio of 1:8.11.

![Waveform](image)

**Figure 6.** Fractal model of an electroglottographic signal. (a): original and synthetic (overlapped); (b): segmental SNR; (c): zoomed portion of (a)

### Acknowledgments

This work was supported by MURST 60% and by MURST 40%, "Algoritmi e Strutture di Calcolo", Rome, Italy.

### References


An Optimization Algorithm for Fractal Encoding of Graytone 2D Images

Anna Maria CARMINELLI GREGORI, Enzo MUMOLO and Igor BONAT

DEEI, Universita' di Trieste, via A.Valerio 10, 34127 Trieste, Italy
Fax: +39.40.6763460, E-Mail: carmin@univ.trieste.it

Abstract. In this paper we describe an algorithm to find an Iterated Function System with associated probabilities (IFS) in order to generate a given coloured image. The proposed optimization algorithm, based on the Gradient method combined with Simulated Annealing, is much faster than the Simulated Annealing optimization algorithm. Actually, more than one order of magnitude improvement in computational time has been gained. The overall algorithm can be viewed as a Simulated Quenching (SQ) method, which is quite efficient for the problem under consideration. Convergence curves are reported and discussed and an example of coded image is included.

1. Introduction
In the context of the fractal geometry introduced by B.Mandelbrot the Iterated Function System (IFS) theory [1] represents a powerful technique for the generation of fractal objects. Fractal image encoding requires the solution of the following "inverse problem": given an image A, find the IFS which reconstructs A. This problem has attracted much attention in the recent literature [2],[3], [4],[7],[11]. Since we want to reconstruct a synthetic image as close as possible to a given image, the inverse problem can be formalized in a mean squared error framework. Nevertheless, the mean squared error is a non linear function with many local minima and therefore the optimization problem is quite difficult to solve. In [4] two different algorithms, based on Steepest Descent (SD) and Simulated Annealing (SA) approaches respectively, were described. Some drawbacks of such approaches, such as the local minimum problem or the high computational time, have been pointed out. In this paper we describe an optimization algorithm for the solution of the described IFS inverse problem. Several different images have been analyzed with this algorithm and we found that a good minimum of the mean squared error is always reached in much less time (more than one order of magnitude) than SA. Since our approach is based on the SA paradigm with an annealing schedule controlled by SD, it can be considered a Simulated Quenching algorithm [8].

This paper is organized as follows: in Section 2 the basic IFS theory is described, and in Section 3 the proposed algorithm is described. In Section 4 some results concerning convergence are discussed and a practical example is reported.

2. Probabilistic IFS and the Optimization Problem
An IFS is defined as a complete metric space K and a finite set of affine contractive mappings W = {w1, w2, ..., wN}. In the context of IFS theory the auto-similarity property, which is typical of fractal objects, can be expressed as follows:

$$A = \bigcup_{k=1}^{N} w_k(A)$$

where A is a compact set, A∈K. An IFS with probabilities, or the probabilistic IFS, consists of an IFS, defined on a compact space K, together with a set of probabilities {p1} associated with {w1}. In this way each affine contractive mapping wk occurs with probability pk. With these data it is possible to define a mass distribution and in turn a finite measure μ on the Borel subsets of K, on the field generated by the IFS and on the attractor of the IFS. Let P(K) denote the space of normalized Borel measure on K.

If μ ∈ P(K) and w:K→K is continuous, then also v(w−1) is continuous.

The Markov operator associated with the probabilistic IFS (with contractivity factor s), is the function M: P(K)→P(K) defined by:

$$M(\psi) = p_1 \cdot \psi(w_1^{-1}) + p_2 \cdot \psi(w_2^{-1}) + ... + p_0 \cdot \psi(w_N^{-1})$$

for all ψ ∈ P(K).

It can be shown that M is a contraction mapping with the same contractivity factor s, with respect to the Hutchinson metric on P(K). This same theorem about the operator M, proves the existence of the unique fixed point μ such that:

$$M\mu = \mu$$

where μ is called the invariant measure of the probabilistic IFS.

Another theorem proves that the support of μ associated with the probabilistic IFS is the attractor of the simple IFS {K; w1, w2, ..., wN}. Let {X; w1, w2, ..., wN; p1, p2, ..., pN} be a probabilistic IFS in the Euclidean 2-dimensional space X. The Random Iteration Algorithm is described as follows:
starting from any point \( Z \in X \), select randomly, according to the \( \{ p_k \} \) distribution, one of the \( w_k \). Apply this affine contractive mapping to \( Z \) and plot the new point \( Z = w_k(Z) \). Iterating this process, a "Markov chain" asymptotically stationary is obtained \([7]\), as it converges in distribution to a unique fixed distribution with support the attractor \( A \) of the simple IFS.

After \( M \) iterations, the chance of visiting pixel \( Y \) is:

\[
P(Y) = \text{value of the pixel } Y/M.
\]

Assigning to each pixel \( Y \) a visiting probability \([9]\), with the following conditions:

\[
\begin{align*}
\text{value of the pixel } Y/M & \geq 0 \\
\text{value of the pixel } Y/M & = 1
\end{align*}
\]

we can obtain a probability measure proportional to \( \mu \).

Thus, given an image \( A \) defined by a matrix of normalized values, the optimization problem can be formulated as follows. According to (2) and (3) the following relation can be written:

\[
\mu(E) = \sum_{k=1}^{N} p_k \mu(w_k^{-1}(E))
\]

for every \( E \) belonging to \( K \) and where \( \mu \) is a normalized Borel measure on \( K \). This relation represents the invariance of the measures with respect to the Markov operator associated to the probabilistic IFS: where \( A(i,j) > 0 \), we have

\[
\mu = \lambda A.
\]

The numerical representation of the union of the cover elements \( A_k = w_k(A) \) is given by

\[
B = \sum_{k=1}^{N} p_k A(w_k^{-1}(i,j)) = \sum_{k=1}^{N} p_k A(w_k(i,j))
\]

(6)

The minimum of the distance \( F \) between \( A \) and \( B \) can be found with the following equation:

\[
\min \{ F \} = \min \{ B - A \}^2
\]

where

\[
F = \sum_{i,j} \left[ \sum_{k=1}^{N} p_k A(i,j) - A(i,j) \right]^2
\]

In this equation the unknowns are the \( w_k \) functions and the probabilities \( p_k \), while \( A \) is the given image. This last equation has been written by considering the measures \( \mu \) for the probabilistic IFS.

It has to be noted that the shape of the fractal image is given by \( W \), while the gray levels, or the colours, of each pixel are controlled by the probabilities \( p_k \).

3. The Proposed Algorithm

The minimization procedures described in \([4]\) and \([11]\) are based on steepest descent (SD) and on simulated annealing (SA) methods respectively. It was reported that SA works better than SD because it always reaches the global minimum of the cost function or a local minimum close to the optimum one, but it requires a very high computational time.

3.1 The SD method

The SD method suffers of the typical problem of the iterative methods: if the starting point of the iteration is not sufficiently close to the optimum, the point reached at the end of the iteration is a local minimum, and there is no way to control the goodness of such minimum. A way to face this problem is to use a multi-start approach, but it is not assured to reach the global minimum or to know how to reach a good local minimum.

In order to minimize the function described in eq.(7) we compute its partial derivatives with respect the coefficients \( c_k \) of the affine contractive mapping \( w_k \) for \( i = 1,2,...,N \) and \( J \) with respect the probabilities \( p_k \). The \( w_k \) is the general affine contractive mapping with \( k = 1,2,...,N \). The derivatives of \( F \) with respect to \( c_k \) are:

\[
\frac{\partial F}{\partial c_k} = 2 \sum_{i,j} \frac{\partial}{\partial c_k} \left[ A(w_k^{-1}(i,j)) \right] \sum_{k=1}^{N} p_k \left[ A(w_k^{-1}(i,j)) - A(i,j) \right]
\]

(9)

The derivatives of \( F \) with respect to the probability \( p_k \) are:

\[
\frac{\partial F}{\partial p_k} = 2 A_k(i,j) \sum_{k=1}^{N} p_k \left[ A(w_k(i,j)) - A(i,j) \right]
\]

where \( A_k(i,j) = A(w_k^{-1}(i,j)) \).

SD is performed with the algorithm described in \([6]\), with eq.(9) and (10), and with the constraints that each \( w_k \) is a contraction and that each \( p_k \) must be a probability.

3.2 The SA method

Simulated annealing is a generalization of the Metropolis Monte Carlo integration algorithm. It includes a temperature schedule for improving the searching efficiency, it become recently very popular as optimization technique. SA has indeed many interesting features, such as the existence of sufficient conditions for the convergence, the ability to process functions with arbitrary degree of nonlinearities, discontinuities and stochasticity and, last but not least, the possibility to derive Simulated Quenching techniques, very powerful for certain classes of problems.

SA is defined by three functional relationship:

* a probability density of the parameters;
* a probability density for acceptance of new cost functions given the previous value;
* a schedule of Temperature annealing.

It is known that he SA method always reaches the global minimum or a good local minimum. On the other hand, it has other drawbacks: for example its convergence is very slow. As a consequence, the SA optimization method takes a very long time to produce good results.

3.3 The proposed algorithm

The general structure of our algorithm is similar to the (SA) one: both the algorithms consist of a loop controlled...
by a C parameter with a function analogous to the
Temperature in the Simulated Annealing process. The
algorithm's basic idea can be described as follows:
starting from a set of randomly chosen \( w_k \) and \( p_k \) (the
parameters of the IFS code), a minimization of the cost
function \( E \) (i.e. our function \( F \)) with respect to one of
the parameters themselves along the gradient direction is
performed. The IFS obtained after this gradient
minimization is stored if it leads to a lower value of the
cost function. At this point the parameters are randomly
perturbed for allowing the cost function to exit from the
local minimum. In this way a new value of the cost
function is obtained. The perturbation is accepted only if
\( (\Delta E < 0) \) OR \( 0 < \Delta E < \text{Threshold} \) \( (11) \)

The perturbations are accepted according to a threshold
related to the control parameter C. This approach is
somewhat similar to the standard acceptance criterion
based on the Boltzmann distribution, because at high
values of the Temperature also big perturbations are
accepted, while at low values only small perturbations
would be accepted. Neverthelees, the deterministic
acceptance criterion aims to reduce the amount of
computational efforts.

The pseudocode of the algorithm is described in Box 1.

```
Initialization;
WHILE ( (C not final) OR (local minimum not good ) ) DO
    BEGIN
        Apply SD from the current state;
        IF (new minimum < local minimum) THEN
            Update the local minimum;
        Update the current state;
        Update the control parameter;
        REPEAT
            Perturb the current state;
            Accept the perturbation according to eq.(11)
        UNTIL ( (new minimum < local minimum) OR
              (new minimum · local minimuml < Thr2) OR
              (number of iterations exceeded) )
        IF (new minimum < local minimum) THEN
            Update local minimum;
        Update the current state;
        Update the control parameter;
END
```

Box 1. Pseudocode of the proposed algorithm

The deterministic acceptance criterion can be analyzed by
considering the time needed to reach a termination of
SA [6]. Because it cannot be shown that the time to reach
a termination of the algorithm is polynomially bounded
for a general NP-hard problem, the conditions for a
stochastic convergence to be polynomial can be analyzed.
Thus, the expected time \( E(S) \) to reach the global
minimum of a function \( f : X \rightarrow R \), where \( X \) is a space
with a finite number \( |X| \) of points (the states) labelled
1,2,...,\( |X| \) is introduced. Rather than working with the
minimisation problem of a function \( f \), in [6] the
following related recognition problem is considered:

```
Given c value \( F \) off, is there a state \( j \) with \( f_j < F \)?
```

According to [6], the depth \( d_j \) of each state \( j \in X \) as the
minimum number of steps from state \( j \) to state \( k \leq j \) with a
transition probability \( \tau_{j,k} > 0 \) can be defined. Then the
depth \( d_F \) of \( F \) is defined as:

\[
d_F = \max(d(j):f_j > F).
\]

A polynomial bound proportional to \( \log |X| \) is only
guaranteed to \( E(S) \) if:

\[
d_F = O(\log n).
\]

where \( n < |X| \). For example, in the traveling salesman
problem with \( n \) cities and \( |X|=(n-1)! \), \( df \leq n-3<10 \) [6].

If we decrement \( F \) we have to consider states with fewer
neighbours giving negative \( df \), and moreover the depth
will be greater because of monotonicity of \( d_F \). The
control value must be smaller to ensure that uphill steps
are usually rejected. Therefore, as we become more
ambitious in our target value \( F \), we are forced to decrease
the control and wait longer. A pragmatic approach to
terminate after a number of steps proportional to \( \log |X| \)
adopted in [6], where a small bound on the depth is
fixed. This method is heuristic even if it performs well,
i.e. good values of \( f \) are reached.

In our situation \( |X|=\infty \), but we adopted the idea of
considering a Threshold limit of neighbours' depth \( df \)
and also a number of steps chosen for the control
parameter to adjust the process's run which behaved
well.

It has to be noted also that in our situation with \( |X|=\infty \),
we can not use any deterministic algorithm that simply
examines all the solutions; this type of algorithm in [6]
is compared with (SA) method. This last, in certain types
of problem, can have a longer time to termination than
the deterministic algorithm. On the contrary we
compared the convergence characteristics of our
algorithm with SA, as shown in Fig.1.

As a matter of fact, the process realizes a constrained
optimization. In fact, all the modifications of the \( w_k \)
are obviously subjected to the constraint that the modified \( w_k \)
must be contractions. Moreover, \( \sum_{k=1}^{N} p_k = 1 \) since
the \( p_k \) are probabilities.

In the next Section it will be shown that successful
results are obtained: good minima have been obtained
with a reduced computational time.

4. Experimental Results and Discussion

To test the algorithm, we first used the Sierpinski
triangle. The results were as follows: the Sierpinsky
triangle with 16 gray levels and a dimension of 32x32
pixel, were encoded with 16 IFS maps. The resulting
error was equal to 1.02% and, considering only the
stochastic part of the algorithm, the proposed
optimization algorithm converged in 7500 iterations with
respect to 175000 needed with the SA method. It should
be noted that the contribution of deterministic part of the algorithm, i.e. the part concerning SD minimization, to the overall computational complexity can be neglected.

![Figure 1. Average convergence curves (percentual error versus number of iterations) of the proposed algorithm (SQ) with the standard SA one.](image)

We then analyzed several graytone and coloured images. The original image (256x256 pixels, 8 bits/pixel) was partitioned into 64 non-overlapping (32x32 pixels) blocks. Each block was separately modeled by means of 16 IFS maps. In Fig.1 the convergence behaviour of the proposed algorithm compared with SA is shown. The 'Lena' image has been used. The data in Fig.1 has been obtained by averaging, at each iteration, the resulting percentual error of each of the 64 sub-blocks. The encoding results are preliminary, since we tested only a limited number of quantization ranges, the focus of this paper being concentrated mainly on the efficient solution of the inverse problem. As said before, each of the sub-image was modeled with 16 IFS. Each IFS is defined by six parameters plus one probability, in total seven parameters. We found that the quantization of the probability parameters is quite critical, and therefore we used 16 bits for them. The remaining parameters were quantized at 16, 14 and 12 bits, with very little difference in reconstruction results. At 16 bit of quantization the image is compressed at 1.75 bit/pixel, and at 12 bit at 1.375 bit/pixel. In Fig.2 the reconstructed 'Lena' image, with 16 bit of quantization for each parameter and a final average error equal to 1.2%, is reported.

![Figure 2. Reconstructed 'LENA' image, encoded at 1.75 bit/pixel and a final error equal to 1.2%.](image)

5. Conclusions
In this paper, the solution of the Inverse Problem has been formulated as an optimization problem. An efficient Simulated Quenching algorithm, based on a combination between Steepest Descent and Simulated Annealing, has been proposed. The initial results are encouraging. In fact a very small error between original and IFS-reconstructed images is obtained with a number of iterations much less than that required by Simulated Annealing.

Acknowledgements
This work was partially supported by ALENIA SpA and by MURST 60%.

References
Trainable Hybrid Filter Structures in Particle Detector Readout Systems

Sami J. INKINEN † and Yrjö NEUVO †

† European Laboratory for Particle Physics, CERN/ECP, CH-1211 Geneva 23, Switzerland
Tel/fax: +41-22-7676350 / 7673600, Email: sinkinen@cernvm.cern.ch
† Nokia Mobile Phones Ltd., P.O. Box 83, FIN-02601 Espoo, Finland
Tel/fax: +358-0-510401 / 5103742

Abstract. In this paper an FIR-order statistic (FIR-OS) hybrid filter is applied to the energy and timing extraction of pulses acquired from high energy physics detectors. The filter rank and the subfilter tap coefficients are obtained using a training approach based on conjugate gradient optimisation. The FIR-OS filter is shown to outperform a single FIR filter when the detector signal is corrupted by electronics and pile-up noise components, sample timing jitter and the superposition of several high-energy pulses. By an efficient use of the FIR subfilters the presented filter structure achieves simultaneously a good time-domain resolution and a high amplitude precision.

1 Introduction

The particle detectors used in high energy physics are multichannel instruments, with which the events generated in accelerators are recorded. By reconstructing the interesting events the particles present in the interactions can be identified and their physics properties measured. The data acquisition systems are usually divided into stages called trigger levels: each level of the processing chain reduces the amount of information by removing irrelevant data. In the first trigger stages the position and the height of the pulses caused by particle hits are computed using 1-D time-domain data. The analog signal obtained from a single detector channel is converted into digital domain using a fast A/D converter, preceded by an analog amplifier having a transfer function $g[.]$. The sampling is synchronized with the collider bunch crossing. However, due to the non-zero length of the particle bunches and the geometry of a detector system the pulse timing relative to the sampling position is slightly different for each pulse. Here this deviation is assumed to be Gaussian-distributed with parameters $G(0,\sigma)$ where $\sigma$ is called the sample timing jitter. The discrete-time signal also contains noise, divided into electronics and pile-up noise. The source of pile-up noise are the minimum bias events, i.e., small energy deposits collected by the detector cells. As the noise is injected into the system prior to the analog amplifier it becomes correlated in the time domain. The Gaussian-like electronics noise component can be effectively attenuated using a matched FIR filter; a discussion of optimal linear filtering can be found in [1].

The linear filters optimized for noise attenuation achieve a low performance in the presence of sample timing jitter and the superposition of high-energy pulses. In this paper an FIR-OS filter is applied to the pulse energy and timing measurement. The filter

$$y(i) = F[x(i-n_a), x(i-n_a+1), \ldots, x(i+n_b)]$$

of length

$$N = n_a + n_b + 1$$

is combined with a three-point maximum finder for pulse timing extraction and an energy threshold $e_t$ for pile-up noise suppression. The energy measurement $y(i)$ directed to the first-level trigger unit is given by

$$z(i) = \begin{cases} y(i) & \text{if } y(i) > e_t \wedge y(i) > y(i \pm 1) \\ 0 & \text{otherwise} \end{cases}$$

(1)

The filter structure is illustrated in Figure 1. Control signals $ctrl0 - ctrl3$ enable different system configurations where the FIR filters are used either independently or as the subfilters of the FIR-OS operator.

2 FIR-OS filter training

A hybrid structure of linear and nonlinear operators, a general order statistic operator on the topmost level using the outputs of linear FIR subfilters as input, is called the FIR-OS filter [2]. The FIR subfilters form a bank of filters from which the order statistic operator selects one in each input sequence point $z(i)$. The filter is given by

$$y(i) = OS[\phi_1(i), \phi_2(i), \ldots, \phi_m(i)]$$

(2)

where $r$ is any rank between 1 and $m$ and $\phi_k$’s are the FIR subfilters, with the tap coefficient vector $h^k = \{h^k_1, h^k_2, \ldots, h^k_r\}$, given by

$$\phi_k(i) = h^k_1 x(i-n_a) + h^k_2 x(i-n_a+1) + \ldots + h^k_r x(i+n_b)$$

(3)

where $N = n_a + n_b + 1$. The filter structure can be optimized using a training sequence to find the rank $r$ and the FIR filter tap coefficient vectors $h^k$ which give the minimum total error between the filter output and the desired output. For this purpose we define the cost function using the mean squared error (MSE) norm to be

$$E = \frac{1}{N_q} \sum_{i=1}^{N_q} E^2 = \frac{1}{N_q} \sum_{i=1}^{N_q} (y(i) - \hat{y}(i))^2$$

(4)

where $\hat{y}(i)$’s are the desired filter outputs and $N_q$ is the number of samples in the training set $Q$. By minimizing the cost function $E$ the optimal filter for set $Q$ is
obtained in the MSE norm sense. In the training approach the optimization is based on the minimization of the total cost function over the whole training set, while in adaptive filtering the coefficients are adjusted according to the error computed for individual samples. An adaptation algorithm for hybrid order statistic filters has been presented in [3]. Weighted order statistic (WOS) and FIR-WOS filters have been optimized using the mean absolute error (MAE) criterion [6, 7]. All these algorithms are modifications of the standard least mean square (LMS) optimization of linear FIR filters [5]. The iterative optimization process presented here is based on the division of the original training set into subsets using the order statistic operator. This algorithm is considered especially powerful in cases where the training set contains several separable components, e.g., single and partially overlapping pulses. To allow the FIR subfilters to specialize on different parts of the sequence during the optimization, the hard-limiting order statistic operator is replaced by a soft order statistic operator. If the \( r \)-th biggest FIR filter output is denoted by \( \phi_{(r)}(i) \), the soft order statistic operator for rank \( r \) is given by

\[
y(i) = w_a \phi_{(r)}(i) + \sum_{k=0}^{m} w_d \phi_{(k)}(i)
\]

where \( w_a \) and \( w_d \) are the acceptance weight and rejection weight, respectively. As the weights \( w_a \to 1 \) and \( w_d \to 0 \) asymptotically the soft order statistic operator approaches the hard order statistic operator (2). The optimization algorithm is shown in Table 1. It should be noted that due to the system nonlinearities the resulting filters are not guaranteed to be globally optimal. However, by performing several optimization runs the best of the local minima can be considered as a good estimate of the global minimum.

3 Pulse amplitude extraction

To evaluate the filter performance two sets of simulated detector signals were created. Set \( A' \) contains single pulses with a bipolar shaping function \( g[i] \) and 15 ns sampling interval. The pulses are corrupted by sample timing jitter \( \sigma_p = 2 \) ns and the electronics and pile-up noise components. In addition to these artifacts set \( B' \) includes overlapping pulses. If the pulse to be measured originates from event \( i \), the energy deposit in event \( j \) causing the pulse superposition is approximated using a logarithmic energy distribution function

\[
e(j) = \begin{cases} 
  e_t \ln[P_m/P] & \text{if } P < P_m \wedge |i-j| \geq d \\
  0 & \text{otherwise}
\end{cases}
\]

where \( e_t = 1 \) GeV is an energy scaling term, \( P_m = 0.03 \) is the maximum probability of having any significant energy deposit in a single event and \( d = 2 \) is the minimum distance for the two pulses to be separable. The probability value \( P \) is uniformly distributed between zero and one. The pulse distance criterion \( |i-j| \geq d \) is included due to the use of the three-point maximum finder. A five-tap FIR filter and an FIR-OS filter with two five-tap subfilters were trained using sets \( A' \) and \( B' \). Pulses of 2, 20 and 200 GeV were used in the training, 500 pulses per energy, and the obtained filter tap coefficients are shown in Table 2.

The normalized mean squared error MSE[y/y] values in the pulse amplitude are shown in Figures 2 and 5. The evaluation sets \( A \) and \( B \) contain 0.5–2000 GeV pulses, 2000 pulses c; each energy. The FIR filter trained with set \( A' \) performs well for the energy region above 20 GeV where the sample timing jitter is the dominant source of error. The FIR filter trained with set \( B' \) is slightly better for the low-energy region. Clearly the FIR-OS filter is an efficient combination of these two operators giving a precise amplitude estimate over the whole energy range. It also obtains a similar performance for set \( B \) whereas the performance of the FIR filters starts to degrade. Note that the pulse amplitude gain of the FIR-OS subfilter 2 is about 3% below unity. The pulse gain value is obtained by filtering single pulses without electronics and pile-up noise. As the FIR-OS rank \( r = 2 \) the subfilter giving the larger output value is selected in each sample position. For this reason subfilter 2 is used only when subfilter 1 extracts an energy estimate which is too small, e.g., when a superposition of several pulses occurs. During the training procedure the systematic error of subfilter 2 is adjusted so that subfilter 1 is used for non-overlapping pulses.
Table 1: FIR-OS training algorithm

1. Assign random weight vectors $h^k_1, h^k_2, \ldots, h^k_m$ to subfilters $\phi_1, \phi_2, \ldots, \phi_m$. Set the acceptance weight $w_a$ and the rejection weight $w_d$ to $1/m$.

2. Divide the original training set $Q$ into subsets $Q_1, Q_2, \ldots, Q_m$ so that each set contains all the points $x(i)$ where the corresponding FIR subfilter is selected by the order statistic operator. Note that in this phase some of the sets may be empty sets.

3. Optimize each FIR subfilter inside the training set $Q$ so that the cost function value $(y(i) - \hat{y}(i))^2$ is multiplied by either the acceptance weight or the rejection weight according to the following criterion:

$$E^k_i = \begin{cases} w_a(y(i) - \hat{y}(i))^2 & \text{if } x(i) \in Q_k \\ w_d(y(i) - \hat{y}(i))^2 & \text{otherwise} \end{cases}$$  \hspace{1cm} (6)

The optimization method is the conjugate gradient search [4]. As the weighted cost function value is given by

$$E^k_i = \sum_{j=1}^{N} w_k (h^k_j x(i + j - n_a - 1) - \hat{y}(i))^2$$  \hspace{1cm} (7)

where $w_k$ equals to either $w_a$ or $w_d$, the partial derivatives needed in the minimization are obtained directly from (7):

$$\frac{\partial E^k_i}{\partial h^k_l} = \sum_{j=1}^{N} 2w_k x(i + l - n_a - 1)(h^k_j x(i + j - n_a - 1) - \hat{y}(i))^2$$  \hspace{1cm} (8)

where $l = 1, 2, \ldots, N$. The direction of adjustment and the step length are defined by the partial derivatives, and by iterative adjustment of the FIR coefficients the local minimum of the cost function is found.

4. Update sets $Q_1, Q_2, \ldots, Q_m$ using the new filter weight vectors obtained in the previous step. Update the acceptance weight and the rejection weight so that $w_a(j + 1) = 1 - (m - 1)Cw_d(j)$ and $w_d(j + 1) = Cw_d(j)$ where $C$ is a constant dependent on the selected number of iterations. The application tests suggest that good results are obtained with 100 iterations and $C = 0.9$.

5. Repeat steps 3 and 4 until the weights $w_a$ and $w_d$ are close to the limiting values 1 and 0, respectively, and the equation (5) corresponds to the hard order statistic operator (2).

6. Repeat steps 1–5 going through all the rank values $1, 2, \ldots, m$. Select the rank and the corresponding FIR subfilter weight vectors which give the smallest total cost $E$. As the obtained filter parameters correspond to a local minimum perform a few optimization runs to check whether the result changes substantially.

The trained filters were also applied to a complete signal sequence containing 5 GeV pulses and corrupted by all the artifacts. Part of the signal is shown in Figure 3. The FIR filter trained with set $A'$ has a relatively wide response on the pulses: its output is less affected by the sample timing jitter than the output of the FIR filter trained with set $B'$ (Figures 4 and 6). However, it is unable to produce a correct output for overlapping pulses. The FIR-OS filter uses different subfilters for single and overlapping pulses and achieves simultaneously a good time-domain resolution and a high amplitude precision (Figure 7). In all three cases the background noise level is almost identical.

4 Conclusions
An FIR-OS filter and a training algorithm for high energy physics detector readout systems was presented. To increase algorithm stability and efficiency in data subset determination, the hard-limiting order statistic operator is replaced by a soft order statistic operator during the training process. The FIR-OS filter performs especially well when the event occupancy is high, i.e., when there is a substantial probability of having overlapping pulses. Compared to FIR filters, optimized to suppress the effect of certain artifacts, it needs to compromise less between the time resolution and the amplitude precision. The hybrid structure is more complex to implement than a single FIR filter but the extra hardware is justified when a high performance is required.

References
Table 2: Filter parameters (for FIR-OS, rank $r = 2$)

<table>
<thead>
<tr>
<th>Filter</th>
<th>Pulse gain</th>
<th>Tap coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIR A</td>
<td>1.0236</td>
<td>-0.4331 -0.8861</td>
</tr>
<tr>
<td>FIR B</td>
<td>1.0136</td>
<td>-0.3738 -0.2183</td>
</tr>
<tr>
<td>FIR-OS 1</td>
<td>1.0070</td>
<td>-0.1193 -0.7853</td>
</tr>
<tr>
<td>FIR-OS 2</td>
<td>0.9707</td>
<td>-0.0052 0.0237</td>
</tr>
</tbody>
</table>

**Figure 2:** Normalized amplitude error, set A

**Figure 5:** Normalized amplitude error, set B

**Figure 3:** Simulated signal sequence

**Figure 6:** Response of the FIR B filter

**Figure 4:** Response of the FIR A filter

**Figure 7:** Response of the FIR-OS 2 filter
Skeleton Redundancy Reduction
Based on a Generalization of Convexity

Renato KRESCH and David MALAH

Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel, Tel/Fax: +972 (4) 204 745 / 323 041, E-Mail: malah@ee.technion.ac.il

Abstract. We present a generalization of the concept of Convex Sets, based on the Morphological Closing operation, and study some of its properties. We also define Extreme Points of such Generalized Convex Sets, which generalize the notion of Extreme Points of Convex Sets.

Moreover, we apply the above notions to skeleton redundancy removal, and present an algorithm for obtaining an Error-Free Skeleton representation with reduced amount of redundant points, using morphological operations only.

1. Introduction

The concept of Convexity is of great interest in several areas, such as Shape Analysis, Pattern Recognition, Image Decomposition, and others. Many properties and relations concerning Convex Sets have been extensively studied and analyzed, and a number of generalizations of Convexity were proposed before (see [1] for example), in order to extend some of these properties and relations to sets which are not strictly convex.

The latter is the purpose of this work too.

The generalization proposed here is based on the Morphological Closing (one of the four Morphological basic operations). The proximity between the Convex-Hull operator and the Closing operator has already been pointed out in [2], but a close look at the structure of their definitions shows that there is more than a proximity; the Convex-Hull is actually a particular case of Closing. Therefore, some important properties of the former operator are naturally extended for the latter one.

In section 2, we remind the definition and some properties of Convex Sets and present the proposed generalization, based on the Closing operator.

In Convex Sets, one can find Extreme Points, which have several properties. Among those properties, there is the ability of fully representing the whole set (if it is bounded), i.e., one can recover a bounded Convex Set from its Extreme Points only (this is done by simply applying the Convex-Hull operator). In section 3, we define Extreme Points for the previously generalized Convex Sets, and present Morphological Formula for calculating them. We also study conditions for perfect reconstruction of the generalized Convex Sets from their Extreme Points.

Section 4 presents an application of the proposed generalization for reducing the redundancy in Skeleton representations. It consists of an entirely morphological algorithm for removing most of the redundant points of a given Skeleton.

Throughout the paper, we use the following notation: For \( A \) and \( B \) sets in \( \mathbb{R}^2 \), \( A \oplus B \), \( A \ominus B \) and \( A \ast B \) are, respectively, the Morphological Dilation, Erosion and Closing of \( A \) by \( B \);

\[ B^c = \{ b \in \mathbb{R}^2 \mid b \notin B \} \] is the complement of \( B \), \( B^t = \{ -b \mid b \in B \} \) is the transposed set of \( B \), and \( B_z = \{ b + z \mid b \in B \} \) is the translation of \( B \) to the point \( z \in \mathbb{R}^2 \).

2. Convex Sets and Proposed Generalization

There are several acceptable definitions for Convex-Hull and Convex Sets. They are all equivalent, up to topological differences concerning the points on the boundary of the shapes. We can also define first the Convex-Hull and then use this definition for defining Convex Sets, or we can do the opposite.

The definitions of the Convex-Hull and Convex sets we choose to work with are the following:

- **Convex-Hull**: \( CH(X) \) is the Convex-Hull of a set \( X \) iff it is the intersection of all the half-planes that contain \( X \).

- **Convex set**: A set \( X \) is Convex iff it is identical to its Convex-Hull, i.e., \( X = CH(X) \).

The generalization we propose is obtained by replacing the half-plane used in the above definition of the Convex-Hull by a generic set \( (B^t)^c \), which is the transposed of the complement of any structuring-element \( B \). We denote the generalized Convex sets as B-Convex sets and the generalized Convex-Hull as B-Convex-Hull because of the dependence on the structuring-element \( B \):

- **B-Convex-Hull**: \( CH_B(X) \) is the B-Convex-Hull of \( X \) iff it is the intersection of all the translations of \( (B^t)^c \) that contain \( X \).

- **B-Convex set**: A set \( X \) is B-Convex iff it is identical to its B-Convex-Hull, i.e., \( X = CH_B(X) \).

Actually, the B-Convex-Hull, as defined above, is not a new operation; it is known in Mathematical Morphology as
the Morphological Closing. In other words:

$$CH^B(X) = X \ast B.$$  \hspace{1cm} (1)

If we choose \(B\) to be a disc, and make its radius go to infinity, then the above Closing converges to the conventional Convex-Hull (as pointed out in [2, p. 100]), meaning that the conventional Convex-Hull is indeed a particular case of the generalized Convex-Hull.

Table 1 shows that some of the basic properties of the Convex-Hull and of Convex sets are naturally extended to the \(B\)-Convex-Hull operation and to \(B\)-Convex sets.

<table>
<thead>
<tr>
<th>Known Property</th>
<th>Property of the Proposed Generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>(CH(\cdot)) is idempotent.</td>
<td>(CH^B(\cdot)) is idempotent.</td>
</tr>
<tr>
<td>(CH(X)) is the &quot;smallest&quot;</td>
<td>(CH^B(X)) is the &quot;smallest&quot;</td>
</tr>
<tr>
<td>convex set that contains (X).</td>
<td>(B)-Convex set that contains (X).</td>
</tr>
<tr>
<td>(X) is convex iff any two</td>
<td>If (X) is (B)-</td>
</tr>
<tr>
<td>points (x) and (y) belonging to</td>
<td>Convex, then (\forall {x, y} \subseteq X,</td>
</tr>
<tr>
<td>(X) are connected by a segment</td>
<td>(CH^B({x, y}) \subseteq X.)</td>
</tr>
<tr>
<td>contained in (X). In other words: (X) is convex iff (\forall {x, y} \subseteq X, ) (CH^B({x, y}) \subseteq X.)</td>
<td></td>
</tr>
<tr>
<td>The intersection of convex sets is a</td>
<td>The intersection of (B)-Convex sets is a (B)-Convex set.</td>
</tr>
<tr>
<td>convex set.</td>
<td></td>
</tr>
<tr>
<td>(X) is convex iff every point</td>
<td>(X) is (B)-Convex iff every</td>
</tr>
<tr>
<td>outside (X) can be separated</td>
<td>point outside (X) can be</td>
</tr>
<tr>
<td>from (X) by a half-plane, i.e.,</td>
<td>separated from (X) by a</td>
</tr>
<tr>
<td>(x \notin X \Rightarrow \exists )</td>
<td>translation of (B^*), i.e., (x \notin X \Rightarrow \exists )</td>
</tr>
<tr>
<td>a half-plane that contains (x) and</td>
<td>((B^*)^n), contains (x) and does</td>
</tr>
<tr>
<td>does not intersect (X).</td>
<td>not intersect (X).</td>
</tr>
</tbody>
</table>

Table 1: Properties of Convex-Hull and Convex sets.

3. Extreme Points

3.1 Definition and Calculation

Like the Convex-Hull and Convex Sets, there are many ways to define Extreme Points of a Convex Set. Table 2 shows one of the classical definitions of Extreme Points for conventional Convex sets, and presents its generalization for \(B\)-Convex sets. We denote the set of Extreme Points of a given Convex Set \(Y\) by \(\mathcal{E}(Y)\) and the set of Extreme Points of a given \(B\)-Convex Set \(X\) by \(\mathcal{E}^B(X)\).

<table>
<thead>
<tr>
<th>Extreme Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convex sets</td>
</tr>
<tr>
<td>A point (t) is an Extreme Point of a Convex set (X) if (t \notin X) is also convex.</td>
</tr>
<tr>
<td>(B)-Convex sets</td>
</tr>
<tr>
<td>A point (t) is an Extreme Point of a (B)-Convex set (X) if (t \notin X) is also (B)-Convex.</td>
</tr>
</tbody>
</table>

Table 2: Extreme Points

The following Morphological closed-form formulae provide two ways of calculating the set of Extreme Points of a given \(B\)-Convex Set \(X\):

$$\mathcal{E}^B(X) = X - \bigcap_{y \in X} (X - \{y\}) \oplus B.\hspace{1cm} (2)$$

$$\mathcal{E}^B(X) = X - \bigcap_{y \in B} (B - \{y\}) \oplus B.\hspace{1cm} (3)$$

The outline of the proofs of (2) and (3) are given in appendices A and B, respectively.

If we consider the computational efficiency of the above equations, when implemented on a computer, then (2) is preferable over (3) if \(X\) contains fewer elements than \(B\), and (3) is preferable over (2) otherwise.

3.2 Reconstruction from Extreme Points

If a conventional Convex set \(Y\) is bounded, then it can be reconstructed back from its Extreme Points by performing the Convex-Hull operation, i.e., \(CH(\mathcal{E}(Y)) = Y\). The set of Extreme Points can be seen as a compact representation of a Convex set.

For a \(B\)-Convex Set \(X\), a necessary condition for perfect reconstruction from its set of Extreme Points \(\mathcal{E}^B(X)\) is: \(X = \mathcal{E}^B(X) \oplus B \neq \emptyset\). This suggests that \(X\) should be "smaller" (in a certain way) than \(B\). Notice that the erosion of any bounded shape by a half-plane is always empty.

The above considerations motivate the definition of a Reconstruction Window for a given structuring-element \(B\), inside which every \(B\)-Convex Set can be reconstructed from its Extreme Points. A \(B\)-Convex Set \(W\) is called a Reconstruction Window for \(B\) iff \(\forall X \in B\)-Convex, \(\mathcal{E}^B(X \cap W) = X \cap W\).

For example, if \(B\) is a rectangle, then \(B\) itself is a Reconstruction Window for \(B\). If \(B\) is a discrete rectangle of integer sides \(n\) and \(m\), then any discrete rectangle of sides \(i\) and \(j\), such that \(0 \leq i \leq (n+1)\) and \(0 \leq j \leq (m+1)\), is a Reconstruction Window for \(B\).

4. Application: Skeleton Redundancy Reduction

It is well-known that the Skeleton representation of images usually contains redundant points, i.e., some of its points may be discarded and still a perfect reconstruction can be obtained [3].

In [3], Maragos and Schafer introduced the concept of Minimal Skeleton, which is defined as any subset of the Skeleton containing no redundant points, from which perfect reconstruction of the original image is possible. They also presented an efficient algorithm for obtaining a Minimal Skeleton, from a given Skeleton. However, this algorithm is not fully morphological, and therefore cannot be implemented on a parallel morphological machine. Fully morphological methods for reducing the Skeleton redundancy are studied in [4] and [5].

In the sequel we present an algorithm for morphologically obtaining a redundancy-reduced skeleton, based on the \(B\)-Convexity theory discussed above.
4.1 The Algorithm

The algorithm is presented below, together with an example. Figure 1 shows the steps of the algorithm for the example.

Fig. 1: Proposed algorithm. (a) A discrete binary shape (black dots: foreground, white dots: background), (b) $S_1$, (c) $Z_1$, and the partition blocks, (d) Extreme Points of the blocks, (e) $\hat{S}_1$, and (f) resulting reduced skeleton (black points) compared to the original skeleton (black and grey points).

1. Let $X$ be a given binary image. Choose a structuring-element $B$, and a family of Reconstruction Windows $\{W(nB)\}$ for all the dilations $nB$ of $B$. (In the example, $X$ is the digital binary shape shown in Fig. 1(a) [described by the black dots], $B$ is a $3 \times 3$ square, and $W(nB)$ are $(2n+2) \times (2n+2)$ squares. Set $n = 0$.

2. Calculate the skeleton subset $S_n \triangleq X \ominus nB - (X \ominus nB) \circ B$, and the set $Z_n \triangleq X \ominus nB$. If $Z_n$ is empty then stop. (In the example, for $n = 1$, $S_1$ is shown in Fig. 1(b) and $Z_1$ is seen in Fig. 1(c)).

3. Obtain a partition of $Z_n$ into blocks $Y^n_p$ such that: $Y^n_p$ is the contents of $Z_n$ inside the Reconstruction Window $W(nB)$ centered at $p$, i.e., $Y^n_p \equiv (W(nB))_p \cap Z_n$, and the blocks cover the whole set $Z_n$, i.e., $\bigcup_p Y^n_p = Z_n$. (In the example, the blocks $Y^n_p$ were obtained by translating the Reconstruction Window horizontally and vertically by steps of $p = 2n + 1$ pixels, so that there is a 1-pixel-wise overlapping between the blocks. The overlapping by one pixel contributes for the redundancy reduction. Fig. 1(c) shows the first block in grey, and the thin solid lines indicate the position of the other blocks.)

4. Calculate the Extreme Points of every block $Y^n_p$, according to $nB$, $E^{nB}(Y^n_p)$. Note that $Y^n_p$ is a $(nB)$-Convex set, since it is the intersection of two $(nB)$-Convex sets. (Fig. 1(d) shows the result of this operation in the example).

5. Define $C_n \triangleq \bigcup_p E^{nB}(Y^n_p)$ to be the set of the resulting Extreme Points of all the blocks, and intersect it with the skeleton subset $S_n$, obtaining $\hat{S}_n = C_n \cap S_n$ (Fig. 1(e) shows $\hat{S}_1$).

6. Increment $n$, and go to 2.

The collection of sets $\{\hat{S}_n\}$ is the Redundancy-Reduced Skeleton. For comparison between $\{\hat{S}_n\}$ and the original skeleton $\{S_n\}$, in the scope of the above example, Fig. 1(f) shows the reduced skeleton composed of black dots, and the original skeleton, composed by both the black and grey dots. The grey dots are redundant points removed by the above algorithm.

Exactly as for the conventional Skeleton, the following relation holds:

$$\bigcup_{n \geq k} \hat{S}_n \ominus nB = X \ominus kB$$

which guarantees partial ($k > 0$) and perfect ($k = 0$) reconstruction of the original image.

4.2 Simulation

Figure 2(a) shows a binary image (Most-significant bit-plane of 256 x 255-pixel "House"), and its morphological skeleton, calculated with a $3 \times 3$ squared structuring-element. The skeleton contains 3173 points. Fig. 2(b) shows the result of applying the above algorithm to the same binary image. The structuring-element and the Reconstruction Windows are the same as in the example of Fig. 1. The resulting skeleton fully represents the original binary image, and contains 1533 points, i.e., only 48% of the points in the original skeleton.
For comparison, a Minimal Skeleton of the above image, using the non-morphological algorithm given in [9], was calculated. It contains 1362 points, i.e., 45% of the points in the original skeleton, and 89% of the number of points in the proposed reduced skeleton. According the above numbers, the proposed skeleton was able to remove 91% of the redundant points in the original skeleton.

5. Conclusion

A generalization of Convexity is presented, where some of the properties of Convex Sets are extended to sets which are not convex in the traditional sense. Extreme Points of the generalized convex sets are defined, and their ability to fully represent the original set is considered.

Furthermore, an algorithm, based on the above notions, is proposed for morphologically reducing the amount of redundant points in the skeleton. Simulation results indicate that most of the redundancy in the skeleton is removed by the proposed algorithm, which is fully morphological.

The proposed approach is also suitable for morphological calculation of the set of Essential Points of the Skeleton, which is the set of points none of which can be removed from the original skeleton if a perfect reconstruction is desired [6].

Acknowledgment

This work was supported by the Fund for the Promotion of Research at the Technion.

References


Appendix A

Proof of equation (2)
The set \((X - \{x\}) \bullet B\), for \(x \in X\), is a \(B\)-Convex set, can be equal either to \((X - \{x\})\) or to \(X\). This is because:
\[
X - \{x\} \subseteq (X - \{x\}) \bullet B \subseteq \subseteq X \bullet B = X
\] (A.1)

By definition of Extreme Points, \((X - \{x\}) \bullet B\) is equal to \((X - \{x\})\) iff \(x\) is an Extreme Point. Otherwise it is equal to \(X\). Therefore:
\[
X - \left[ \bigcap_{x \in X} (X - \{x\}) \bullet B \right] \subseteq B^c = X - \left[ \bigcap_{x \in E_B(X)} (X - \{x\}) \right] = \bigcup_{x \in E_B(X)} \{x\} = E_B^c(X)
\] (A.2)

Appendix B

Outline of the proof of equation (3)
It is enough to prove that for any sets \(A\) and \(B\):
\[
\bigcap_{a \in A} (A - \{a\}) \oplus B = \bigcap_{b \in B} A \oplus (B - \{b\})
\] (B.1)

First, let us denote the left hand of the above equation as \(\mathcal{H}\), and then write the dilation explicitly in the following way:
\[
\mathcal{H} = \bigcap_{a \in A} \bigcup_{b \in B} \bigcup_{a \neq \hat{a}, \bar{b} \neq \hat{b}} \{a + b\}
\] (B.2)

Then, after some logical and set manipulations, we notice that a point \(z = \hat{a} + \hat{b}\) belongs to \(\mathcal{H}\) if there is another pair of points \(a\) and \(b\) in \(A\) and \(B\) respectively, such that \(a + b = z\). In other words:
\[
\mathcal{H} = \{z = \hat{a} + \hat{b} = a + b \in A \oplus B \mid a \neq \hat{a}, b \neq \hat{b}\}
\] (B.3)

Since equation (B.3) is symmetric, i.e., the roles of \(A, a, \hat{a}\) and \(B, b, \hat{b}\) are respectively interchangeable, then we can interchange the above sets and elements also in the original expression, which provides (B.1).
Improvements to the 'Top Hat' transform, used for analysing pigmented patches on flower petals.

Stephen J. IMPEY and J. Andrew BANGHAM.

School of Information Systems, University of East Anglia, Norwich NR4 7TJ, UK.
Tel/Fax. +44 (0) 603 592245 / 507720, Email: sji@sys.uea.ac.uk, ab@sys.uea.ac.uk

Abstract. It is frequently necessary to preprocess images to remove the background. In this paper a controlled study of background removal by different types of filter is presented. The bandpass characteristics of each type of filter are optimised, and it is shown that the morphological top-hat transform as usually implemented [2,3] is very sensitive to noise but that replacing the opening (or closing) step by a median sieve (a cascade of median filters with structuring elements of increasing size [5,6]) removes the background in a much more robust fashion.

1. Introduction

In this paper mathematical morphology and median based methods are used to analyse the patterns of pigmented cells in flower petals. It has been shown that the patterns on one particular variety of Antirrhinum-major (Snapdragon) can be controlled by the temperature at which the flowers are grown [1]. This is of huge importance because it allows the contribution of cells from different parts of the petal to the overall shape of the flower to be followed. The results will provide an insight into the development of the morphology of other living things. Individual petals are dissected out, pressed flat and scanned (using a 35 mm slide scanner) into the computer as red, green and blue images, each with 8 bit precision, Fig. 1a shows an example.

Analysis follows the overall strategy outlined by [2] in which the background is subtracted by 'highpass' (extended to 'bandpass') filtering. Then the signal is thresholded and the dark pigmented patches located and analysed by a modified flood-fill algorithm. The attribute vector associated with each patch includes measures of amplitude, area and centroid. To ensure reliable estimates of the attributes it is important to remove the background causing as little distortion to the pigmented patches as possible.

2. Methods

Morphological 'Top-Hat' transformation [3] is defined as:

\[ t_m(f) = f - \psi_m(f) \]  
\[ t_m(f) = \varphi_m(f) - f \]

where \( f \) is the initial image and \( \psi(f) \) (resp. \( \gamma(f) \)) is a morphological closing (resp. opening). The choice of operation is data dependent. The opening and closing are defined by:

\[ \gamma_m(f) = \delta(e(f)) \]  
\[ \varphi_m(f) = \varepsilon(\delta(f)) \]

where \( \delta(f) \) is a dilation and \( e(f) \) is an erosion [4] with a flat structuring element, of size referenced by \( m \). The smallest structuring element (\( m=1 \)) is a 2 by 2 square, the next (\( m=2 \)) a 3 by 3, larger elements are of bigger radius and are increasingly circular (\( m=5, \ldots \)).

In the same way, a median sieve 'Top-Hat' is defined in terms of a lowpass filter formed from a cascade of median filters. These have been shown to have similar properties to M and N filters, but be more robust [5,6]. The sequence of median filtering operations begins with 3 by 3 centre weighted medians. First centre weight 7, second 5, then 3 and finally uniform weights of 1. Larger elements are of bigger radius and are increasingly circular (\( m=5, \ldots \)).

\[ t_m(f) = \mu_m(\ldots \mu_3(\mu_1(f)) \ldots, ) \]

The filter at a single stage \( m \) of the cascade is defined as:

\[ \mu_m(A_{\ell k}\{x\}) = \text{Med}( A_{\ell k+j=1}; (k,l) \in B_m) \]

where \( B_m \) defines the points in mask \( m \). The output of the \( n \)-th stage of the cascade is defined as:

\[ S_n(f) = \mu_n(\ldots \mu_3(\mu_1(f)) \ldots, ) \]

The median sieve 'Top-Hat' is defined as:

\[ T_n(f) = f - S_n(f) \]

The extended 'Top-Hat' [7] has a similar response to the standard morphological 'Top-Hat', but also removes density fluctuations, that is noise. This is defined by:

\[ t(f) = f \cdot \chi(\varphi(f))\chi(f) \]  
\[ \tau(f) = \varphi(\chi(f))\varphi(f) \cdot f \]

The above 'Top-Hat' transforms have the effect of removing the background, and the result can be thought of as details which are smaller than the mask. It is a highpass filter that preserves details and random noise present in the original image. The highpass filters can therefore be extended to bandpass by filtering the 'Top-Hat' output by a lowpass filter, characterised by \( n \). This has the effect of removing small details and noise. For
the morphological filters the bandpass filter becomes a opening operating on the closing (resp. opening) 'Top-Hat':
\[ b_{mn} = \gamma_l(t_m(f)) \quad \text{or} \quad b_{mn} = \gamma_h(t_m(f)) \]  
(n=cm)

Since the median sieve bandpass 'Top-Hat' is defined as a cascade of filters, the bandpass version is defined by:
\[ B_{m,n}(f) = S_{m}(f) - S_n(f) \]  
(n=cm)

The final type of filter to be considered is a linear bandpass convolution filter
\[ Y_{mn}(f) = (f * m) * h_n \]  
(n=cm)

where the * operator defines convolution of the signal with \( h_m \), a square mask of size \( m \) containing circularly symmetrical Gaussian coefficients with standard deviation of 0.5 \* \( m \).

Having bandpass filtered the image it is then thresholded \( t \) to produce a binary image in which the dark patches are counted. The three parameters, \( m, n \), and \( t \), are chosen, with the assistance of a standard mininising routine (dpmin [8]) and a model. The error metric, between the processed test image and reference, is calculated by summing the distances from each patch in test image to the nearest equivalent patch in the reference to the difference in areas and the difference in the actual number of patches counted.

3. Results
Fig. 1a shows a typical petal (2550x1650 pixels), the pigmented patches are clearly visible. Fig. 1b shows another petal in which the pigmented cells are missing. It is found that the range of pixel values observed in the natural pigmented patches overlap those in the rest of the petal. Fig. 2a shows a detail of a pigmented patch, and illustrates the range of values in the petal, notice the white 'salt' pixel. Consequently, simple thresholding cannot separate the two and a bandpass filter is required.

Fig. 2b shows a model patch, diameter 15 pixels that has been added to the blank petal. Model patches are added at random positions with a probability density, \( p \) (probability of placing a patch at each location \( \leq 0.15 \)). Fig. 4a shows an image containing patches \( (p=0.04) \). There is a chance that overlapping patches will occur, producing larger patches. Fig. 3a shows the relation between the number counted and the number added when analysing clean images as shown in Fig. 4a, and the recovery rate for patches reduces as the number added increases.

Fig. 4b shows a 500x400 pixel portion of the blank petal, Fig. 4c shows the result of adding a set of model patches (Fig. 4a) and Fig. 4d shows the result of adding further salt and pepper (10%) noise. Fig. 5 shows the result of bandpass filtering Fig. 4c and 4d followed by thresholding (using best possible parameters, \( m, n, t \)). The error signal, in each case, is obtained by comparing these images with the reference Fig. 4a.

Fig. 5a shows the result of filtering with a morphological 'Top-Hat' \((n=0, m=41, t=78)\), the output patches are very fragmented and most of the detail is removed if \( n>0 \) and this is not desirable. Fig. 5b is the result of filtering with the extended 'Top-Hat' \((m=51, t=16)\). The patches are much less fragmented. Fig. 5c shows the result of filtering with a median sieve 'Top-Hat' \((n=8, m=30, t=35)\), this result is almost identical to the reference. Likewise the linear bandpass filter \((n=8, m=45, t=30)\). Fig 5d. The experiments shown in Fig. 5c and 5d are repeated three times at each of a number of patch densities and the error rate associated with the median sieve and linear bandpass filters are shown in Fig. 3b and 3c. Fig. 5aii and 5bii show that in the presence of noise both the morphological based filters completely fail to preserve any of the patches. The median sieve based filter output (Fig 5cii), is virtually unaffected by the noise, and the linear bandpass filter allows some noise to pass (Fig. 5di). The median sieves is clearly more robust.

Fig. 6a shows a 1020x550 pixel portion taken from a complete petel image, this contains 243 patches when counted by eye. Fig. 6b shows the result of filtering with a median 'Top-Hat' with the above stated coefficients, this gives a patch count of 253. Fig. 6c shows the same area filtered with the stated coefficients for a morphological 'Top-Hat', it is observed that many patches are either too small or completely missing. Reducing the threshold to \( t=64 \) produces Fig. 6d, more of the patches appear together with noise. Any noise increases the patch count, area thresholds count be used to reject this noise.

Fig. 6e shows the result of filtering with an extended 'Top-Hat', the patches are of the correct size, but spurious small scale noise increases the count to 327.

4. Conclusions
The standard 'Top-Hat' transform is applied to a biological problem. It does not, in some ways, perform as well as a standard linear bandpass filter. A bandpass filter based on the median sieve is shown to be more robust in the face of impulsive and natural noise and appears to be the preprocessor of choice.

Acknowledgements
Thanks to E. Coen, John Innes Institute, Norwich, UK, for supplying the flower petals.

References
Fig 1a
Petal with Patches.

Fig 1b
Blank petal.

Fig 2
30 x 30 pixel area of a petal showing:
a) natural pigmented patch, b) Modelled patch.

Fig 2a
Best Count - No Background

Fig 2b
Error in Median TopHat Count

Fig 2c
Error in Linear Bandpass Count

Fig 3a

Fig 3b

Fig 3c

Fig 3d

Fig 4a. Artificial patches

Fig 4b. Blank petal

Fig 4c. a)+b) petal

Fig 4d. c)+10% noise
Fig 5: Filtered with: i) no noise, ii) noise added, a) Morph. Top-Hat, b) Extended Top-Hat, c) Median Top-Hat d) Linear Bandpass.

Fig 6
a) Original Area Of Petal.
b) Median Top-Hat,
c) Morph. Top-Hat (t=78),
d) Morph. Top-Hat (t=64),
e) Extended Top-Hat
2D Pattern Recognition using 1D Sieves based on alternating sequential, root median and self-dual filters

C. Jeremy PYE and J. Andrew BANGHAM

School Of Information Systems, University Of East Anglia, Norwich, NR4 7TJ, U.K.,
Tel/Fax: +44 (0603) 56161, E-mail: cjpy@sys.uea.ac.uk, ab@sys.uea.ac.uk.

Abstract. It is shown that simple two dimensional (2D) pattern matching based on the 1D non-linear decomposition of images is, under certain circumstances, more effective than one based on linear decomposition. It appears that in the absence of uncorrelated noise, mathematical morphology based sieves (alternating sequential and self-dual filters) that yield decompositions that are invertible, perform better than median based sieves. However, in the presence of noise the latter appears to be more robust. The results from 1D sieves appear to be better than from 2D sieves.

1. Introduction

Pattern recognition in two dimensional (2D) images is a long standing problem that has been addressed in many ways. In many the first step is to first perform a multiscale (multiresolution) decomposition of the image. Much interest has recently been centred on wavelet decomposition [1], however, alternatives have recently been described [2, 3, 4, 5] that bind edges intimately with scale related features in a way that is not possible with linear decompositions.

Since edges are important, for it is they that describe the shapes and relationships of objects, these non-linear decompositions, are very attractive as a pre-processing step in image pattern recognition. Non-linear morphological methods that have been applied to pattern recognition include alternating sequential filters, \( N \) and \( M \), [6, 5], median filters [7], simple open and close filters [8, 9], however, more complex filters show some advantages. Studies suggest, for example, that 2D sieves can outperform linear cross-correlation, as a method for pattern recognition in images [6, 10]. The choice of structuring elements (masks) influences the extent to which 2D features are decorrelated in the transform, namely granularity, domain, and so the ease of pattern matching. One implication is that problems will need different shaped masks or adaptive masks [5]. A simpler alternative is considered in this paper namely that of using 1D rather than 2D decomposition (for pattern recognition but not for image smoothing or removing random noise, [11]). The recognition step is still performed in 2D. Of course, it is necessary to repeat the decomposition at a number of angles, but there are a number of advantages to set against this. For example, changes of apparent shape causing contra-lateral rotation simply cause the feature to be resolved at different scales by sieves set at different angles. Another, more fundamental, advantage is explored in this paper.

An important property of decomposition by 1D sieves, is that certain of them, including alternating sequential \( M \) and \( N \) filters but not median filters, yield invertible transforms \([12, 13]\). Let \( \vartheta \) be a signal and let the sieve perform a mapping \( G \) (the set of granules that represent \( \vartheta \) in the granularity domain), \( f : \vartheta \rightarrow G \). It appears that where \( G \subset G \), there is a mapping from \( G \) to \( \vartheta \) and back to \( G \), i.e. \( \hat{G} = f^{-1}(G) \). Consequently there is a 1:1 mapping between \( G \) and \( \vartheta \) and so patterns recognised in \( G \) domain will be representative of those in \( \vartheta \) domain. However, 2D \( M \) and \( N \) filters are not invertible, so presumably there is a many to one mapping, and this should reduce the effectiveness of pattern recognition by this route. This paper provides exploration of pattern recognition using a decomposition in 1D.

2. Method

The \( M, N \) and median based sieves are described in detail [5, 6]. In essence they consist of a set of increasing scale bandpass non-linear filters that decompose a signal into a set of increasing scale granules. The decomposition yields a set of granules at each scale, \( G_m = \{g(x,a,m)\} \); where \( x \) is position, \( a \) is amplitude, \( m \) is scale. The set of granule sets, \( G = \cup G_m \) is used for pattern recognition. In this study images are sieved in one orientation only, namely line by scan line, however, a full recognition system would use granules obtained from decompositions at a number of angles. Unlike the \( M \), \( N \) and root median sieves, a multiscale cascade of self-dual morphological filters [6] spreads pulses of a given width to many scale channels, and to this extent should not be considered a sieve. However, it does appear to exhibit some useful invariance properties and so is used here as another benchmark.

Scenes were digitised using a Sony SSC-M370CE (768 by 512 pixel CCD, fitted with 16mm lens) using a
DT2800 board. The resulting images are decomposed line by line by each of the filters to a scale of $m = 60$, the maximum necessary to completely decompose the target pre-set potentiometer.

Pattern recognition is performed by counting the number of granules in the target granularity, $G^r$, for which there is a granule in $G^b$ at the same $m$ and position offset to $x$ and expressing this as a percentage of the total number of granules in $G^r$ to yield a $P_b$, [10]. The resulting surface is similar to, but not the same as, the cross correlation surface obtained by linear cross-correlation between the target and the image (locally normalising both to mean zero and standard deviation one). Neither of these surfaces (examples given in Fig. 3) is easily interpreted so suitable thresholds, $T_r$ and $T_b$, are chosen to allow well matched regions to be selected. The result of, in the linear case convolving the surface with the target image to produce the 'matched filter output' and in the non-linear case summing $G^r$ that match granules in $G^b$ (at their respective positions) to produce the 'matched sieve output', is a more easily interpreted result.

3. Results
First consider an abstracted example of 1D pattern matching using linear and non-linear methods. The 'image' consists of the signal shown in Fig. 1a. The first goal is to find all objects exactly two samples wide. Fig. 1b shows the matched sieve output and Fig. 1c shows the cross-correlation between Fig. 1a and a filter with taps [001100]. A simple thresholding operation on this result is not enough to resolve the two sample wide objects. The problem becomes more serious with larger objects. Fig. 1f shows the cross-correlation obtained between Fig. 1a and an eight sample wide pulse. Replacing the tap value of $(0000000111111000000000)$ by those that minimise the r.m.s. error makes little improvement. Contrast this with the matched sieve output, Fig. 1c. In both cases the matched sieve outputs exactly the correct signal.

Fig. 4c shows an image containing six pre-set potentiometers, their positions, 1 to 6, are indicated by the numbers in the panel below. The image also includes clutter consisting of other items some of which partially occlude the targets and a discontinuous background. Fig. 2a shows the cross-correlation surface between Fig. 4c and Fig. 4g. The white areas represent places where the two match, thresholding and recorrelation produces Fig. 3a. The equivalent surface, $P$, obtained by counting granules in $G^{a4}$ that match those in $G^{b4}$ is shown in Fig. 2b. Fig. 2c shows the same surface with well matched regions highlighted (reproduction is rather poor). The associated matched sieve output is shown in Fig. 4a. Unlike the matched filter, all six pre-sets are located correctly. In this case decomposition is performed using the self-dual cascade. Fig. 4b, d, e show similar results obtained with the root median, $N$ and $M$ based sieves. The invertible, self-dual, $N$ and $M$ methods yield marginally better results. Fig. 5 shows results obtained in the presence of added Gaussian noise. The root median based sieve, Fig. 5b, unambiguously picks out the three dark pre-sets (nos. 1, 2 and 3) and also identifies no. 6. Fig. 5a, d, e shows that the other filters are less successful. It is noticeable that the erroneous matches around no. 1 obtained by the self-dual decomposition are rather comparable to that obtained by the linear filter, Fig. 3b. They are due to the dark capacitor, the d.i.l. chip and the mica resistor. These results are quantitatively compared to those using 2D decompositions [5].

4. Conclusions
The 1D granularity decomposition of 2D images can be successfully used for pattern recognition. The median based sieves are more robust, whilst the others are better under non-extreme conditions presumably because they yield invertible transforms and therefore there is a 1:1 mapping between the granularity domain and the spatial domain. [12, 13].

References

Figure 1. Trace A shows a signal characterised by sharp edged objects. B shows the matched sieve output when the target is a pulse of the form \(f(001100)\), but any amplitude. C shows equivalent cross-correlation. D, E, F same as A, B, C but the target is an eight sample wide pulse.

Figure 2. Matching surface (a) normalised cross-correlation surface. (b) self-dual, clutter noise only, match surface. (c) object identity for (b).

Figure 3. Matching result using normalised cross-correlation, in presence of (a) Clutter noise. (b) gaussian noise. (c) Impulsive noise.
Figure 4. Clutter noise match result from (a) self-dual, (b) root median, (d) N, (e) M sieves. (c) Image containing clutter noise, (f) object ident for match, (g) target pattern.

Figure 5. Gaussian noise match result from (a) self-dual, (b) root median, (d) N, (e) M sieves. (c) Image containing clutter noise and gaussian noise, (f) object ident for match, (g) target pattern.
A CLASS OF NONLINEAR VECTOR FILTERS
APPLICATION TO MULTI-SPECTRAL IMAGE FILTERING

Wael ALSHATTI, Philippe BOLON, Patrick LAMBERT

Laboratoire d’Automatique et de MicroInformatique Industrielle
LAMII / CESALP - Université de Savoie - BP 806 - F-74016 Annecy Cedex
(CNRS-GdR 134 Traitement du Signal et Image)
e-mail: bolon@univ-savoie.fr or lambert@univ-savoie.fr

Abstract

Multi-spectral images can be regarded as vector valued signals. Filtering this kind of images is considered in this paper. Methods based on ordering multi-variate data have been found to be noise model dependent methods. They are defined by using sub-ordering techniques. The aim of this work is to propose two families of vector filters based on an approximation criterion. The first one concerns smoothing. These filters are optimal in the sense of minimizing the proposed criterion. The proposed filters are compared to existing methods. Applications to color image filtering are presented.

Keywords: Image processing, Color, Image filtering, Order statistics.

I. INTRODUCTION

Multi-spectral (MS) images are very useful to get a good description of scenes. Generally, the different components of MS images are highly correlated data [1], especially in the vicinity of the contours which can be regarded as deterministic signal. The correlation between noise components is shown to be lower compared with that of the signal.

Filtering MS images did face a large difficulty because of the vectorial nature of these images. Order statistics based filters have found a large success in gray-levels image processing. They have good properties of both smoothing and edge preservation. Extending these methods to the MS case cannot be made in a straightforward way. Since filtering is applied to ordered data, a method of ordering multi-variate data should be adopted. Four sub-ordering methods were proposed in [2]. Only two of them are applicable to MS image filtering. They are called marginal and reduced ordering methods.

In marginal ordering, data of each component of the image are ordered independently. A family of L-filters based on marginal ordering were proposed in [3] using a maximum likelihood (ML) criterion adapted to noise models having independent components. The component wise median (cwm) is one of these filters. It is defined as the ML filter for exponential noise. The output vector is composed of the output of scalar median filters applied on each component of the image. As shown in [3], these filters cannot eliminate impulses nearby transitions. In this situation, new regions could be created in the filtered image.

A better method based on marginal ordering was proposed in [5]. The proposed criterion is the least square error. This method considers the correlation between the image components, but requires knowledge about signal and noise so that applying this method is difficult.

With reduced ordering, data are ordered according to measured distances between data vectors and a centre of measure which is a priori chosen (it can be the cwm or the average vector). Reduced ordering was used in [5] and [6] for deriving an extension of the scalar rank and α-trimmed filters to the MS case. The proposed filters need thresholds for adjusting the filter performances. Using these filters gives rise to two practical problems: the choice of the measure centre and the choice of the thresholds which cannot be made analytically.

In this paper, we propose a class of filters based on an approximation criterion. Two families of filters can be defined. The first one considers rank filtering. The second one is a family of smoothing filters. A description of the proposed criterion will be given in the following section. Section III presents the family of rank filters. Then, we present in section IV the family of smoothing filters and we give some results about their properties. In the last section, results of the application of the proposed filters to color images, as a special case of MS ones, will be presented and compared with those obtained by ML filters.
II. PRINCIPLE AND DEFINITIONS

The principle of the proposed method is to describe filtering of vector valued signals by means of an approximation criterion. The filter output can then be found by minimizing this criterion. The basis of such an optimization approach was introduced in [7], and was applied to gray-level image filtering [8] [9]. A generalization to MS image filtering is presented in this paper.

Let \( W \) be a window of size \( N \) included in a noisy stationary region of a MS image having \( p \) components. Let \( S = \{X_1, X_2, ..., X_N\} \) be the set of vectors inside \( W \). Then, the filter output is the vector \( Y \) which minimizes the criterion defined as a cost function by:

\[
E_{\alpha\beta}(Y) = \left[ \sum_{i=1}^{N} (\|Y - X_i\|^p)^{\frac{1}{p}} \right]^{\frac{1}{p}}
\]

with

\[
\|Y - X_i\|^p = \left( \sum_{j=1}^{p} (y_j^i - x_j^i)^p \right)^{1/p}
\]

and \( \alpha \) and \( \beta \) are the filter parameters having real positive values. In the general case, there are two main situations according to the values of \( \alpha \) and \( \beta \). For parameters values greater than or equal to one, \( E_{\alpha\beta} \) is a convex function. In the other cases, \( E_{\alpha\beta} \) turns out to be piecewise concave. Two families of NL-filters (non-linear relationship between the input and output) can then be defined. They are presented in section II and section III.

II. ADAPTIVE RANK FILTERS

If at least one of the two parameters \( \alpha \) and \( \beta \) have values lower than one the cost function is piecewise concave. For MS image processing purposes, an interesting situation occurs if \( \alpha \) is lower than one and \( \beta \) is greater than or equal to one. In this case the cost function has \( N \) local minima associated to each of the \( N \) input vectors. Figure 1 shows the cost function for three bidimensional vectors.

![Figure 1: Cost function for adaptive rank filter](image)

Minimizing the cost function results in an optimal output corresponding to one of the input vectors. Then the defined filter can be regarded as an adaptive rank filter. It is adaptive since the selected rank depends on the data inside \( W \) and can be controlled by \( \alpha \) and \( \beta \). These filters are studied in [10]. It is shown that they have edge enhancement properties.

IV. CONVEX OPTIMISATION FILTERS (C.O.F)

These filters are defined for values of \( \alpha \) and \( \beta \) greater than or equal to one. Then the cost function, which is the accumulated Minkowsky distance for the set of data in \( W \), is convex, having only one minimum which represents the optimal output to be calculated. Figure 2 shows the cost function \( E_{\alpha\beta} \) (for two components vectors) for some values of \( \alpha \) and \( \beta \).

![Figure 2: cost function for C.O.F](image)

It can be noticed that for special values of \( \alpha \) and \( \beta \), we obtain some well known ML estimators [4]. For \( \alpha = \beta = 1 \), we obtain the component-wise median. When \( \alpha = \beta = 2 \) we have the average filter. And if \( \alpha \), \( \beta \to \infty \) we obtain the midpoint filter. In fact, we find that performing marginal processing corresponds to choosing identical for values for \( \alpha \) and \( \beta \). This result is obvious from the definition of the cost function.

Then the optimal filter output \( Y \) is found by setting the \( p \) partial derivatives of \( E_{\alpha\beta} \) to zero:

\[
\frac{\partial E_{\alpha\beta}}{\partial y^i}(y^1, ..., y^p) = 0
\]

In the general case, we can not solve explicitly the system of equations above, but it is easy to do it numerically by using a gradient method. In the case of \( 8 \) bits color images, the computing time is found to be 7 ms per pixel using a SUN-SPARC10 workstation.

IV. PROPERTIES OF C.O.F

a. variance

One of the advantages of this family of filters is a
better noise reduction compared with the ML estimators. This can be shown by studying the variation of the output variance versus $\alpha$ and $\beta$. Figure 3 shows the results for exponential, gaussian and uniform noises. The noise, independent between components and between vectors, was added to a synthetic 64x64 constant color image.

\begin{itemize}
  \item exponential noise
  \begin{equation}
  \text{minimum variance} = 0.28 \\
  \text{for } \alpha = \beta = 1.4.
  \end{equation}
  \item gaussian noise
  \begin{equation}
  \text{minimum variance} = 0.32 \\
  \text{for } \alpha = \beta = 2.2.
  \end{equation}
  \item uniform noise
  \begin{equation}
  \text{minimum variance} = 0.28 \\
  \text{for } \alpha = \beta = 3.0.
  \end{equation}
\end{itemize}

Figure 3: variance for filter size = 3.

The results show that the minimum variance is reached for values of $\alpha$ and $\beta$ different from those corresponding to the ML estimators. The variance obtained using our filters is lower than that obtained using ML estimators. The noise reduction gain approaches 20%. Going on increasing the filter size, it can be shown that the results of the ML filters converge to those obtained using C.O.F. Figure 4 shows the variation of the variance with the filter size.

\begin{itemize}
  \item exponential
  \begin{equation}
  \text{variance} = 0.32 \\
  \text{for } \alpha = \beta = 1.4.
  \end{equation}
  \item gaussian
  \begin{equation}
  \text{variance} = 0.32 \\
  \text{for } \alpha = \beta = 2.2.
  \end{equation}
  \item uniform
  \begin{equation}
  \text{variance} = 0.28 \\
  \text{for } \alpha = \beta = 3.0.
  \end{equation}
\end{itemize}

Figure 4: evolution of the variance versus the filter size. (independant exponential noise).

This means that we can obtain, with C.O.F filters, the same amount of noise reduction given by the ML estimators, but with smaller filter sizes. This results in a better preservation (from a geometrical point of view) of the shapes in the image.

The proposed NL-filters are robust against correlation between the image components. In fact, experimental studies of the variance show that the value of the variance is practically independent on the correlation between the different components. However, the optimal values for $\alpha$ and $\beta$ can be modified. The table below shows the values of $\alpha$ and $\beta$ obtained when varying the correlation between components.

<table>
<thead>
<tr>
<th>noise correlation</th>
<th>independent ($\alpha$/var)</th>
<th>medium ($\alpha$/var)</th>
<th>high ($\alpha$/var)</th>
</tr>
</thead>
<tbody>
<tr>
<td>exponential</td>
<td>1.4/1.4/0.282</td>
<td>1.3/1.5/0.291</td>
<td>1.2/2.1/0.294</td>
</tr>
<tr>
<td>gaussian</td>
<td>2.0/2.0/0.326</td>
<td>2.0/2.1/0.328</td>
<td>2.2/6.0/333</td>
</tr>
<tr>
<td>uniform</td>
<td>3/3/0.286</td>
<td>3/3/0.290</td>
<td>3/3/0.295</td>
</tr>
</tbody>
</table>

b. mean absolute error

Marginal filtering of MS images produces false regions nearby transitions where the signal is highly correlated. The proposed NL filters have the nice property to give a better response nearby transitions if appropriate values of $\alpha$ and $\beta$ are used. In order to illustrate this property, we calculated the mean absolute error (MAE) for a transition having a 3 pixel width contaminated by an exponential correlated noise. We compared C.O.F filter with cwm filter and standard vector median filter (svm) defined in [4]. Figure 5 shows the MAE for these filters.

\begin{itemize}
  \item exponential noise with SNR = 1.4
  \begin{equation}
  \text{MAE} = 0.28
  \end{equation}
\end{itemize}

Figure 5: the MAE for different filters (size = 7).

It should be noticed that both svm and cwm filters produce greater errors than C.O.F filter in the vicinity of the transition. CWM filter is slightly better than C.O.F filter in stationary areas. In fact, parameters values minimizing the MAE criterion are not the optimal ones for the variance.

c. deterministic properties

Let us consider a sharp transition (step edge) between two adjacent regions, $R_1$ and $R_2$, characterized by vectors $V_1$ and $V_2$ respectively. For the sake of simplicity, we consider a 1D vector signal, the transition being located at pixel $k=0$. Assume that $V_1=0$ and $V_2=H$. 

862
The filter output at pixel $k$ is:

$$Y_k = H \cdot \theta(u)$$

with

$$u = \frac{N + 1 + 2k}{2N}$$

and $\theta(u) = 0 \quad \text{for } u \leq 0$

$$= 1 \quad \text{for } u = 1$$

$$= \left[ 1 + \left( \frac{1 - u}{u} \right)^{\alpha - 1} \right]^{-1} \quad \text{otherwise}$$

The expression above gives the transition blurring effect caused by the filtering. Figure 6 illustrates this phenomenon.

![Figure 6: The blurring effect of C.O.F.](image)

By setting parameter $\alpha$, the behaviour of COF filters can be adjusted between the one of a cwm ($\alpha=1$) and the one of an average filter ($\alpha=2$).

4. Choice of $\alpha$ and $\beta$ values

Choosing identical values for $\alpha$ and $\beta$ corresponds to the marginal processing. Practically, we obtain good results if $\alpha < \beta$. We observe that the smoothing effect is controlled by parameter $\beta$ and parameter $\alpha$ has effects on transition preservation.

VI. APPLICATIONS

We present here an application to color image filtering as a special case of MS images (Fig.7). It can be noticed that, given a filter size, COF filtering (7e) produces a more efficient smoothing effect than cwm filtering (7b). The blurring effect is limited by choosing $\alpha < 2$.

By choosing $\alpha < 1$, we can obtain an edge enhancement effect (7e). This is interesting for preprocessing images before a segmentation step.

VII. CONCLUSION

In this paper, two families of vector filters are presented. The first one can be regarded as an extension of scalar adaptive rank order filtering. The second one, COF filters, can be regarded as an extension of scalar L-Filters.

Figure 7: color image

Multivariate Median and Trimmed Mean Filters

Kari-Pekka ESTOLA

Nokia Research Centre
Heikkiläntie 7, SF-00210 Helsinki, FINLAND
Tel. +358-0-43761, Fax. +358-0-4552082

Abstract. This paper introduces a new vector median that has similar characteristics as the classical vector median. The proposed vector median finds the sample that is surrounded evenly by the other samples, i.e., relative to the median the rest of the input set is distributed evenly in all directions. The new vector median does not depend on the distances between the vector median and the rest of the input set but on the angles where the rest of the input set is seen. The structure of the new direction and magnitude trimmed mean filters correspond to multistage filtering where the first stage is a robust nonlinear operator followed by a linear operator in the second stage. Examples show that the performance of the proposed vector medians compares favourable to the reference methods.

1. Introduction

Standard Median (SM) filters defined by Tukey [1] perform simple operations of sorting the samples in the filter window and picking the output sample from the mid position of the sorted data window. SM filters reject impulsive noise efficiently and they do not smear the edges when the amount of additive noise is moderate. However, the edge preserving property degrades along with the increased variance of the additive noise i.e., edges in noisy signals are not preserved very well [2]. The median filters computing the raw estimate of the true median using the Tukey's method can also be expressed as a special case of Order Statistic (OS) filters or L-filters. Other well known members of OS filters include rank-order filters, mean filters and alpha trimmed mean filters [3]. OS filters have also been optimized for certain noise distributions such as Laplacian noise [4].

Traditionally median operations have been applied to one or two dimensional signals. However, there is an increasing demand to use nonlinear operations to multivariate (or vector) multidimensional signals. Such signals are for example color images in television systems, multispectral satellite images and multichannel signals encountered in geophysics and underwater acoustics. There are also a number of applications where position or velocities are tracked in three axis coordinate systems.

Definitions for Vector Median (VM) operators have been proposed for $L_2$ and $L_1$ norms [5], [6]. In these definitions the vector median of $x_1, x_2, \ldots, x_N$ is $x_{vm}$ such that

$$x_{vm} \in \{x_i | i = 1, 2, \ldots, N\}$$

and for all $j = 1, \ldots, N$

$$\sum_{i=1}^{N} \|x_{vm} - x_i\|_l \leq \sum_{i=1}^{N} \|x_j - x_i\|_l,$$

where $l$ denotes the norm. The above definitions are suboptimal in the maximum likelihood sense to Laplacian ($l=1$) and Gaussian ($l=2$) distributions. The definition for generalized vector median for arbitrary distributions is also given in [6]. Design procedures for optimal coefficient sets of multivariate OS filters have been presented in [7]. These designs are based on the concept L-filters exploiting the marginal ordering and using the mean squared error as the optimization criteria. The drawback of some of the designs is that they need the estimate of the signal vector.

Also other definitions for multivariate median filters have been suggested. Such approaches are for example to output the multivariate sample that is closest to the mean or the median operation is computed separately for each component i.e., componentwise filters. The noise suppression capabilities of the multivariate median filters using componentwise median filters and filters with linear
substructures are better than those of the VM filters defined in the previous paragraph. The problem of the exploding computational burden as the dimension of the variable space increases has been addressed in [8]. This approach cuts the dimension of the variable space using optimal projections to one dimensional variable spaces.

2. Proposed Filter Structures

Vector median filters are defined as filters which select the sample from the input set that minimizes the sum of the distances to all the other samples. Thus vector median is a sample that is in a central position among the input sample set. This paper introduces a new vector median that shares the same characteristics of being in a central position among the input set. The proposed vector median finds the sample that is surrounded evenly by the other samples, i.e. relative to the median the rest of the input set is distributed evenly in all directions. The new vector median does not depend on the distances between the vector median and the rest of the input set but on the angles where the rest of the input set is seen.

The proposed Vector Direction Median (VDM) of \( x_1, x_2, \ldots, x_N \) is \( x_{vdm} \) such that

\[
x_{vdm} \in \{ x_i | i = 1, 2, \ldots, N \}
\]

and for all \( j = 1, \ldots, N \)

\[
\left| \sum_{i=1}^{N} x_{vdm} - x_i \right| \leq \left| \sum_{i=1}^{N} x_j - x_i \right|
\]

In the above definition the summing is done over the unit vectors in the directions of \( x_i \) and thus the distance between the multivariate samples has no effect in the computation of the VDM. The pairs of samples for which \( x_j - x_i = 0 \) are not included in the summation. In the actual paper efficient methods for computing the VDM are presented. It should be noted that the definition for the VDM gives, in the scalar case, exactly the same median as the VM. For scalars the unit vectors are 1 or -1 depending on whether the candidate for the median is larger or smaller relative to the sample compared. Thus, for \( N \) being odd, the output of the VDM is the sample for which there is an equal number smaller and larger samples i.e. the classical median. The special case of multiple samples of the same value is also handled correctly.

This paper proposes also new multivariate trimmed mean filters. For the new filters the trimming of the outliers is based on the definitions for the VM and VDM filters and the output of the trimmed mean filter is the mean of \( K \) most central samples. The output of the new trimmed mean filters is obtained by computing the metrics \( y_{m,j} \) or \( y_{d,j} \) defined as

\[
y_{m,j} = \sum_{i=1}^{N} \left| x_j - x_i \right|, \quad j = 1, 2, \ldots, N
\]

and

\[
y_{d,j} = \left| \sum_{i=1}^{N} x_j - x_i \right|, \quad j = 1, 2, \ldots, N
\]

and picking the \( K \) samples as vectors \( x_{m,k} \) or \( x_{d,k} \), \( k = 1, 2, \ldots, K \), from the input set \( x_1, x_2, \ldots, x_N \) that correspond to the \( K \) smallest values of \( y_{m,j} \) or \( y_{d,j} \). The output \( x_{vdm} \) of the Magnitude Trimmed Mean (MTM) filter is

\[
x_{vdm} = \frac{1}{K} \sum_{k=1}^{K} x_{m,k}
\]

and the output \( x_{vdm} \) of the Direction Trimmed Mean (DTM) filter is

\[
x_{vdm} = \frac{1}{K} \sum_{k=1}^{K} x_{d,k}
\]

correspondingly. In the scalar case the MTM and DTM filters are similar to the alpha trimmed mean filters. The distinction is the way how the outliers are removed - alpha trimmed mean filters remove samples from the both ends of the ordered set of the input samples whereas MTM and DTM filters remove the most distant samples based on the metrics not regarding how they are located in the ordered set.

The structure of the new direction and magnitude trimmed mean filters correspond to multistage filtering where the first stage is a robust nonlinear operator followed by a linear operator in the second stage. If the distribution of the multivariate data is heavily tailed the nonlinear stage removes the possible outliers from the data and the linear operator is applied to the data which has been reshaped to a Gaussian-like distribution. Examples show that the proposed direction trimmed mean filter outperforms the vector median and magnitude trimmed mean.
filters in random noise suppression and it preserves the deterministic shapes in the data well.

3. Numerical Examples
In order to show the superiority of the new hybrid multivariate trimmed mean filters over the VM, componentwise median (CWM) and mean filters two numerical examples are considered. The first example compares the noise suppression capability of the new filters with the VM (z=1), componentwise median and mean filters. The two-variate data is generated in such the way that diagonals in the angles π/4 and -π/4 are independently distributed Gaussian or Laplacian noise. In the example the number of two-variate data points is 4500 and the length of the mean and medians are 15. For the direction and magnitude trimmed mean filters the number of points used in averaging is 3.

Table 1. Random noise attenuation in dB.

<table>
<thead>
<tr>
<th>Noise</th>
<th>DTM</th>
<th>MTM</th>
<th>VM</th>
<th>CWM</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplacian</td>
<td>-12.3</td>
<td>-11.9</td>
<td>-11.1</td>
<td>-11.4</td>
<td>-11.5</td>
</tr>
<tr>
<td>Gaussian</td>
<td>-9.9</td>
<td>-9.4</td>
<td>-8.9</td>
<td>-10.3</td>
<td>-12.2</td>
</tr>
</tbody>
</table>

It is seen in the Table 1 that the new direction trimmed mean filter outperforms the other filters in the Laplacian noise attenuation. For the Gaussian noise mean is of course the optimal estimator and VM filter is worst. For the both noise types the new direction trimmed mean filter is 1 dB better than vector median.

The second example addresses the ability of the compared median filters to preserve deterministic shapes in the data. The input two-variate data forms a unit circle with a neck where the number of the points on the circle, displaced by π/16, is 31 and in the neck 10 in both ends. The length of the median filters are 9. Figure 1 depicts the input data and Figures 2a to 2d the filtered data using CWM, VM, mean and DTM filters, correspondingly. It is seen that the new direction trimmed mean and VM filters preserve the shapes best. Mean filter has the worst behaviour. Componentwise median crushes the sides of the circle but preserves the neck whereas the direction trimmed mean and vector median preserve the circle but miss points in the neck.

![Figure 1. Two-variate input data for the Example 2.](image1)

![Figure 2. Two-variate filtered data for the Example 2. a) CWM, b) VM, c) Mean and d) DTM.](image2)
4. Conclusions
This paper introduced a new vector median having similar characteristics as the classical vector median. The proposed vector median found the sample that was surrounded evenly by the other samples, i.e. relative to the median the rest of the input set was distributed evenly in all directions. The new vector median did not depend on the distances between the vector median and the rest of the input set but on the angles where the rest of the input set was seen. The structure of the new direction and magnitude trimmed mean filters corresponded to multistage filtering where the first stage was a robust nonlinear operator followed by a linear operator in the second stage. Examples show that the performance of the proposed vector medians compared favourable to the reference methods.

References


Knowledge-based segmentation of medical images

M.J. CARREIRA, D. CABELLO, A. MOSQUERA, M.G. PENEĐO and J.M. PARDO
Departamento Electrónica e Computación, Facultade de Física, Universidade Santiago de Compostela.
15706 Santiago de Compostela, SPAIN. Tel/Fax:+34 [81] 563100 / 520676, E-Mail: electronica@gaes.usc.es

Abstract. The obtention of significant regions from a scene segmentation process requires the use of knowledge about that scene. An important task is how to inject this knowledge into the segmentation process. In this work we present a knowledge-based system for the automatic segmentation of chest X-ray images in which knowledge is used in a progressive way: In the early process stages the knowledge is implicit and not very elaborated; in the following stages it is explicit and represented by means of production rules. Thus an initial segmentation is progressively refined until an image in which all regions have clinical meaning is achieved.

1. Introduction
An important task in automatic systems for image analysis is the segmentation process. It consists in the partition of the image into a set of homogeneous elementary regions. Nevertheless, the building of significant entities from these elementary regions presents problems related to the kind of images to analyze and so it’s required the use of specific knowledge about the domain. The segmentation must be based in a formal mechanism of reasoning around the scene (biomedical image), objects in it (organs and/or pathological structures) and its representation (regions). Several authors have proposed the combination of segmentation and artificial intelligence technics in order to implement image interpretation systems. An important aspect is the insertion of knowledge in the segmentation process. Sherman et al. [1] have introduced the idea of progressive segmentation. Thus, an initial partition is successively refined using the knowledge about the domain in a progressive way. Our proposal for implementing a knowledge-based system for automatic interpretation of chest-radiographic images follows this strategy [2]. Thus, in a first level, operations independent of the domain are done, while in the second level the explicit knowledge we have about it is injected in the system.

We do the progressive segmentation in the following steps: Initially, an oversegmented image is obtained in base on statistical information (histograms). The next step consists in refining this segmentation using spatial information to confirm that the boundaries previously obtained are really edges. The boundaries are modified or eliminated in base on criteria that include in an implicit way a not very elaborated knowledge about the domain. This low-level algorithm, which integrates statistical and spatial information, leads to a segmentation in a reasonable number of regions, which is a good starting point for the high-level block. This first makes a split-and-merge of regions guided by knowledge about organs morphology in order to identify them. For doing this, regions are characterized by means of a set of morphometric, densitometric and relational properties and the explicit knowledge is represented by means of production rules. This process allows to extract the lungs, and to label correctly their contours (heart silhouette, aortic node, ...). A latter analysis will arise to a symbolic description of the pathological structures found in the image.

2. Initial segmentation
The initial segmentation of the image is based on the statistical information given by its histogram. For this we use a histogram division iterative algorithm based on the analysis of the local clustering centers [3]. In a first stage, we apply this segmentation algorithm to the original image taking as a parameter the minimum percentage a cluster must have to be considered region. With this we obtain a rough segmentation of the image made up of two or three significant regions. In the second stage, we apply the algorithm to the previously obtained regions, introducing as a parameter the minimum length of the interval which must exist between two adjacent local clustering centers, controlling this way the level of detail desired for each area and leading to an oversegmentation of each region of the image. In each iteration the elementary regions arise by the transfer of class information from the feature space into the image space. From this translation we obtain a boundary image. In the kind of images we consider (radiographic images), the two main different kinds of regions are ‘dark’ (lungs and background) and ‘light’ (mediastinum and lateral tissue), and we take a second parameter for each one.

Before next process for refining this segmentation, it is necessary to organize in an efficient way the information obtained. This implies labeling both the vertex and the boundaries between vertex. After this process, the information is structured as a data matrix $B (\text{label}, \text{len(label)}$) in which all the contours are stored, where label is the label for each contour and len(label) is the location of each point in it. This matrix will be the starting point for the stage of eliminating no-edge boundaries.
3. Elimination of no-edge boundaries

The main error observed in the segmentation provided by this clustering algorithm in chest X-ray images consists in the multiplicity of boundaries in those areas in the image with smooth transitions in grey level which basically correspond to transitions between organs. The refinement of the initial segmentation, tending to confirm, modify or eliminate the previously obtained contours, must penalize that parallel boundaries, establishing, if possible, a unique contour in the greatest contrast area. The refinement procedure consists in calculating for each boundary a merit function that images contrast, length and surrounding information criteria and eliminating, if there are no topological restrictions which prevent it, those boundaries with small merit functions [4].

For each boundary \( b \) we define a merit function \( f_m(b) \) as the contrast between the regions \( i \) and \( j \) it separates. In principle, the boundaries for which this merit function is small must be eliminated, thus penalizing the artifacts of the initial segmentation. But to eliminate a boundary might imply the later elimination of a similar one. Therefore, we must include a series of topological restrictions which must be verified and we will define a global merit function that must be maximized [5] and which is expressed in terms of \( f_m() \) in the following way: If boundary \( b \) must be eliminated, it contributes to the sum with \( \theta - f_m(b) \), and if it must be maintained it contributes with \( f_m(b) - \theta \), where \( \theta \) is the expected contrast across the edges of the image. If we denote the set of indexes for the two just mentioned groups of boundaries as \( I_n \) and \( I_k \) respectively we will have the global merit function of eq.(1).

\[
F(b_1, b_2, ..., b_n) = \sum_{i \in I_n} (\theta - f_m(b_i)) + \sum_{i \in I_k} (f_m(b_i) - \theta)
\]

(1)

If we forget the restrictions imposed by the topological validity, \( F() \) would be maximum if we eliminated all the boundaries with \( f_m(b) \) less than \( \theta \) and kept the rest. Nevertheless, in order to maintain the topological validity we must keep some boundaries with a low \( f_m() \) value. The boundary elimination strategy follows a 'greedy' algorithm: every time we eliminate the weakest boundary, we check that its merit function is less than \( \theta \). The process finishes when there are no more boundaries which verify the condition. This process can be modified by predicting the effect of \( F() \) if the boundary we are considering is eliminated and not eliminating it if the effects are no desirable. Let \( b \) be the candidate; if we eliminate it we must reorganize the configuration of boundaries and regions. If during this process, we join boundaries with very different values of \( f_m() \) or the length of \( b \) is much shorter than the length of the contours to join, then the elimination of \( b \) is cancelled. All these constraints are introduced by a restrictive function that must be true in order to eliminate the boundary. This function \( \text{better}(b) < \text{value} \) is true when the following conditions are verified: if eliminating \( b \) does not imply the union of other boundaries, if the length \( b \) is not much shorter than those of the boundaries to be joined and it is not a closed and isolated boundary, probably associated with a lesion in the kind of images we consider. In order to define these conditions we fix some parameters at the beginning of the process. These parameters will be the minimum percentage in order to consider the merit functions as similar, the minimal ratio between the lengths and the contrast threshold \( \theta \). With this we have finished the low-level analysis of the image in which domain knowledge appears in an implicit way in criteria and thresholds selection.

4. Knowledge-based analysis

The elementary regions obtained from the low-level block make up a first image description which is given a clinical meaning by means of the high-level block. For doing this, these regions are described by a set of densitometric, morphometric and relational properties: area, minimum bounding rectangle (MBR), regularity, mean grey value, mass center coordinates and relational aspects (inside of, at the right of, ...).

As regards to the kind of images we're considering, we have so anatomical descriptive knowledge, represented by production rules, as heuristic knowledge, which indicates the strategy to use for analyzing the regions. Thus, first a 'regularization' process on elementary regions is done. Regions with 'low regularity' and 'large area' are split in to obtain other regions with 'more regularity' if in the contour image provided by a Canny filter there are contours 'strong enough' to divide them. The next step in the analysis procedure consists in identifying characteristic organs (lungs), what take us to make an 'image plan'. A later process explores the remaining regions, giving them a proper significance (hilia, heart silhouette,...) or merging them to another already identified region. The described strategy is implemented by means of a control structure which triggers the execution of region-split, organ recognition or region merge/assignment rules depending on the recognition stage. The process finishes when all regions have a label assigned.

4.1 Region-Splitting

The process of splitting some regions in the image is needed because of the irregularity of some important regions. This irregularity is caused by the loss of some important weak edges in the previous stage. In order to restore them, we apply a Canny filter to the original image which first convolve the image with a Gaussian filter of standard deviation sigma. Points not at the peak of a strength gradient are suppressed and then, a hysteresis with thresholds \( t_1 \) and \( t_2 \) is carried out over this image in which we keep on edges stronger than \( t_2 \) and edges stronger than \( t_1 \) connected with the first ones. A latter process is applied for connecting terminating pixels that are within specified limits of distance and orientation.

Once the Canny image is obtained, we compute the area and regularity of all regions in the image and look for the canny edges in regions large enough and with regularity lower enough. So we define the regularity of a region \( k \) as in eq.(2), where MBR represents the minimum bounding rectangle containing region \( k \).

\[
\text{regularity}_k = \frac{\text{area}_k}{\text{area}_{\text{MBR}_k}}
\]

(2)

Thus a new image is obtained, in which irregular regions are split into several more regular regions. This labelled image will be the starting point for the following stage.

4.2. Region-Merging

From the image obtained in the previous stage, we compute the features of the regions. These features are the area, mean grey value, regularity and coordinates of the mass center. Then, we look for the three largest regions in the image, which usually will correspond to right lung, left lung and mediastinum with lateral tissue. The identification of these three regions will lead us to a preliminar interpretation of the image.

After this step, we define grey level thresholds depending on the maximum and minimum grey values of the whole image as can be seen in eq.(3), and position thresholds in coordinates \( x \) and
y as indicated in eq.(4), where dimx and dimy are the sizes of the image in horizontal and vertical directions respectively.

\[ u_k = (s_{\text{max}} - s_{\text{min}}) \cdot \frac{k}{\frac{5}{3} \cdot \text{min}}; \quad k = 1, \ldots, 4 \tag{3} \]

\[ dx_k = \frac{\text{dimx}}{3} \cdot k; \quad k = 1, 2 \quad dy_k = \frac{\text{dimy}}{4} \cdot k; \quad k = 1, \ldots, 3 \tag{4} \]

We also define two more thresholds in dimension y as indicated in eq.(5).

\[ uy_0 = \frac{1}{20} \cdot \text{dimy} \quad uy_1 = \frac{19}{20} \cdot \text{dimy} \tag{5} \]

The reason for introducing the thresholds uy0 and uy1 is for considering the merging of regions positioned below uy0 or above uy1 under more restricted conditions arising from the desire of no merging regions in lower diaphragm or in upper clavicle or lateral tissue.

Once computed the grey and position thresholds, we try to identify the lungs in the three major regions of the image. The rules for identifying them will be:

1. Mean grey level of candidate region (g_k) must be less than u1.
2. Coordinate y of candidate region mass center (cmy_k) must be in the interval between dy1 and dy2.
3. If coordinate x of candidate region mass center (cmx_k) is less than dx1, then region k is right lung, else if it is greater than dx2, region k is left lung.

In these rules we have injected a very simple knowledge about the position and grey level value of lungs in radiographic images.

After identifying lungs, we look for the regions similar to right and left hilia and heart silhouette. The reason for doing this before merging is because we will only merge no-labelled regions, and so we will not obtain erroneous merged resultant regions. The rule for identifying left hilia is as follows:

**IF**
\[ \text{grey} > u_1 \quad \text{AND} \quad \text{cmx}_k \text{ is around dim} \cdot \text{y}/2 \quad \text{AND} \quad \text{cmx}_k \text{ is in (dim} \cdot \text{y}/2, \text{cmx}_{\text{left lung}}) \quad \text{AND} \quad \text{area} < 7000 \]

**THEN** k=left hilia and for identifying right hilia, the rule is symmetric.

The next step is the proper merging process. We first merge regions similar to right and left hilia, then regions similar to heart silhouette and finally regions which can be associated to right or left lungs. The merging of hilia regions is a little more restrictive than those for merging lung neighbors, considering only regions with the same characteristics as the rule which identify them but with grey values between the grey thresholds u1 and u4.

For example, one rule for merging a neighbor region of the right lung would be:

**IF**
\[ \text{cmx}_{\text{neighbor}} \text{ in (uy}_0, \text{uy}_1) \]

**THEN** **IF** grey_{neighbor}=u_{right} \text{ OR}
\[ (\text{grey}_{\text{neighbor}} < u_4 \text{ AND position}_{neighbor} = \text{left right lung}) \text{ OR} \]
\[ (\text{grey}_{\text{neighbor}} > u_4 \text{ AND position}_{neighbor} = \text{right right lung}) \]

**THEN** merge(right lung, neighbor)

So we allow merging dark regions (grey<u4) or regions with medium-high grey level (grey>u4) positioned at the left of right lung (external regions) or with medium grey value (grey>u4) and at the right or right lung (internal regions). We distinguish between internal and external regions because of the graduation in the grey level of the whole image from its center to its borders. The internal regions usually have higher grey values than external ones.

For left lung, the first condition can be extended to regions in the interval (uy0, dy4) because the diaphragm is now positioned lower than in the left half of the image, and so regions in (uy4, dy4) could belong to left lung. The other conditions will be symmetric for left lung.

Each time we merge a neighbor region, the properties are updated and next we try to merge another neighbor. The process finishes when there are no more regions verifying the conditions.

Finally, what we obtain is a new boundary image in which we have achieved the opening of lung borders, and so we have a 'cleaner' interpreted image. In some cases it is possible, as a latter step, the identification of the aortic node, with some simple rules about its position and grey value.

5. Preliminary Results

We have applied the progressive segmentation process described in the work to several digitalized chest radiographic images (512x512 pixels, 256 grey levels). In Figure 1a it can be seen the original image over which we have applied the whole process. Figure 1b shows the results corresponding to the initial segmentation after the clustering algorithm. In the first stage, we have considered regions greater than 300 and with minimum percentage of the histogram 13% and in the second stage we have segmented each region classifying them as 'dark' (minimum distance between cluster centers 15) or 'light' (minimum distance between cluster centers 10). The final number of regions after this process is 64 and the matrix B has 139 contours stored.

Figure 1c shows the boundary image after eliminating no-edge boundaries. The multiplicity of boundaries after the initial segmentation in transition areas between organs disappears in large measure with the refinement process. Now, the contours are placed in the greatest gradient areas. The parameters we have used were \theta=40 as the minimum value of contrast for a boundary, relen=0.05 as the minimum ratio between the boundary to be eliminated and the boundaries related with it and diffmin=0.05 as the minimum similarity percentage between the merit functions of the boundaries to join. In the final image we have 23 regions and 36 boundaries stored in matrix B. This segmentation is the starting point for the high-level block.

For the region-splitting, we have applied a Canny filter with sigma=0.3 and hysteresis thresholds of 10 and 25 respectively. Finally, in order to obtain connected edges to divide irregular regions, we have connected terminating pixels with limit distance 15 and limit angle 90. Thus, we consider the canny edges only for regions with regularity less than 0.25 and area greater than 7000.

These irregular regions can be seen in Figure 1c as darker regions. After splitting these areas, we have 84 regions that can be seen in Figure 2a. Following the knowledge-guided split-and-merge process, both regions and contours have clinical significance. Lungs, hilia, heart silhouette, aortic node, mediastinum, diaphragm and lateral tissues have been rightly identified. After this identification of regions, the final number of regions is 27, although this number would be smaller if we also merge regions in mediastinum and lateral tissue areas. Figure 2b shows the resulting labeled regions after approximating the lateral external contours of lungs to curves. A later analysis of the different organs will allow us to obtain a sinbolic description of the possible pathologies found in the image, equivalent to the medical report given by a human specialist, final objective of the knowledge-based system for automatic interpretation of chest images in which the described progressive segmentation process is integrated.
Figure 1. (a) Original chest radiographic image. (b) Boundary image after the initial segmentation. (c) Segmentation obtained after no-edges elimination with 'irregular' regions in dark.

Figure 2. (a) Image after region-splitting. (b) Image after region-merging and interpretation.

Acknowledgements
This work was supported by the Spanish CICYT under Grant No. TIC91/0816.

References
Application of the optimized Canny-Deriche filter for edge detection on cardiac scintigraphic images

Marc GIRARDIER, Dominique MICOLLET, Claude MILAN, Jean-Louis PELLETIER*

Processing language assistant, Computer science lecturer, Laboratoire G.E.R.E, Université de Bourgogne, 
6, Bd. Gabriel, 21000 Dijon, France, Tel/Fax: 80 39 66 43 / 80 39 50 66 
* Supervisor, Centre de Médecine Nucléaire du Parc, 18 cours Général de Gaulle, 21000 Dijon, France, 
Tel/Fax: 80 65 24 44 / 80 65 25 18

Abstract. We have developed a method for quantification of the left ventricular (VG) and ventricular ejection fraction (VEF) from gated single-photon emission computed tomographs. We evaluate the heart activity in diastole (dilatation) and systole (compression). The result is a percentage. Using previous results about generalization of the Canny-Deriche filter on every edge form, we describe a boundary closing method by labelling the edges which we apply just after the filter. Then, we explain a method to fit the contours with a set of 21 masks, the aim of which is to close the remaining holes by polynomial approximation. Finally, we present two examples. In the first case, previous ventricular ejection fraction results were always over 100%. Now the final result shows that the patient is in a good health. The second one is to validate previous results.

1. Introduction.
In order to improve the cardiac scintigraphic image of edges for the purpose of better diagnostics, the heart boundaries have to reach reality with accuracy. An approximate edge leads to wrong diagnostics. Further more, the left ventricle, which shape is almost circular, may be masked by other organs, like right ventricle or aortic cross, which increase the problem.

For all those reasons, we have applied the optimized Canny-Deriche [1][2] filter developed by the GERE group [3] followed by an edge labelling closing boundaries method with thresholds. This filter is composed by the 3 criteria which are the signal to noise ratio, the localization and the cancelling of multiple responses. The criteria must be maximized. Then, fitting and polynomial approximation give a good contour for quantification. Results are better than those obtained with very fast process already validated and sometimes too approximate.

2. Previous results of applying the Canny-Deriche filter to any type of edges and its implementation.
There are many edge detection methods and most of them use first or second order derivates local operators.

But this class of operator is not sufficient for images we deal with. In order to get a usable contour, thresholding and fitting operators are necessary, for noisy images and intensity changes which are not always very sharp. That is why, for a few years, optimal algorithms have been developed, in reference with a predefined edge shape and maximization of criteria. This is the case of the Canny-Deriche filter.

2.1 Previous results.
The shape of the edge, defined by the GERE group, is as follows:

\[ C(x) = \begin{cases} 
1 - \frac{\text{e}^{-|x|}}{2} & \text{for } x \geq 0 \text{ with } \sigma \geq 0 \\
\frac{\text{e}^{5|x|}}{2} & \text{for } x < 0
\end{cases} \]

The optimal filter function is that which maximizes the 3 basic criteria mentioned below:
- signal to noise ratio SNR.
- localization L.
- cancelling of multiple responses MRC.

Mathematical expressions of these criteria are in Appendix A.
It gives for the 3 criteria:
- signal to noise ratio

\[ \text{SNR} = K_1 \frac{\int_0^\infty f(x)(1-\text{e}^{-|x|})dx}{\sqrt{\int_0^\infty f^2(x)dx}} \quad \text{with } K_1 = \frac{1}{n_0 \sigma^2} \]

- localization L

\[ L = \frac{s^2 \int_0^\infty f(x)e^{ax}dx}{\sqrt{\int_0^\infty f^2(x)dx}} \]

- cancelling of multiple responses does not change because of its independence beside the shape of the edge.

The filter function f must maximize the product L*SNR under the constraint MRC. To find f, the variationnal method (Lagrange Multiclitors) leads to the resolution
of a differential equation (Euler equation) which solution is:

\[ f(x) = a_2 ((a_1/a_2) e^{\alpha x} \sin(\alpha x) + e^{\beta x} \cos(\alpha x) - e^{\delta x}) \]

2.2 Filter implementation.
The optimization and parameter search is achieved numerically.
After normalization and Z-transformation, the calculation gives a recursive stable third order filter described by the following equations:

\[
\begin{align*}
Y^+(i) &= b_1 X(i-1) + b_2 X(i-2) + b_3 X(i-3) - b_4 Y(i-2) - b_5 Y(i-3) \\
Y^-(i) &= b_1 X(i+1) + b_2 X(i+2) + b_3 X(i+3) - b_4 Y(i+2) - b_5 Y(i+3)
\end{align*}
\]

3. Closing the boundaries by labelling the edges [4].
Closing boundaries is a capital step in image processing because it allows to give a quantitative approach.
The method developed below is a method with thresholds applied to "grey scale" images.
At first, the image is scanned line by line by labelling each pixel, according to their values with thresholds chosen by the operator (low and high). The lower range of the lower threshold constitutes background noise which we label as 0. The high range of the upper threshold constitutes the real edge which we label as 1. If the pixel is between those two thresholds, it will be labelled as an arbitrary value which is neither background noise nor real edge. Each new region is labelled at a different value and is automatically reassigned a value which is the value of the other that it touches.
An extreme is defined by a pixel which has only one neighbour. We have to discern real and wrong contours among labelled regions. If a region is linked to at least 2 extremes (labelled as 5) belonging to a real contour, it will be reassigned as a real one (1). On the contrary, it will be labelled as 0.

4. Fitted edges and polynomial approximation.
4.1 Edge fitting method.
An extreme has only one neighbour, thus the thickness of the edge must be of one pixel so as to apply, after that, a polynomial approximation.
This method consists in scanning line by line the image and applying a 3x3 mask list to each pixel. There are 21 masks. They put certain directions forward, eliminate lone pixels and are strictly ordered. An example is shown on Figure 1. The reference pixel is the central one. The 6 masks described on Figure 2, put one direction forward. Crosses indicate pixels at 255 and empty cells represent background noise.

\[
\begin{array}{|c|c|c|}
\hline
X & X & X \\
\hline
X & X & X \\
\hline
\end{array} \quad \Rightarrow \quad 
\begin{array}{|c|c|c|}
\hline
X & X & X \\
\hline
\end{array}
\]

Figure 1.

\[
\begin{array}{|c|c|c|}
\hline
X & X & X \\
\hline
X & X & X \\
\hline
\end{array} \quad \begin{array}{|c|c|c|}
\hline
X & X & X \\
\hline
X & X & X \\
\hline
\end{array} \quad \begin{array}{|c|c|c|}
\hline
X & X & X \\
\hline
X & X & X \\
\hline
\end{array}
\]

Figure 2.

4.2 Polynomial approximation.
The aim of this method is to tag, line by line, holes of incomplete edges and to note coordinates of their extremes added to those of a neighbour.
The approximation is going to link the 3 points by a parabolic branch whose coefficients are calculated by the resolution of a 3 equation system.
Fitting was essential to discern extremes and neighbours.
The parabolic approximation does not generate an important error for the boundaries because shapes are small and holes represent few pixels (2 or 3). After this algorithm is executed, if there are extremes left, we realize an image symmetry with the "y=x" axis. The method is then started again.

5. Application on cardiac scintigraphic images.
The images we are dealing with are cardiac scintigraphic images [5]. They are obtained from tests on patients who have been injected with a tracer Tc99m with HMPAO which will localize itself in the heart, and which we detect with a gamma camera. A test board can be seen in Appendix B. The heart cycle is cut into 16 small images.
Each one is a 64x64 pixel large. Dilatation steps are at the first and at the sixteenth image and compression is at the eighth one. The ventricular ejection fraction is calculated with the first and eighth images with the equation:

\[ FEV = \frac{D - S}{D} = \% \]

with D: left ventricle surface in diastole.
S: left ventricle surface in systole.

It must be noted that 60% to 70% is the range of values relative to a good healthy patient. If it is less, the heart is like a pump which has a low efficiency.

6. Results.
Figure 3. and figure 4. show an example of a good healthy patient (VEF=73%).
Figure 5. and figure 6. show an example of a deeply sick patient. There is almost no compression and dilatation (VEF=18%).
7. Conclusion.
The method described above has shown that the results are better than the previous ones but not in all cases because processing an automatic application with thresholds is very difficult.

In order to get better working conditions, this algorithm should be implemented into specialized circuits (ASIC, ... ). Furthermore, from the 16 images, the aim is to get a Fourier development of the left ventricle surface in order to visualize his magnitude and his phase so as to obtain a sector distribution. That is why the algorithm has to be fast.

A better way of optimizing the method would be to include Poisson noise into calculations because until now, it was only tested with a Gaussian white noise.

Acknowledgements
The authors thank ANVAR (National Association for Research Valorization) for a grant to young researchers.

References

Figure 6. Image after treatment


Appendix A
- c(x) is the shape of the chosen edge.
- f(x) is the signal.
- signal to noise ratio
\[
SNR = \frac{\int_{-\infty}^{+\infty} c(-x)f(x)dx}{n_0 \sqrt{\int_{-\infty}^{+\infty} f^2(x)dx}}
\]
- localization
\[
L = \frac{\int_{-\infty}^{+\infty} c'(-x)f''(x)dx}{n_0 \sqrt{\int_{-\infty}^{+\infty} f'^2(x)dx}}
\]
- cancelling of multiple responses
\[
MRC = \left( \frac{\int_{-\infty}^{+\infty} f'^2(x)dx}{\int_{-\infty}^{+\infty} f'^2(x)dx} \right)^{1/2}
\]

Appendix B
Here is a test board. The circular part has got one, two or three cameras and turns around the body of the patient.
Image Sequence Analysis for on line Observation of Thrombus Formation in Blood Vessels

Ralf TÖNJES

Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung,
Division: Automatic Image Interpretation (Prof. C.-E. Liedike)
Universität Hannover, Appelstr. 9A, 30167 Hannover, Germany,
E-Mail: toenjes@tnt.uni-hannover.de

Abstract. An image analysis system for in vivo observation of thrombus formation in skin blood vessels of small mammals is presented. A model based approach is applied, allowing automatic segmentation of thrombus, blood vessel, and tissue in low contrast images. Image sequence analysis takes advantage of preceding results in three ways: data reduction, improvement of algorithm efficiency, and robustness. This is achieved by a fast motion compensation that takes the image movements in the in vivo observation into consideration. For clinical use in a biomedical laboratory a graphical user interface was developed.

Key words: biomedical image processing, image sequence analysis, model based image analysis, motion compensation, real time analysis

1. Introduction

In thrombus research the in vivo observation of blood vessels in the skin of small mammals is used to study thrombus formation. After triggering thrombus formation by a controlled local injury at the inner blood vessel wall the generation and temporal development of the thrombus can be observed under the microscope and recorded via camera. The subsequent evaluation requires the segmentation into the three regions vessel, thrombus, and surrounding tissue. Till now evaluation was done manually, since automatic segmentation of the images suffered from low contrast.

Automation is achieved in three steps:
1. Vessel and thrombus are emphasized by a special mixed illumination with halogen and laser light [1][2].
2. A priori knowledge is integrated in the image analysis by using a geometric vessel and thrombus model.
3. Image sequence analysis is applied to deal with the uncertainties caused by the analysis of single low contrast images.

2. The Model

Two types of a priori knowledge can be distinguished, describing static and dynamic features concerning thrombus formation.

For segmentation (chapter 3) the static features F1 to F3 are relevant:

(F1) Blood vessel and thrombus are three-dimensional. However for the observation of the thrombus formation the 2D projection is sufficient. Figure 1 shows a schematic view of (a) the real vessel and (b) the 2D projection model.

(F2) Due to the short length of the considered vessel segment, the outer vessel boundaries can be approximated by two almost parallel straight lines, named vessel walls in the following. It is assumed, that both vessel walls lie within the image at the beginning of the sequence.

(F3) The thrombus is situated within the vessel at the injured vessel wall and forms a compact region.

Figure 1. (a) Vessel with thrombus and (b) 2D Model
For image sequence analysis (chapter 4) the dynamic features F4 to F6 are applied:

(F4) The illumination conditions vary slowly.

(F5) The temporal change of thrombus geometry is limited by the physiology of thrombus formation and the thrombus position is fixed till it is torn off by the blood stream.

(F6) The geometry of the vessel is constant, but the observed scene exhibits a common translatory movement which is caused by breathing and jerking of the marmal.

3. Segmentation

Goal of the segmentation is to divide the single image into the regions vessel, thrombus, and surrounding tissue. Characteristic features for segmentation of the original image (figure 2a) are:

- brightness differences between vessel and surrounding tissue
- spatial inhomogeneities of intensity in the tissue due to cell contours and speckle pattern caused by coherent illumination.

3.1 Vessel Segmentation

Vessel segmentation exploits the speckle pattern, which is caused by the chosen illumination [2]. The coherent light shows an interference i.e. speckle pattern. However, while taking a picture the blood flow averages the speckle pattern. This texture is tested by a local difference operator and the vessel label is assigned to the largest non speckle region.

The local difference of the grey values \( g_{x,y} \) for the image coordinate \((x,y)\) is measured by a modified roberts gradient \( r_{x,y} \) as defined in eq. (1).

\[
r_{x,y} = \max( |g_{x-1,y} - g_{x+1,y}|, |g_{x,y-1} - g_{x,y+1}| )
\]

(1)

This operator enhances the speckle features better than averaging gradient operators. Further in opposition to difference operators of higher order it avoids suppression of low contrast speckle regions due to overproportional amplification of high differences.

The local difference image is smoothed by low pass filtering. The histogram based thresholding and subsequent shrinking and blowing for smoothing and small blob closing yields a binary image where the largest non background region represents the vessel.

The detection of the vessel walls is made insensitive to false segmentation by applying the tube model (F2). The vessel walls, i.e. contours between the vessel and tissue regions, are approximated by straight lines using linear regression [2].

3.2 Thrombus Segmentation

Thrombus segmentation is primarily model based, because thrombus and surrounding tissue exhibit no significant texture differences. Assuming a tubular vessel (F2) the thrombus is defined by the part of the thrombus–tissue region that projects into the vessel at the injury (F3).

Initially vessel geometry is unknown and thrombus and vessel must be determined together. Segmentation is achieved in two steps. First the free blood vessel, i.e. the non speckle region, is approximated by a tube. Due to the thrombus the tube is estimated too small with a low regression coefficient for the wall at thrombus side. Hence in a second step the linear regression is repeated neglecting the thrombus part of the contour, i.e. the contour projecting into the current tube. The regression is iterated until convergence of the contour approximation is obtained.

Thrombus segmentation suffers from low contrast within the vessel and a solely texture based segmentation approach fails. Hence a weighted sum of the brightness and speckle features is used. The weights are automatically adapted to varying illumination conditions.

The results of vessel and thrombus segmentation are shown in figure 2b with the straight lines approximating the vessel walls and the thrombus being the area beneath the white curve.

![Figure 2](image-url)  
(a) Original image and (b) results of vessel and thrombus segmentation.
4. Image sequence analysis

Image sequence analysis is used to profit from preceding results in three ways: data reduction, improvement of algorithm efficiency, and robustness (table 1).

<table>
<thead>
<tr>
<th>a priori knowledge</th>
<th>profit</th>
<th>required control</th>
</tr>
</thead>
<tbody>
<tr>
<td>intermediate results of previous image analysis</td>
<td>robust computation</td>
<td>verification</td>
</tr>
<tr>
<td>location and contour of thrombus</td>
<td>data reduction</td>
<td>thrombus window</td>
</tr>
<tr>
<td>location and contour of vessel</td>
<td>algorithm efficiency</td>
<td>motion compensation</td>
</tr>
</tbody>
</table>

Table 1. Profit of a priori knowledge for image sequence analysis

Illumination and contrast vary slowly (F4). Thus previous results are used for parameter prediction. The image analysis parameter, which are automatically adapted to varying illumination and contrast conditions, are verified and, when not satisfying the predictions, modified.

The vessel geometry is assumed to be constant (F6), but form, size and position of thrombus vary with time (F5). Thus only the image part including the thrombus must be analysed. The thrombus window is determined by the location of the thrombus in the previous image and a border for growth and movement.

Due to the motion of the animal during the in vivo experiment, a global shifting is superimposed on the image (F6). Thus thrombus and vessel are shifted during the image sequence. To take advantage of known thrombus location and vessel geometry a fast motion compensation is required.

For motion estimation a hierarchic blockmatching algorithm is applied [3]. Because the whole image is shifted uniformly it is sufficient to determine the displacement of a single image section in consecutive images. An image section with characteristic pattern [4] is selected and matched with neighbouring sections within the maximal displacement range of the consecutive image using mean absolute intensity difference as a quality measure. For matching a truncated calculation is applied, that terminates the test of a displacement, if it gets worse than the hitherto optimum. The search is accelerated by a coarse to fine strategy with logarithmic raster reducing computational cost. To increase efficiency further multiple pixels are matched at once by comparing binary images and collecting 16 pixels in a word.

Figure 3 shows the motion compensated updating of the vessel and thrombus window location. From the match windows the displacement of image(t) to image(t−1) is calculated and added to the location of the vessel and thrombus window in the previous image(t−1), thus reducing image analysis to segmentation of the thrombus within a window.

Figure 3. Motion compensation for localization of vessel and thrombus window
5. Implementation and results

The algorithm was implemented on a UNIX operated workstation with a frame grabber. For a SPARCstation 10 processing of 3 frames per second was achieved. Because the image sequence analysis must not be interrupted to wait for user inputs, the user interface is realized as independent process. Data transfer via shared memory is controlled by two communication mechanisms. For inter process communication the UNIX signal mechanism guarantees asynchronous process control and the X-Window selection model provides the update of the windows for image output of the process. The graphical user interface conforms to the OPEN LOOK standard [5]. Figure 4 shows the graphical user interface with windows for process control and visual feedback displaying segmentation results and temporal development of thrombus size and vessel blocking. The thrombus formation graph shows a thrombus growing for 70 s until it is torn off and a new thrombus formation starts.

6. Conclusion

An image analysis system was realized, that provides an on line evaluation of image sequences with in vivo microsopy images showing the formation of a thrombus after a vessel injury. Segmentation of low contrast images is achieved by exploiting a priori knowledge and results from analysis of preceeding images. The latter is based on a fast motion compensation.

Acknowledgement

This work has been supported by a grant of the THOMAE GmbH, which also provided the image data.

References


Pattern Recognition in Normal and Dyskaryotic Cervical Cell Images Using Statistical Texture Analysis

Alan E. ADAMS, Farid HALLOUCHE and Oliver. R. HINTON

Department of Electrical and Electronic Engineering, University of Newcastle upon Tyne, NE1 7RU, United Kingdom.
Tel/Fax +44 [91] 222600/2228180, E-Mail alan.adams@uk.ac.newcastle, oliver.hinton@uk.ac.newcastle

Abstract. Cervical cells are known to be rich in texture information which cytopathologists routinely use for analysis and screening purposes. In cervical cell images, the quantification and modelling of the chromatin texture of well stained nuclei represent an important resource for the automatic recognition of cell condition. In this paper the implementation and application of a method for nuclear texture analysis, based on the Markovian chain process, are discussed. A number of Markovian parameters are computed along the four principal directions. Following a normalisation process these features may be used to discriminate between cell classes. Experiments using a database of 571 samples have been carried out, the aim being to test the mechanism’s ability to discriminate between six cell classes. The results show that the mechanism employed provides a powerful cell discriminator with correct classification rates varying between 70% and 96%.

1. Introduction

Texture analysis has been used for different applications in many areas of science and technology, including the recognition of various types of crop fields and terrain, analysis of the structure of reservoir or soil rocks, analysis of satellite images, analysis of material surface and microstructures and recognition of x-ray and cell imagery [1].

In the analysis of the texture of biomedical imagery and microstructures such as cells, techniques based on the statistical approach seem to be the most often used as reported by Pressman [2] and Haralick [3]. Cervical cells are known to contain important texture information which cytopathologists and histologists use in the screening process. In describing the nuclear chromatin structure, some of the qualitative adjectives used routinely by cytopathologists are fine, coarse, condensed, irregular and clumped. The analysis of cell texture is believed to provide useful cyto-diagnostic markers, and is badly in need of quantification. The reasons behind such quantification are basically to reduce elements of subjectivity and produce automated measurements. Computer-based analysis of texture has the advantage of allowing the objective assessment of the accuracy of the extracted numerical textural features as well as their reproducibility, through, for instance, their statistical classification.

In this paper, a statistical method based on the Markov process, has been developed and applied to cell texture analysis [4]. This method involves the computation of some texture parameters, based on the joint probabilistic distribution of grey levels of neighbouring pixels. The choice of such an approach is based on the knowledge acquired through the visual observation that the overall pattern of the nuclear chromatin texture is such that it can be readily described quantitatively on the basis of the relative arrangement of pixels (i.e. the textural constituent elements or units) along the different directional scan lines. Such a technique has been developed so that it would, in principle, result in global quantitative analysis and modeling of the nuclear chromatin micro-structures. The application of this method is confined to the nuclear region because it is known that the latter holds most of the diagnostically important texture information.

2. Materials and Methods

2.1 Cervical Cell Preparations

In this work, the cervical smear used were those routinely prepared and submitted to the laboratory for visual screening. The material was collected from the uterine cervix with a spatula, usually a wooden Ayre’s or Aylesbury spatula. The material on the spatula was put onto glass and was fixed immediately by immersion in 95% methanol. The smears were then transported to the laboratory where they were stained with Papanicolaou stain, cleared in xylene and coverslipped using DPX. The experimental data base consisted of a total of 571 cell images, of which 113 were squamous superficial (SUP), 103 squamous intermediate (INT), 70 squamous metaplastic (MET), 73 endocervical (END), 103 mild dyskaryotic (CIN1) and 109 severe dyskaryotic (CIN3). The main morpho-densitometric and textural features of these cell types are described below.

The superficial squamous cells are polygonal in shape and are made of abundant and transparent cytoplasm. The nuclei are central, small in size, round, pyknotic and stain very densely. The intermediate squamous cells are similar to the superficial cells, particularly with respect to cell size and cytoplasmic characteristics. The main differences from superficial cells reside in their nuclear properties. The nuclei of the intermediate cells are round or oval in shape, comparatively larger in size than the superficial nuclei and are vesicular in texture, with the chromatin material being finely granular and containing occasional chromocenters. The metaplastic squamous cells are smaller than the intermediate or superficial cells. The nuclear size and nucleo-cytoplasmic ratio are larger for metaplastic than for intermediate cells. The metaplastic nuclei are generally centrally located with round or oval shape, vesicular pattern with finely granular chromatin and small chromocenters. Small nucleoli may also be present in the metaplastic nuclei. Endocervical cells occur occasionally in the vaginal smear, but are more frequent in
the cervical smear or endocervical aspiration. They may appear in smears as single cells or as a sheet of cells. The cell sheets often form a honey comb pattern or a palisade of cells. Unlike the squamous epithelium, which is strong because of its protective function, the endocervical epithelium, which secretes mucus, is fragile and cell morphology is often poorly preserved. Endocervical cells are distinctly characterised by their columnar and tall configuration. They are also known to present eccentric nuclei that usually occupy one end of cell cytoplasm. The endocervical nuclei are characterised by their marked variation in size, particularly between single cells and those in sheets, by their round or oval shape and by their chromatin material that is fine and smoothly granular, regular in distribution and with occasional clumps of chromocenters as well as the presence of nucleoli. Endocervical cells can be mistaken as suspicious (malignant) cells, particularly because of their varying size.

The two other cell types employed in these experimental studies are of the dyskaryotic (pre-malignant) type. These cell groups represent two abnormal grades which are mild and severe dyskaryosis. Mildly dyskaryotic cells, which will be referred to throughout the rest of this text as CN1, have, by comparison to normal cells, a higher nucleo-cytoplasmic ratio, with the abnormal nucleus occupying up to 50% of cell area. The nucleus is usually irregular in shape and the nuclear region is characteristically denser and hyperchromatic, exhibiting various features such as coarse granularity, condensation of material into blobs, chromocenters or strands and uneven or irregular chromat in arrangement, although not very marked.

The severe dyskaryotic cells, which will be called CN3, have some similar properties to CN1 but differ from them mainly in having thicker and denser cytoplasm and more pronounced nuclear abnormalities. The abnormal nucleus fills at least two thirds of the cell area, resulting in a much higher nucleo-cytoplasmic ratio. As far as nuclear chromatin features are concerned, these are more or less similar, though more marked, to those describing CN1.

2.2 Texture Analysis Method

As mentioned above, cell texture analysis was carried out by using a method based on the Markov process. This method works by making use of the frequency of joint occurrence of different possible pairs of nearest-neighbour grey levels present in an image [1]. Two main steps of computation are involved. First, the spatial grey level dependence matrix (also called the co-occurrence matrix) is computed, then a set of statistical texture features are derived from such a matrix.

The co-occurrence matrix is made of elements with each element being the relative frequency \( F(i, j, d, \theta) \) with which two neighbouring pixels are present in the image, one with grey level \( i \) and the other with grey level \( j \), separated by distance \( d \) and along orientation \( \theta \). Thus the joint grey level probabilities in the co-occurrence matrix are functions of the distance separating the neighbouring points as well as of the orientation along which one occurs relative to another. For a given distance \( d \), up to four directional co-occurrence matrices, along one of the main directions (i.e. 0, 45, 90 or 135 degrees), may be computed.

The co-occurrence matrix for an image of 64 grey levels will consist of 64 x 64 elements. The idea behind the computation of the co-occurrence matrix is that such a matrix is assumed to contain the main textural detail, with each of its elements making a contribution towards such extracted information. Obviously this represents a large amount of texture data to analyse, and it will be redundant to describe textural properties merely in terms of these single matrix elements. Hence, with a view to reducing such large data, some further parameters are computed in order to consistently measure and summarise the textural information contained in the co-occurrence matrix. Nineteen statistical texture features which represent different mathematical combinations of the co-occurrence matrix components were employed in this work.

The rationale behind such mathematically derived features is to use all co-occurrence matrix elements in different weighted combinations with a view to extracting and summarising various important textural properties assumed to be contained within such a matrix. While some of these features are known to measure specific texture characteristics such as contrast, linear structure or homogeneity, others are abstract and generally model the distribution, dominant frequencies or complexity of the different grey level transitions present in the co-occurrence matrix. For instance, some of the characteristics in need of quantification are to measure similarity or dissimilarity of different pairs of grey levels and how frequently they appear in an image. This is why measures of similarity, such as \( 1/(1+(i-j)^2) \), and dissimilarity, such as \( (i-j)^2 \), between two grey levels \( i \) and \( j \) have been introduced in some of the texture features. If a grey level \( i \) recurs sufficiently often, a regular texture is produced. This particular case is reflected in the co-occurrence matrix by a concentration of high frequency values on or near the left diagonal of the matrix.

3. Experimental Studies and Results

The different computational steps involved in the Markovian approach for texture analysis have been discussed above. It was mentioned that in the implementation of the co-occurrence matrix two variables were included; one is related to the step size and the other to the angle. The step size, selected in the computation of the co-occurrence matrix, is defined as the number of 'pixel-distances' that separate any two neighbouring points in the image. In this work a step size of one (i.e. the neighbouring points are immediately adjacent to each other) was selected for the computation of the co-occurrence matrix. This decision was made because the nuclear image was relatively small and also the system resolution was not particularly high. Using larger step sizes would result in faster processing but at the risk of missing critical texture features. For the angle variable, it was considered that since the chromatin material distributes in different patterns along different directions then the extraction of textural information along all directions would provide more useful diagnostic information than if a single angle were used. As was mentioned above, up to four directional co-occurrence matrices may be computed. The computation of angle-dependent texture parameters may result in some inconsistency, particularly if the images are captured at varying orientations. To alleviate this, it was decided to compute texture parameters in all directions then normalise these such that they are no longer angle-dependent. This was done by computing, for each original Markovian parameters, four directional features, these are, in turn, used for the derivation of five statistical parameters, which are the mean (MEAN), the standard deviation (STD), the minimum (MIN), the maximum (MAX) and the range (RNG=MAX-MIN). Thus for each of the nineteen Markovian parameters, five statistical values, corresponding to the angle normalisation option were developed, making the total number of textural features ninety-five.
After the computation of the Markovian texture parameters discussed above, for each one of the 571 cervical nuclei, statistical multivariate analysis was then applied for cell classification. The details of the techniques used for cell discriminant analysis have been described in a previous publication [5]. Given the large number of computed parameters and the amount of high correlation existing between some of them, it was deemed necessary to first proceed through feature selection. Feature selection was carried out mainly by applying full stepwise discrimination on the original set of variables, with some further minor adjustments.

<table>
<thead>
<tr>
<th>Experimental Cell Classification</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True Cell Type</strong></td>
<td><strong>SUP</strong></td>
</tr>
<tr>
<td><strong>SUP</strong></td>
<td>94.46%</td>
</tr>
<tr>
<td><strong>INT</strong></td>
<td>03.88%</td>
</tr>
<tr>
<td><strong>MET</strong></td>
<td>00.00%</td>
</tr>
<tr>
<td><strong>END</strong></td>
<td>00.00%</td>
</tr>
<tr>
<td><strong>CIN1</strong></td>
<td>00.00%</td>
</tr>
<tr>
<td><strong>CIN3</strong></td>
<td>09.92%</td>
</tr>
</tbody>
</table>

Table 1. Classification Matrix for Cell Types using QDF

Subsequent to feature selection, discriminant analysis was then applied in order to classify cells in the database using the selected feature model. The covariance matrices were found to be unequal so that a quadratic discriminant function (QDF) was employed. Table 1 shows QDF classification results for the six cell groups. Table 2 summarises the total error values of abnormal and normal cell groups as well as the false positive and false negative estimates.

<table>
<thead>
<tr>
<th></th>
<th>Normal Cells</th>
<th>Abnormal Cells</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Error Rate</strong></td>
<td>0.1242</td>
<td>0.2827</td>
</tr>
<tr>
<td><strong>False Positive</strong></td>
<td>0.1872</td>
<td>0.0453</td>
</tr>
</tbody>
</table>

Table 2. Total Probability of Incorrect Classification for Abnormal and Normal Cells

For graphic display of cell group discrimination, canonical discriminant analysis (CDA) and principal component analysis (PCA), both described in [5] were carried out. Figure 1 shows the plot of the first main canonical variables. Only the first two canonical variables (CAN1 and CAN2) were chosen because they resulted in the highest eigenvalues and thus accounted for the largest percentage of variation between groups; i.e. 57.38% for CAN1 and 19.35% for CAN2.

4. Discussion

The experimental findings presented in this paper may be summarised and discussed as follows:

The scores in the classification matrix indicate that 96.46% of squamous superficial, 87.38% of intermediate, 80% of metaplastic, 86.49% of endocervical, 72.82% of mild dyskaryotic and 70.64% of severe dyskaryotic cells were correctly classified. The overall rate of correctly classifying these six cell groups was thus 82.30%, with the highest true classification rate being attributable to the SUP cells whereas the

![Figure 1. Cell Classification using Canonical Variables](image)
CIN3 cells were the least correctly classified. Table I also shows that 12.86% of METs were classified as INT, 10.81% of END cells were classified as CIN1, 13.59% of CIN1 cells were classified as CIN3, 11.93% and 10.09% of CIN3 cells were classified as INT and MET, respectively. The relatively low correct classification of CIN1 cells is due to the high number of mis-classification of these cells into CIN3 and MET. Similarly, a large number of mis-classifications were shown in the case of CIN3 group due to the unexpected large false classifications of these cell types into INT and MET. It may be worth mentioning that lower error rates than the ones presented here were obtained with the resubstitution method [6,7].

These experimental studies have demonstrated that the extracted Markovian parameters for texture analysis together with the classification methods employed have resulted in relatively high and very promising rates of correct differentiation between the six normal and abnormal cervical cell types. The best feature model made of the selected Markovian textural parameters is specifically valid for the experiments described here, therefore, one must be cautious about making too sweeping a judgement in regard to their general discriminatory powers. Further experimental work, particularly with standardised preparations, will need to be used to determine the reproducibility of the strength of these Markovian parameters, individually and in combination. In general, it is recommended that feature selection will have to be performed for every application, specially if different specimen preparations are used.

The present studies have shown that of the six cell groups, SUP and INT categories were the most highly correctly classified, whereas CN3 and CN1 have shown most overlap between them, and with END and MET cell types. An example of an expected cell overlap is that between INT and MET; this is because it is known that the mature metaplastic cells have some features which are similar to the intermediate squamous cells. On the other hand, an example of an unexpected cell overlap is that of the 11.93% of CIN3 cells classified as INT. Given the measurement and discrimination methods employed in the present experimental studies, the mis-classification rates of some of these cell categories might be attributed to a number of factors. For instance, in some cases, cell discrimination might have been wrongly made on the basis of texture measurements extracted from overlapping nuclei or specimen artefacts (such as objects or dirt overlapping nuclei) rather than from single nuclei. Another important factor which might be a possible cause of the amount of mis-classification reported here is the fact that specimens used in the present experiments were stained using the PAP staining. The PAP staining, unlike the Paulgen staining, for example, is not specific for DNA and thus chromatin material analysis, because it is known to produce variations in the staining results that can lead to inconsistencies in the subsequent related quantitative measurements [8]. On the other hand, the classification percentages of the different cell groups found in these experiments were made with reference to an initial visual screening carried out mainly by the experimenter and then verified by a gynaecologist and a cyto-pathologist. It is possible that during such a reference visual screening, which contains elements of subjectivity, cell selection errors were introduced within the true classes used. Finally, since the Markovian-based texture information is dependent on the frequency of co-occurrence of grey level pairs, it seems that using a higher spatial resolution than the 256×256 images employed in this work might result in better cell classification. On the other hand, it may also be speculated that the textural characteristics of some of these cell groups are such that the Markovian parameters used are not sufficiently powerful on their own to produce any higher cell group discrimination rates. In this respect, it is anticipated that the addition of other textural and non-textural features to the Markovian ones might result in improving the classification rates of at least some of the cell groups.

Acknowledgements

Acknowledgement is made of the contribution to this work of Dr. G.V. Sherbet of the Cancer Research Unit, University of Newcastle upon Tyne and Dr. V. Wadhera of the Dept. of Pathology, Newcastle General Hospital.

References


Image Analysis Tool for Embryo Qualification
Based on Morphological Indices

Lilla BÖRÖCZKY*, Razi DIANAT*, András DINNYÉS#, István ERÉNYI*, Zoltán FAZEKAS*

* KFKI Research Institute for Measurement and Computing Techniques, Budapest, P. O. Box 49, Hungary, H-1125,
Tel: (36 1) 169 6279, Fax: (36 1) 169 5532, E-Mail: h1715@e11a.hu
# Institute for Animal Sciences, Agricultural Biotechnology Centre,
Gödöllő, Hungary, H-2103

Abstract. A new application is presented in the field of computerized microscopy, namely the qualification of mammalian embryos. The objective and reliable evaluation of viability for livestock-embryos is an important step in minimizing the economic risk, when using new bio-technological methods -- such as transplantation of high-quality embryos into low-quality recipient animals -- in husbandry. This evaluation has to be carried out at various segmentation phases of the embryos. In this paper we mainly deal with mammalian embryos being in the 8-cell stadium of their segmentation. In this stage meaningful circles (or circle-arcs) can be fitted to the silhouette of the whole embryo and its constituent cells. Based on these circles morphological indices are calculated. These morphological indices had been used in non-automated viability evaluations and were reported as applicable features (together with other features for different stages) for this purpose.

1. Introduction

As the need for high-quality ("well-engineered") livestock has increased, the transplantation of mammalian embryos became a routine -- but still relatively expensive -- practice in the husbandry. The worldwide trade with embryos of the best quality parent animals has emerged forming a new business branch of biotechnology.

The embryos of top-quality parent-animals are extracted for later transplantation into recipient animals. In order to minimize the economic risk and maximize the likelihood of the future birth of healthy animals, the viability of the embryos selected for transplantation must be checked thoroughly using microscope.

The key issue in the evaluation of embryos is objectivity. It has two major aspects. Firstly, embryos are evaluated 'as they are'. Secondly, observations and measurements have to be repeated in different segmentation stages (the word 'segmentation' is used here in biological sense) and the morphological data of a certain embryo have to be treated as a temporal sequence. One part of the measurements regards the shape of the embryos as a complex, while the other part investigates the individual cells within the embryo.

So far, embryos have been evaluated 'manually' via observation using microscopes. This process is prone to the subjective judgment of the human expert.

Objectivity of the evaluation process may be improved by the usage of 'measurable' indices for each embryo, and also by recording images of different segmentation stages on films or videotapes. Further improvement may be achieved by using computers and computer vision systems, that can -- on one hand -- enhance images, classify and select different parts of the embryo, and so, provide basic support for human evaluators, and -- on the other hand -- do automatic calculation on the indices. Further, digitized and stored images can be later retrieved, and so, the segmentation stages of each embryo can be collated for temporal evaluation, or the evaluation of the development of the segmentation process.

2. Morphological indices for mammalian embryos

For the segmentation-stadia up to the morula stage -- when the number of internal cells stays relatively low -- (not exceeding 16), some applicable and relatively simple morphological indices have been reported in [1,2]. These indices may serve as empirical guidelines for the human experts checking the viability of the embryos.

Only the embryos classified as 'good' should be used in transplantation. The evaluation procedure can be divided into two phases, namely, into a qualitative and a quantitative one. In the qualitative phase, embryos that are 'obviously' dead or degenerated are separated from
live ones. In the quantitative phase, a series of morphological indices are calculated from the measured geometrical properties of the embryos. These indices are:

- the volume of the embryo: \( V_{\text{out}} = \frac{d_{\text{out}}^3 \pi}{6} \)
- the intrazonal volume: \( V_{\text{int}} = \frac{d_{\text{int}}^3 \pi}{6} \)
- the intrazonal index: \( i_{\text{IZ}} = \frac{V_{\text{int}}}{V_{\text{out}}} \)
- the zona pellucida index: \( i_{\text{ZP}} = 1 - i_{\text{IZ}} \)
- the fullness index: \( i_f = \frac{\Sigma V_{\text{cell}}}{V_{\text{int}}} \)
- the perivitelline sine index: \( i_{\text{PVS}} = 1 - i_f \)

where \( d_{\text{out}} \) and \( d_{\text{int}} \) are the outer and the inner diameters of the embryo, respectively, and \( V_{\text{cell}} \) is the volume of the internal cells. In the original figure and figures obtained during processing, one can see the structure of an 8-cell embryo. The embryo is covered by a special thin membrane, then outside of that, by a stronger membrane. This area is called zona pellucida. Within the embryo, the cells can be observed. The microscopic images contain some irrelevant details, such as hidden edges caused by overlapping cells. The translucidity of the embryos makes distinguishing between the background and the embryo difficult.

3. Enhancement and segmentation of images

The microscopic images often are of rather poor quality. Computerized image processing may make the 'manual' evaluation much easier in the way of enhancing the original microscopic images and providing interactive tools to image segmentation and classification of the details in the images. This eliminates errors caused by poor quality, irregularity of shapes, and the partially invisible shapes. Errors may result in calculating misleading indices and incorrect classification. So, a procedure has been developed to support qualification and eliminate evaluation errors. We note, that in our experiments we used images of mouse embryos which are similar to the livestock embryos in the early stadia of segmentation and are remarkably cheaper than the bovine ones.

For our experimental research work, we selected segmentation stadium with 8 cells in the embryo. In this stage, the structure of the embryo already is complex ('matured') enough and so, sophisticated qualification algorithms can be worked out and tested for checking the viability.

The analysis of the image starts with a preprocessing phase. Firstly, the embryo has to be localized in the original image (Figure 1.a). For this purpose, we have adopted a non-linear filtering technique to enhance the image, together with a windowing technique (originally used for localizing white blood cells in blood-smears [3,4] and also used in localizing pollens in honey-smears [5]). This windowing technique uses a priori information available about the objects appearing in the image (in the present case the size of the embryos is in a well-defined range). Then after median filtering, edge detection and histogram equalization, the enhanced image is obtained (Figure 1.b). This image is much more informative for the human experts.

Further help may be offered for the experts in a way of finding and segregating the components of the embryos. In our research work, we developed methods to find the 'border lines' of the whole embryo and its cells. This process starts with an automatic 3-class thresholding, after which the resulting picture contains the contours of the outer membrane (zona pellucida) and also the cells inside the embryo (Figure 2.a). To find a finer structure of the constituent cells, a further thresholding and skeleton operation obtains the 'contour' lines of the shapes (Figure 2.b). Visual observation of these two images may offer a closer look at the regularity of shapes and the size of the cells.

Figure 1.a. Original image of the embryo  
Figure 1.b. Enhanced image of the embryo
More precise estimation and evaluation of the formation of the internal cells and also of the zona pellucida is provided via segmentation of the cells and zona pellucida. This can be achieved via 'opening the outer circle' and then labeling the components, after which geometric filtration is applied to filter out irrelevant areas (small objects). In this way, the obtained image contains the internal structure i.e. the cells of the embryo, where cells are labeled by different gray levels (Figure 3.a). Separating the background via simple thresholding from the cells we can get a mask, with the help of which the cells may be masked out from the enhanced image, leaving there only the zona pellucida (Figure 3.b). Irregularities in the shapes and size of the cells and also the whole embryo can be observed and evaluated in an easy way on the base of the classified components or objects. The segmented objects may also provide a basis for index calculation (though another way is described in Chapter 4.). Note, that these images may be used for calculating other indices, as well, i.e. shape indices: area, circumference, their ratio, etc. The relevance of these indices to the vitality of the embryos has not been checked yet in the practice.

4. Extracting morphological indices for embryo-qualification

An automated procedure has also been developed for this stadium of segmentation and is under development for other relevant stadia. Morphological index extraction process starts with the same steps as it is described in Chapter 3. After the steps of fast segmentation scheme via automatic 3-class thresholding, the coordinates of the contour points are extracted. The two essential components of the mentioned image are the outer membrane of the embryo and the individual cells inside the embryo. Ideally, the outer membrane should look like a ring (with circles of inner and outer radii) and also the cells should be of equal size and should appear as circles. In reality, it is just an approximation.

For determining the center points and radii of the circles fitting to the contour of the embryo and of the individual cells, we have selected the algorithm described in [6]. This algorithm provides the best fitting (in mean square sense) circle, using some points of the original curve. (Note, the algorithm allows to use various number
of points within a wide range). Obviously, the algorithm is sensitive to the appropriate selection of the representing points, if only a few points of the curve are used. That is why we decided to choose those points, if that is possible, where there is an abrupt change in the curvature. We mention that an other possibility for determining the radii and the centers of the fitting circles is based on generalized Hough-transformation.

If the curve-circle fitting is relatively good, then one may also consider the mean square errors calculated during the radius-estimation as a further morphological feature, since it characterizes the uncurliness of the curve. Based on the measurement and calculation data, the embryo structure can be reconstructed as it is shown on Figure 4.

Figure 4.a. Reconstruction of the internal cells

Figure 4.b Reconstruction of the Zona pellucida

5. Conclusions

In this paper we present an image analysis tool which can be integrated into complex, interactive or fully automated embryo-qualification system for embryos being in various stadia of segmentation.

For interactive systems, the embryo qualification image processing aid allows the human expert to modify the procedure in an adaptive way, adjusting it to the quality, size, etc. of the original image, that may be obtained using various microscopes. To improve and widen the services of the system, further work should be done in three main directions:

- man-machine communication system, to make the aid more user-friendly, not requiring much image-processing knowledge from the user,
- testing the system in practice by experts and finding how much relevance is between other morphological features and the viability of the embryos,
- works on archiving the results of processing and evaluation to offer ways for the evaluation of the whole temporal process of embryo segmentation.

For the automated index calculation system, a circle-fitting technique has been used to estimate radii and the centers of the embryos. Using these data, we could calculate the approximate volume of the embryo and its cells and also the empirical indices used for morphological description of embryos. Further work is planned on using other techniques, e.g. ring fitting proposed in [7], or Hough transformation.

Acknowledgments

This work is supported by Hungarian Science and Research Fund contract number: F4065 and PHARE Accord Programme, contract numbers: H 9112-0265/266.

References

Texture Analysis of Ultrasonographic Endoscopy Images, Using the Master Classifier Method

Piotr WILIŃSKI†, Basel SOLAIMAN†, Christian ROUX† and Michel ROBASZKIEWICZ††

†Ecole Nationale Supérieure des Télécommunications de Bretagne, dept. Image et Traitement de l’Information, BP 832, 29 285 BREST Cedex, FRANCE Tel/Fax : +33 98 00 13 08/98 00 10 98
E-Mail Piotr.Wilinski@enst-bretagne.fr
††Centre Hospitalier Universitaire de Brest, 5 av Foche, 29 200 BREST Cedex, FRANCE

Abstract. This paper presents a new segmentation method based on the cooperation of several segmentation algorithms. Those algorithms are seen as independent classifiers and their cooperation is completed in the decision domain. This new method is applied to the segmentation of textured images, especially in the segmentation of Ultrasonographic Endoscopy images. In this problem, standard algorithms do not give satisfying results, because of poor image quality. Therefore, it is necessary to combine methods. Obtained results show the superiority of the proposed cooperation method, compared to standard solutions like the "winner takes all method". To show the possibilities of the new technique it is also applied to the segmentation of standard textured images.

1. Introduction

The ultrasonographic endoscopy is an advanced method giving grey scale images of internal organs. It was developed for an early detection of esophageal cancers. Endoscopic images permit to observe several internal organs, like the esophagus, the heart, lungs, the windpipe, lymph nodes. The main role of this technique is tumor detection, appearing the most often between esophageal walls. The physician's role is to estimate the danger of tumor expansion for other organs, mainly for the lymph nodes. To this end, the disease extension should be measured. The diagnostic is mainly based on texture differentiation. The present work is a part of a larger investigation program leading to the automatic region labelling, using different textual, geometrical and syntactical methods.

2. The problem.

Quantitative ultrasound tissue characterization has shown promising results in detection and monitoring of the pathology of tissue. In medical imaging, one of the most important constraints is the diagnosis accuracy. In the ultrasonographic endoscopy case several tests are performed before deciding the appropriate treatment. If the illness is growing fast it may be inevitable to immediately operate the patient. To make a diagnosis, the tumor surface should be measured with a high accuracy. In this objective, taking into account the poor quality of ultrasonic images, new segmentation methods are investigated.

The main idea of the procedure presented in this work, is to assure the visibility of several details in the segmented image. Various methods of segmentation exist [1]. The nature of images, and textures which differentiate organs, suggest the use of statistical methods [2][3]. In successful applications, there is always a trade-off between the resolution and the statistical power of the regions of interest. To compute statistical parameters it is necessary to use large pixel masks [4]. If we want those parameters to be robust, at least 5x5 pixels masks should be taken into account. On the other hand, image details have often 3-4 pixels size. Masks for which the statistical parameters are computed should be therefore 3-4 pixels large, and could not give enough information to a satisfying parameter estimation.

3. The texture analysis methods.

Texture analysis is considered here as a pattern recognition problem. One of the assumption of this work is that in ultrasonic images it is possible to distinguish textures [2][3], which can be classified using statistical methods. Several textural parameters were tested. The best ones are the parameters issued from the modified grey level covariance matrix which were proposed in [4], the mean and the variance of the grey levels, the variance and the contrast of the coccurrence matrix and the variance of the dissimilarity matrix. To obtain satisfying results at least 6 parameters vectors were used to describe each pixel.

4. Proposed solution.

In our problem it is proposed to consider the texture analysis (and hence the image segmentation) as a classification problem. Different classifiers may be used, each one performing the segmentation efficiently in one direction [5]. Different windows shapes: vertical, horizontal, diagonal and anti-diagonal ones serve to 4 different pixel classifications. The problem is to compare their decisions to form one global segmentation decision. The proposed method considers each pixel separately and not regions of pixels forming textures. Each classifier gives in each pixel an independent decision (figure 1).

5. Master Classifier Method

A new method of cooperation, called the Master Classifier Method (MCM) [6] consists of computing the a posteriori probabilities of each classifier decision, and then of forming a global a posteriori probability decision model. Pixels which membership knowledge is certain are defined. For those pixels, after a first recognition, which is called the Master Classifier Learning (MCL), the probabilities of each kind of false decision are computed. There may be two main reasons for false classification. The first one, may be caused by the spatial neighbourhood of different classes. It is obvious that a classifier makes mistakes on the limits between adjacent textures. The second reason may be the proximity in the feature space. This kind of errors is due to the classifiers weaknesses. The MCM can be also seen as a neural method based on computing of posterior probabilities of each decision, conditioned on output values and output decision. In fact, each
classifier gives two kinds of information: the decided class label, and the highest output value. Two separate steps of this algorithm can be distinguished. The first one corresponds to the learning phase (MCL), in which the probability that the i-th output should be used as a value describing the decision of class j, knowing that output i is the highest one, and knowing its value is computed. It can be designed as: \( P(\omega_j / (D_k \cap D_k)) \),

where \( \omega_j \) is the input class label, \( D_k \) is the value of the output i for classifier k, and \( D_k \) is the decision given by the i-th output of that classifier. It is possible to stress a correlation between the decision quality, the output value and its label. In the recognition phase the global a posteriori probability of each class is computed by summing the probabilities of partial events:

\[
P(\omega_j^{pos} / X_m^I) = 1 - \prod_{k=1}^{n} (1 - P(\omega_j / (D_k \cap D_k)))
\]

(1)

where \( \omega_j^{pos} \) is the a posteriori probability of class \( \omega_j \), and \( X_m^I \) is the m-th input vector belonging to class j.

To apply this formula the assumption of the independence of separate methods has to be made. In the present application the same method operating on 4 independent feature sets is used. The MCM gives also the possibility of cooperation between different classification methods. This is due to the fact, that it uses probabilities

Figure 1. Proposed experiment

Figure 2. Classification using different mask forms.

Figure 3. Comparison between the “winner takes all” method and the MCM.
describing each single classifier decision and not directly output values. In the presented application, separate classifiers are applying the same paradigm, but it is possible to imagine applications where the cooperation is done between completely different classification methods.

6. Experiments
To test the MCMs abilities to achieve the cooperation between several segmentation methods, it was first tested on a simple case of natural textures segmentation. An image composed of five different Brodatz textures, respectively: grass (D9), herringbone weave (D15), wood grain (D68), woolen clothes (D19) and pressed calf leather (D24) was synthesized. Those textures present some similarities, and their classification is not easy. The segmentation results are shown in Figures 2 and 3. First, four images present the segmentation obtained using directional masks, respectively: the horizontal, the vertical, the diagonal and the antidiagonal one. Each grey level corresponds to one class. Figure 3 presents the results given by two cooperation methods: "the winner takes all" method and the MCM. The first one can be considered as a reference. It consists of choosing the decision of the classifier which has the highest output. In this case, it means that the classifier which gives the pixel highest potential imposes its own decision.

Second, MCM was applied to ultrasonographic endoscopy images. To avoid texture damage it was decided to avoid any preprocessing. Several images were processed. It was noticed that textures were varying even in the inside of a single organ. To cope with that problem, several learning windows were defined for some textures. Another problem was the choice of different classes. In endoscopic images besides classes like esophagus, esophagus wall, veins, lymph nodes and tumors, which correspond to objects we are detecting, it is necessary to define some other ones, like saturation zones where the image luminosity is high and texture information is often lost, dark zones which are generally situated in parts of image which are far from the centre, and cells lying between different organs. In order to obtain satisfying results, at least 6 or 7 classes should be defined. Segmentation results are presented in figures 4 and 5. At the left side are situated original endoscopic images. The segmentation is carried out only in one part of the image. It is not necessary to do it in regions where there is no relevant information to explore. In the future, an automatic algorithm will choose regions of interests, which are supposed to contain relevant informations. First, 20x20 windows containing only one class of texture are defined. Such windows are shown in right upper images. To each pixel is assigned a texture parameter vector which is classified using the potential method: a single pixel belongs to the class which creates the highest potential in this point of the parameter space. Four independent classifications using different mask forms for computing in parallel textural parameters are achieved. Then two methods of cooperation between those classifiers are applied. Results of the "winner takes all method" are presented in the second image in the left column. Each grey level represents a different class. In the right bottom corner the results of the MCM are presented.

7. Results discussion and conclusions
The MCM has been initially developed as a cooperation strategy taking into account several classifiers. It was first applied to optical characters recognition [6]. In this study, texture analysis is considered as a pattern recognition problem. This allows us to use the MCM to improve texture recognition issued from several classifiers.

The MCM was applied on two different kinds of images. First, standard images containing different textures were processed. In that case separate methods gave respectively

![Figure 4. Esophagus image and its segmentation](image-url)
81.72%, 72.92%, 76.04% and 75.63% of recognition (Figure 2), the majority vote 80% and the MCM 85.54% (Figure 3). The classification imperfections were mainly due to the fact that in the classification process the neighbourhood information was not taken into account.

Second, endoscopic images were segmented. Their quality varies significantly, and the segmentation results depended strongly on several factors like the mode of their storage, endoscope device tuning, and even the illness degree. The best results were obtained for images digitized directly at the output of the endoscope. On those images directional methods gave 69.8%, 71.65%, 67.67%, 63.94% of recognition, the winner takes all method 67.63% and the MCM 73.22%. Such results permit to evaluate the external limits of tumors and lymph nodes, but not to measure their surface. This problems is caused mainly by the fact that in endoscopy images each organ can be represented as a mixture of textures. Therefore, segmentation methods assign to a single class pixels, several labels. That is the case for tumors and lymph nodes. It can be noticed that the MCM increases the rate of the correct class labels, but can not totally solve the mixture problem. To achieve a high accuracy segmentation a cooperation with other methods detecting specific forms and object contours should be done.

Compared to other cooperation strategies the MCM presents several advantages. For example compared to the "winner takes all" method it gives 8-10% of gain in terms of recognition rate. It uses always the result of all separate methods [7]. Taking into account even rejection and false decision, all disposable information is used.

References


Figure 5. Another example of esophagus image and its segmentation

891
Transient Detection by a Time-Scale Representation Applied to Biological Signals

E. Çağatay GÜLER†, Bülent SANKUR, Yasemin P. KAHYA, and Tanju ENGİN††

†Biomedical Eng. Institute, Boğaziçi University, Bebek-80815, İstanbul, Turkey, Tel: +90 (212) 263 15 41-1441, E-Mail: gulerca@irboun.binet
Dept. of Electrical Engineering, Boğaziçi University, Bebek-80815, İstanbul, Turkey, Tel/Fax: +90 (212) 263 15 41-1455, (212) 257 50 30, E-Mail: sankur@irboun.binet, Tel/Fax: +90 (212) 263 15 41-1881, (212) 257 50 30, kahyay@irboun.binet
††Faculty of Electrical Eng., Istanbul Technical University, Ayazaga, Istanbul, Turkey, Tel: +90 (216) 344 00 44-6737, E-Mail: eengin@iteu.binet

Abstract. A novel method for the detection of transients in biological signals is proposed. The method is based on the augmentation of the transient-to-background ratio by the use of time-frequency analysis in association with nonlinear operators. The signal is decomposed into M octave bands via the discrete wavelet transform and these components are first processed separately, and then combined to form a time-frequency landscape. Transients are identified as significant edges in this landscape. The intended application of the method is the detection of crackles in respiratory sounds. The performance tests indicate that the proposed detector is superior to other detectors as applied to the specific case of respiratory transients.

1. Introduction
Detection and analysis of transient events in biological signals, e.g., bursts, abrupt changes, discontinuities, local nonstationarities, continue to be one of the important issues in signal processing. Abrupt changes or discontinuities in these signals may be symptomatic of functional disorders, or may be instrumental in obtaining qualitative and/or quantitative assessments. For example, information regarding the presence or the stage of a pathology or disorder can sometimes be obtained from the shape of the transient, its duration, epoch and arrival rate during a cycle.

In this study we address the problem of detection of transients in a colored noise background. To this effect we propose a novel method based on time-frequency techniques, that yields accurately the epochs and wave shapes.

The advantages of the time-frequency techniques are as follows:

• Time-frequency analysis of a signal avail us of the possibility to observe the signal characteristics both in the time and frequency domains.

• Time-frequency techniques may be used jointly with nonlinear operations for the augmentation of the transient event-to-background ratio.

• Decomposition of the signal spectrum into its subbands allows to operate on separate resolution levels, in which the amount of nonstationarity is reduced. Furthermore each subband can be treated individually, so that the subbands or the combination of subbands yielding best the transients can be used.

• Time-frequency representations can be interpreted to stand between nonparametric and parametric analysis techniques. Most time-frequency representations are nonparametric. However, a priori knowledge on the signal process can be coparated on these representations as in the case of matched wavelet banks.

As a specific example of biological transients, we have chosen to analyze crackles in pathological respiratory sounds. Respiratory crackles are attributed to sudden bursts of air within bronchioles, and they can be perceived against the background of normal breath sounds. Crackles are localized in time, and they do not have a significant bearing on the total spectrum. Auscultation of these transient sounds are utilized by physicians to assess the type and the stage of the pathology in the airways. Their duration is usually not more than 70 ms, and they have a spectrum of frequencies between 100 and 2000 Hz. Figure 1 shows a sample pathological respiratory sound waveform containing 12 crackles.

We have assessed the performance of our transient detector as applied to crackles comparatively with respect to two other crackle detection methods proposed in the literature, respectively, by Ono et al. [1] and Kaisla et al. [2].

In the Ono method transients are separated from the background using a nonlinear filter that separates the stationary from the nonstationary parts. The filter is a prediction error filter (PEF) tuned to the background respiratory sounds, hence, mismatched to the transients. The output of the filter is thresholded so that only the PEF samples above the threshold are considered to correspond to transients [1].

The detection method proposed by Kaisla is based on monitoring deviations from the short-time stationarity of the signal. In other words transient peaks in the sonogram of the signal are detected by searching significant ridges in the time-frequency landscape [2].

In our method based is on the time-frequency analysis of biological signals via the discrete wavelet transform (DWT). The DWT is well suited for the analysis of nonstationarities (e.g., abrupt changes, discontinuities) in signals.

Other examples of signals bearing transients and that can be analyzed with our detector are Doppler measurements of the blood flow and detection of embolies, EEG signals and evoked response analysis, and microevents in the speech signals.

2. Preliminaries: The Discrete Wavelet Transform
In the discrete domain, wavelet transform can be regarded as a mathematically formalized subband coder and usually implemented as a bank of bandpass filters. In this approach, the

![Figure 1. A pathological respiratory sound containing 12 crackles. Crackles are indicated by arrows.](image-url)
original signal space $V_0$ is decomposed into two orthogonal subspaces $V_1$, $W_1$ using analysis and synthesis filters [3] (Figure 2). In Figure 2(a), $f_L$ and $f_H$ are typically FIR lowpass and highpass analysis filters respectively, and $\hat{\uparrow}2$ denotes downsampling by a factor of two. Similarly, $s_L$ and $s_H$ represent projections of $s(n)$ into two, respectively, low-resolution, $V_1$, and high-resolution, $W_1$, subspaces, while $\hat{s}_L$ and $\hat{s}_H$ are their analysis stage counterparts. In Figure 2(a), the synthesis stage filters are denoted by $g_L$ and $g_H$, and $\uparrow2$ represents upsampling by a factor of two. If the analysis and synthesis stages are symbolically denoted by the hatched boxes (Figure 2(b)), then the dyadic splitting of the signal into $M$ bands can be shown as in Figure 3. Here, the choice of $M$ depends on the signal record length and the resolution requirement for the signal under analysis.

In the structure given in Figure 3, $M$ bandpass detail signals, $s_{Hm}$, $m = 1, ..., M$, and the blurred lowpass component, $s_{Lm}$, of the original signal in each $W_m$ and $V_M$ spaces are obtained by closing only one switch at a time. If all switches are closed simultaneously, the original signal is perfectly reconstructed at the output. The analysis and synthesis stages are shown in Figure 3, separately, for the sake of clarity, so as to allow for example for further processing of either the decimated signals, $s_{Hm}$, or their synthesized versions, $s_m$.

In the sequel, for the simplicity in the notation, we will denote the ($M+1$) subband signals by $s_m$, $m = 0, 1, ..., M$, i.e., $s_m = s_{Hm}$, $m = 1, 2, ..., M$, and $s_0 = s_{L}$, i.e., the lowpass components in the $M$th stage. For such a decomposition we will use the notation: $DWT\{s(n)\} = \{ s_0, s_1, s_2, ..., s_M\}$.

### 3. Transient Detection Method

The transient detection method is based on the separation of transients from the background using the DWT in association with nonlinear operators. By decomposing a signal into wavelet components, the transients are put more into evidence in matching passbands. In other words, the use of the DWT as a bank of matched filters enhances the transient-to-background ratio. The number of wavelet components is judiciously chosen depending upon the signal bandwidth and frequency location of transients. The block diagram of the detection method is depicted in Figure 4. The steps of the transient detection method are as follows.

#### STEP 1. Conditioning: Prewhitening and Time-Frequency Expansion

A prediction error filter (PEF) tuned to the background is applied to the measured signal, $s(n)$, in order to decorrelate the background, while preserving as much as possible the nature of the abrupt change phenomena, that is transient waveforms that do not belong to the background activity. The biological signal, $s(n)$, is assumed to consist of the sum of a short-time stationary background which can be modeled by an autoregressive (AR) and various nonstationary components (e.g., transient events).

Since the AR predictor, $\hat{s}(n)$, is fitted to represent the stationary components of the input signal, the prediction residuals are significantly larger in parts where the signal switches from the stationary background to a transient phenomenon. On the other hand, the PEF tends to decorrelate the background component. The error signal, $e(n) = s(n) - \hat{s}(n)$, is thus assumed to be the sum of a decorrelated background signal and filtered transient signals.

The predictor error signal, $e(n)$, is then expanded in the time-scale space using the DWT:

$$DWT\{e(n)\} = \{ e_0, e_1, ..., e_M \} \quad (1)$$

where $e_m$, $m = 0, 1, ..., M$, are the decimated wavelet components. In the analysis we only consider the detail components, $\{ e_1, ..., e_M \}$.

#### STEP 2. Background Suppression

In the second step, Teager's energy operator [4] is applied to each of the $M$ detail signals (e.g., to $M$ wavelet components), $\{ e_1, e_2, ..., e_M \}$, in order to suppress background activity. Teager's energy operator is an energy function which is a very local property of the signal that depends only on the signal and its first two derivatives. For the discrete case, Teager's function, $\varepsilon(n)$, of a given signal $\varepsilon(n)$ is defined as

$$\varepsilon(n) = \varepsilon^2(n) - \varepsilon(n+1)\varepsilon(n-1). \quad (2)$$

Thus, with the application of Teager's operator to all wavelet components one obtains $\{ \varepsilon_1, \varepsilon_2, ..., \varepsilon_M \}$.

Steep changes in the signal amplitude are amplified while smooth behaviour is attenuated by applying the Teager's operator. Thus signal behaviour which is not attenuated by the PEF, and hence supposedly, corresponds to the transients phenomena, are put further into evidence by the Teager's operator. In this method cross terms are assumed to be negligible.

#### STEP 3. Time-Frequency Landscape of Wavelets

In this step, full scale waveforms, $\tilde{E}_m$, $m = 1, 2, ..., M$, are reconstructed selectively from the "Teagerized" wavelet components, $\varepsilon_m$, using the synthesis part of the structure depicted.
Figure 4. Block diagram of the transient detector. PEF: prediction error filter, DWT: discrete wavelet transform, TEO: teager's energy operator, IDWT: inverse discrete wavelet transform, TFL: time-frequency landscape, RED: roof edge detector, APF: all-pole filter.

In Figure 2. In other words, inverse DWT (IDWT) of each \( \tilde{e}_m \) is taken such that,

\[
\tilde{E}_m = \text{IDWT}\{ \tilde{0}, \ldots, \tilde{\tilde{e}}_m, \ldots, \tilde{0}_M \},
\]

where \( \tilde{0} \) represents a complementary null vector of size \( 2^N \) corresponding to the \( j \)th non-switched wavelet component. Thus \( M \) signals each of length \( N \), and, respectively, in the frequency intervals \( [2^{-m}B, 2^{1-m}B] \), are obtained, where \( B \) is the bandwidth of the original signal, \( s(n) \).

These \( M \) signals are plotted in the scale (frequency) vs time plane, in other words, as a scalogram, a time-frequency landscape is obtained. This landscape is constructed as an \( M \)-by-\( N \) matrix, \( I(m, n) \), from the absolute values of the synthesized subband signals. This time-frequency landscape matrix is as follows:

\[
I(m, n) = [ |E_1(n)|, \ldots, |E_M(n)| ],
\]

Since transient phenomena are expected to show up on more than one scale, they are expected as vertical striations in the scalogram. In other words, vertical edges in this "image" are identified as transients. The center of gravity of the edge in the scale axis is an estimate for the scale of the transient, and the edge location becomes the time of occurrence of the transient.

Localization of transients in time is achieved by a line finding (edge finding) algorithm on the scalogram image. The line finder mask, \( C \), is shown in Figure 5.

The edge enhanced image, \( I_{PR}(n) \), is obtained by convolution of the scalogram with the mask along the \( n \) variable:

\[
I_{PR}(n) = I(m, n) * C(m, n)
\]

(5)

The scalogram image, preemphasized with the line finder mask and collapsed into a 1-D function, is then thresholded with \( C_{TH} \):

\[
I_{TH}(n) = \begin{cases} 
I_{PR}(n) & \text{if } I_{PR}(n) \geq C_{TH} \\
0 & \text{if } I_{PR}(n) < C_{TH} \end{cases}
\]

(6)

The nonzero values of \( I_{TH}(n) \) are indicative of the transient locations. At this stage there are two alternatives to proceed with the transient marks in order to recover the exact transient waveforms, \( T(n) \). Either \( I_{TH}(n) \) is put through an inverse PEF filter, or the original signal \( s(n) \) is windowed with \( I_{TH}(n) \). The first alternative results in

\[
T(n) = - \sum_{k=1}^{P} a_k T(n - k) + I_{TH}(n),
\]

(7)

while the second alternative is a cut-and-paste approach on the signal described as follows:

\[
T(n) = \begin{cases} 
s(n) & \text{if } I_{TH}(n) > 0 \\
0 & \text{if } I_{TH}(n) = 0 \end{cases}
\]

(8)

4. Results and Discussion

This section presents the experimental results of the proposed transient detection method as applied to respiratory crackles.

The value of the AR model order, \( P \), used to implement the prediction error and all-pole filters of the detector is chosen as 2, as it was shown that low model order values are adequate for modelling respiratory sounds [5]. The AR coefficients are calculated using the Burg method.

The signals are expanded in the time-frequency space using the DWT. The number of resolution levels, \( M \), was chosen as 5. In other words, since the signal bandwidth is 2500 Hz, signals are analysed in five bands: 1250-2500 Hz, 625-1250 Hz, 312.5-625 Hz, 156.25-312.5 Hz, 78.125-156.25 Hz. The lowpass subband component, i.e., \( s_p \), is ignored since it corresponds to the 0-80 Hz frequency band which was filtered out to avoid the heart sounds and muscle noise interference. The applied wavelet in the analyses is the Daubechies filter [7] with 6 coefficients. The length and one-sided, asymmetric nature of lengths 6 Daubechies filter suit well to crackle phenomena.

The value of the threshold, \( C_{TH} \), used to localize transients in the time domain is evaluated automatically using the "Otsu method" [7].

Application of our method to the signal, \( s(n) \), shown in Figure 1 is illustrated in Figure 6. The error signal, \( e(n) \), at the output of the second order prediction error filter is given in Figure 6(a). The image, \( K(m,n) \), i.e., the time-frequency landscape of \( e(n) \), obtained from the wavelet components, \( m = 1, 2, \ldots, 5 \), is depicted in Figure 6(b). The output of the detector, \( d(n) \), is depicted in Figure 6(c). In this example, the detector has located 12 of 12 crackles existing in the waveform correctly with no erroneous detections.

Figures 7(a) and (b) show the outputs of Ono and Kaisla detectors for the same input signal (Figure 1), respectively. As can be seen from Figure 7, the Ono detector detects 12 of 11 crackles correctly and misses 1 crackle, while Kaisla detector detects 12 of the 12 crackles but gives 3 false alarms.

Performance measurements were conducted using 10 pathological sound records containing in total 123 crackles. For the validation of the method, an expert observer counted crackles from the time-expanded waveforms of signals. The numbers of crackles detected by our detector, and by the Ono and Kaisla detectors were thus compared. For detection performance we used two parameters; namely the sensitivity (S), and the positive predictivity (PP) parameters. Figure 8 gives the performance test results of three detectors. The sensitivities and positive predictivities of our detector, and of the Ono and Kaisla detectors with respect to the observer are (95.45%, 90.52%, 82.03%), (90.51%, 78.36%, 86.78%), respectively.

Performance comparisons indicate that the three detectors perform close to each other. Our detector gave better results, for the following reasons: (a) Nonlinear operators enhance the transient-to-background ratio in the individual bands, (b) The filter used in the DWT analysis matches more or less transient waveforms. The other detector methods rely only on the local nonstationarities of transients. We observed that our detector detects not only individual respiratory transients, but overlapping transients as well.

The Ono and Kaisla detectors cannot handle overlapping transients. Also the Ono detector, while yielding successfully the
first two or three deflections of crackles, fails to represent the late deflections. On the other hand the Kaisla et al. method fails to detect low amplitude crackles since it assumes that the crackles have higher energy content compared to background. Finally if the length of the time window does not match the duration of crackles, background components are also detected within a window. In other words, the improper choice of the length of the window causes false alarms.

5. Conclusion
In conclusion, we have described a novel clinically applicable transient detection scheme based on the use of time-scale analysis in association with nonlinear operators. This detector is then tested on crackle transients encountered in pathological respiratory sounds. Using the method, nearly all of the respiratory transients, which were identified visually, have been correctly detected with their exact wave shapes.

Figure 7. (a) The output of the Ono detector for the input signal in Fig 1, (b) The output of the Kaisla detector for the input signal in Figure 1. Missed crackles and false alarms are indicated by arrows.

Figure 8. Performance comparisons of three detectors.

Acknowledgment
This work has been supported by the Boğaziçi University Research Fund.

References
Stochastic Model of Awake and Sleep EEG

Agostinho C. da ROSA, Ana L. N. FRED and José M. N. LEITÃO

Instituto de Telecomunicações / Dep. Eng. Electrotécnica e de Computadores
Instituto Superior Técnico, Lisbon, Portugal
Tel/Fax: +351 (1) 8419361 / 3923014, E-Mail: d1812@beta.ist.pt

Abstract. The main features of a stochastic model of the electroencephalogram (EEG) are presented. The model parameters characterize the tonic activity (basic brain rhythms) while the phasic events are taken as responses to excitatory impulses.

An optimal detector of K-complexes and Vertex sharp waves, based on the stochastic model, is considered and briefly characterized.

Another application consists in converting the EEG signal into a sequence of symbols which form the input data of a higher level syntactic EEG description.

1 Introduction

The electroencephalogram (EEG), being the voltage evolution sensed at an array of electrodes located (for instance) on the scalp (with respect to a common reference) is commonly used to assess the brain activity. Some functional or cognitive states can be clearly read from the EEG. Other working patterns are only detectable and monitored if adequate signal processing and recognition methods are applied; these, to be efficient, demand some prior knowledge of the processes to be estimated and of the events to be detected.

The EEG signal is generally viewed as a combination of two different components (activities), see [1]):

- The background or ongoing activity is a combination of several rhythmic signals (resonant generators) with slowly varying amplitude and relatively stable frequency contents, known as delta (.5 - 4.0Hz), theta (4 - 8Hz), alpha (8 - 12Hz), sigma (12-14Hz) and beta (14 -20Hz) rhythms.

- The phasic activity is composed by several types of short duration signals or transient responses (phasic events). They are regarded as specific responses of the brain to internal or external stimuli. In awake EEG most of these responses are not visually discernible. Averaging the repeated evoked responses is the usual method to enhance them. On the other hand, the transient responses in sleep are in general clearly visible and stimuli independent, i.e. the amplitude, spatial distribution and morphology of the responses are relatively insensitive to the origin (internal or external) and to the type (acoustic, visual and somatosensory) of the stimuli. Phasic events like K-complexes and Vertex waves are very common in sleep recordings, there are several hundreds per night with a mean density of 2 events per minute. These events appear spontaneously (internal stimuli) or may be evoked (external) by sound, flash, light pattern or electrical stimuli.

This paper summarizes the main features of a stochastic model of the EEG (single electrode) and shows how it can be used in detecting phasic waveforms while tracking the basic tonic rhythms.

Furthermore, this model is also used to convert the EEG message into a sequence of symbols. At a higher level, a syntactic model, based on stochastic context-free grammars describes the temporal organization of these symbols.

2 Model Description

The model herein presented takes into account the two activities (tonic and phasic) simultaneously. It has been developed in [2], being essentially a refinement and generalization of the one studied in [4, 3] which is based on a blend of neurophysiological data [5] with pragmatic engineering criteria. As shown in figure 2, its main structure is a bank of feedback loops, each one formed by a second order bandpass filter (\(h_d(t), h_o(t), \ldots\), with the band limits of the corresponding EEG rhythm) with a variable gain (\(K_d, K_o, \ldots\)). This gain models the amount of respective activity. A first order lowpass filter (\(h_p(t)\), with cutoff frequency of 2Hz) at the output models the frequency selective attenuation effect on the EEG from cortex (\(e(t)\)) to the scalp (\(e(t)\)).

The input of the model has two components: a white gaussian noise process, \(n(t)\), responsible for the background activity; and a random sequence of impulses, \(p(t)\), representing the internal or external afferent stimuli of the evoked responses or phasic events.
3 Model-based Estimation and Detection

3.1 Estimation

Figure 2 shows the estimator / detector model-base structure. The cortical EEG, \(c(t)\), is obtained from the observed scalp signal, \(e(t)\), by passing it through the inverted filter modeling the transfer function between the cortex and the scalp \((h_{sp}(t))\). For analytical simplicity, signals \((\delta(t), \theta(t), \ldots)\) are represented as \((r_1(t), r_2(t), \ldots)\). For practical digital implementation the observed scalp EEG is sampled (100 samples/second) and an equivalent discrete model is derived. Considering only the tonic activity, the problem is to estimate the vector of gains \(k = [k_1, k_2, \ldots]\) and the variance of the driving sequence \(\sigma_n^2\).

Since there is no a priori knowledge about the model parameters, a maximum likelihood (ML) estimation criterion is adopted.

\[ r_{nj} = \begin{bmatrix} r_1(n) \ldots r_n(n) \end{bmatrix}^T, \quad Q_n \text{ is the covariance matrix of the noise and } m_n \text{ is the vector of the noise mean value.} \]

The discrete input signal \(\bar{v}_n\) can be reconstructed by using the following relation:

\[ \bar{v}_n = c_n - \hat{k}_m r_n^T \]

where \(c_n(k)\) is the cortical EEG.

3.2 Detection

The detection of the stimulus is based on the detection of impulses in the reconstructed input signal \(\bar{v}(n)\) as given by (2).

A Bayesian criterion may be used; the test statistic is given by the log likelihood ratio [6]

\[ \ln(lr) = (n\mu^2)/2q_d^2, \]

where \(\mu\) is the mean and \(q_d\) is the variance of the reconstructed input process.

The impulses may also be regarded as deviation of gaussianity in the input process. These deviations can be detected by a Chi-squared test. This test compares the normalized sum of a set of squared samples of the input to a threshold which depends on the significance level desired. The power of the test depends on the number of the observations in the set (degree of freedom - df) which has to be adapted to the average duration of the deviations from gaussianity.

\[ \chi^2[\bar{v}_n(k)] = \frac{1}{q_d} \sum_{j=0}^{df} \bar{v}_j(k)^2 \]

A practical application of the model is the development of a model-based detector of K-complexes and Vertex waves. The detector is formed by two stages. The first stage is the ML reconstruction of the input signal \(\bar{v}(t)\), as described previously, based on a fixed number of samples selected according to the typical duration of rhythms (4 secs for delta and 2 secs for the rest). The second stage is the detection of the impulses in noise using a combination of Chi-square and Bayesian tests.

In the analysis of performance of the two detection methods and its combinations it is found that the best solution for the detection of clinical meaningful k-complexes is obtained by the geometrical mean of the statistics of the two tests [2]. This procedure is needed to filter out artifacts and unavoidable contamination of the EEG signal induced by body movements and other electrical sources as eyes movements, muscles and electrocardiographic activities.

The performance of the detector in stages 1 and 2NREM is characterized by a mean detection rate of 90% and a mean false positives of 13% in sleep recordings from a population of 15 normal subjects, using as reference the visual classification of two specialists.

4 Simulations

The model and the model-based estimator / detector (i.e. the analyser) have been evaluated by extensive
simulations and applied for long periods in several situations. By recording the evolutions of the estimated parameters and the localization of the detected impulses, the analyzer builds up a representation of the original signal. Since the estimated parameters remain constant for long periods, only the significant changes have to be stored; the result is a significant data reduction. Thus, besides estimation and detection, the analyzer also has encoding properties. The artificially generated EEG by this process exhibits all the variability and characteristics of the real EEG during sleep in different brain states. This is shown in figure 4 where short (15 seconds) EEG epochs corresponding to stages 1, 2, 3 and 4NREM (see below) are shown. Notice the resemblance of delta tonic activities and the practical coincidence of original and reconstructed spindles and K-complexes.

![Figure 3: "Original" ("Or") and "reconstructed" ("Re") signals in different sleep stages.](image)

A further validation of the model is the generation of a night (8 hours) of sleep EEG, using a sequence of gains and input impulses obtained by the same method described above. The visual classifications in sleep stages according to Rechtschaffen and Kales (R&K) [7] of the real and artificial signals are shown in figure 4 in terms of hypnograms (i.e. graphical descriptions of stage evolution). The R&K classification considers basically 6 stages: Awake ("W"), Rapid Eye Movement sleep (REM) and four stages of non REM sleep (NREM) with increasing degree of deepness (1NREM, 2NREM, 3NREM, and 4NREM).

The main disagreements of the two classifications are found: in epochs were artifacts are present; and in the lack of discrimination between stages REM and 1NREM when using only one channel of the EEG in the classification process. Due to the resemblance of the EEG signal in these two stages, the scoring criteria needs the information provided by other physiological signals like the electromyogram (EMG) and the electrocugram (EOG).

5 Hierarchical Modeling of Sleep EEG

Another application of the model is the description of the EEG as a sequence of symbols. The signal is initially segmented in fixed epochs (ex: 0.5 to 1.0 sec.). For each epoch the log-likelihood ($LL_m$) ratios of the different rhythms are estimated using a single loop model. In this case the expression (4) reduces into

$$\hat{R}_m = \frac{\sum_{j=0}^{n} r_{j}^m c_j}{\sum_{j=0}^{n} r_{j}^m r_{j}^m}$$  \hspace{1cm} (5)

and the corresponding likelihood ratio is

$$LL_m = \sum_{k=0}^{n} \left[ r_{k}^m c_k - \frac{1}{2} c_k^2 \right]$$  \hspace{1cm} (6)

The ratios are subject to a variable multi level discretization according to clinical and physiological relevance criteria. Segments of the same discretized levels in all the LL ratios are linked as variable length epochs.

Figure 5 shows an epoch of 15 seconds of EEG displayed along with the discretized LL ratios for delta, theta, alpha and sigma rhythms. The last trace shows the linked segments and the corresponding symbols.

![Figure 5: Conversion of the EEG signal into a sequence of symbols.](image)
According to this design, the stochastic model (level 1) is used for the extraction of features, which form the basis of a symbolic description of the original signals (level 2). At a higher level, a syntactic model, based on stochastic context-free grammars, describes the temporal organization of symbols.

![Diagram of Hierarchical Model of Sleep EEG]

Figure 6: Hierarchical model of sleep EEG.

Reference [8] concerns the application of this methodology to the prediction of entrance in REM sleep from stage 2NREM. The good results obtained further motivate the adaptation of the hybrid pattern recognition system to a broader range of applications in microstructure analysis.

6 Concluding Remarks

The model is based on the assumption that the Evoked Potentials during awake and the Phasic Events in sleep are essentially transient responses of the same underlying neuronal structure to impulsive type stimuli in different brain states. The form, amplitude and duration of the response depend on the characteristics of the stimuli and mainly of the brain state that is here characterized by the state of the parameters of the model.

The basic structure of the model is obtained by linearization and simplification of a nonlinear distributed parameter physiological model. Due to this physiological background the parameters of the model have close relationship with meaningful clinical and neurophysiological variables. These characteristics enabled realistic simulations of the EEG of normal sleep not only in the different sleep stages but also the sleep dynamics of entire (8 hours) night sleep signal.

The model is simple but sufficiently accurate for the development of high performance model-based estimators and automatic detectors for clinical and research applications. The model-based K-complex detector output performed the existent detectors reported in the literature and is currently in use for routine tasks.

The segmentation method based on a decoupled model is simple and efficient. The method was applied successfully as a pre-processor of a hierarchical model of sleep microstructure.

References


Automatic Left Ventricular Boundary Extraction in Echocardiographic Images using Neural Networks

Iain A Hunter  John J Soraghan
Signal Processing Division, Dept of Electronic and Electrical Eng., University of Strathclyde, Glasgow G1 1XW
Scotland, UK. Tel/Fax: +44 [0] 552 44000/552 2487, E-mail: lain@uk.ac.strathclyde.eee.sdp

Abstract. This paper presents a fully automatic implementation of the radial search algorithm for the extraction of both epicardial and endocardial boundaries in short axis mid-papillary muscle level echocardiographic images. The centre for the radial algorithm in the LV is found using scale-space neural networks and then another neural network is used to detect inner and outer edge points in the image along the radial search lines. Snake energy functions are then minimised using dynamic programming to link the correct edge points on each radii to form the closed epicardial and endocardial boundaries.

1 Introduction

Two-dimensional echocardiography is an important non-invasive clinical tool for cardiac imaging. The epicardial and endocardial boundaries of the Left Ventricle (LV) can provide useful quantitative measures of cardiac function such as ejection fraction and wall motion. To quantify these measures the LV boundaries must be defined, a process usually done manually.

The extraction of these boundaries by standard image processing techniques is complicated by the image's inherent low spatial resolution, high level of speckle noise and the absence of myocardial boundaries in regions where the cardiac muscle wall is parallel to the incident ultrasound beams. The most common algorithm for boundary detection is the Radial Search algorithm. In this each point on a set of equiangular radial search lines emanating from the LV centre is evaluated as a candidate boundary point. The detection of boundary points has been carried out using neural networks [1] and matched filters [2]. These candidate points are then transformed to the polar domain to be linked into closed contours. The radial search method and similar methods involving cartesian to polar transformations have the advantage of reducing the search space to one dimension along each search line from the two dimensions in the image.

All of the radially based methods involve the location of the centre of the LV, in most implementations this has been supplied manually. Automatic detection of the LV centre has been done using of matched filters [3]. This paper presents an improved version of a published radial search algorithm [4], in which two cascaded Artificial Neural Networks in a scale space approach are used to find the centre of the LV cavity. This centre is then used as the centre of a radial search algorithm in which another ANN is used to identify possible inner and outer edge points on each radii. Two snake energy functions are then minimised by dynamic programming to give the epicardial and endocardial boundaries.

2 Neural Networks as Statistical Classifiers

The neural networks used are of the well known feedforward type trained using backpropagation [5]. It has been shown that a feedforward neural network can be trained to approximate a Bayes optimal discriminant function for that classification problem, with the outputs from the neural network giving the a posteriori probabilities for each class. The a priori probabilities of each class presented in the training set are implicitly assumed to be equal to the relative number of patterns from each class in the training set [5] and so there must be an equal number of vectors from each class in the training set.

3 System Architecture

The system comprises three main modules as shown in figure 1, namely LV centre detection, LV edge detection and contour extraction.

3.1 Centre Detection

This takes place on images reduced in resolution by 32:1 and 16:1 by pixel averaging. It uses neural network classifiers to give the probability of each pixel in these images being the LV centre. The inputs to these networks are windows which are scanned across the images, for the 32:1 image a 9x9...
square window with 81 elements was used and for the 16:1 image a 17x17 star window with 72 elements (figure 2). The network topologies used were 81-2-2 and 72-2-2 respectively.

![9x9 Square Window for 32:1](image)

![17x17 Star Window for 16:1](image)

Input pixel being classified by input vector is marked in black

Figure 2: Input vector configurations

The networks classify the input vectors as either feature vectors (LV centre) or background vectors (Not LV centre). Previous work [4] found that the 32:1 image centre classification was more robust, with the 16:1 images having the potential for more accurate detection but that they were liable to erroneously detect spurious centres. The use of scale-space allows the more robust 32:1 detection to guide the 16:1 detection to give a more accurate centre location. The centre at 32:1 is used to define a 4x4 pixel Region Of Interest in the 16:1 image that is searched for a centre. The centre detection classifies the pixel in the image with the highest probability as the centre. The results are given in table 1, the RMS error is given in terms of pixels in the full resolution image.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMS error</th>
<th>Centres Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>32:1 only</td>
<td>20.1</td>
<td>1010</td>
</tr>
<tr>
<td>16:1 only</td>
<td>23.7</td>
<td>1010</td>
</tr>
<tr>
<td>Scale-space</td>
<td>20.4</td>
<td>1010</td>
</tr>
</tbody>
</table>

Table 1: RMS error (in full resolution image pixels) in LV centre location and number of centres found for different methods

### 3.2 LV Edge Detection

The LV Boundary Detection module consists of a neural network to select candidate LV boundary points. In a radial search algorithm each pixel along a radial line is examined as a possible boundary point. Boundary point detection is carried out on a 2:1 subsampled image. Low pass filtering was not used, as it was found experimentally not to improve edge detection. Each pixel on the N=60 radial search lines (6 degree resolution) is then classified by a neural network into one of three categories: non-edge, inner edge or outer edge. The input to the network is a 7x7 window centred on the pixel of interest which is rotated with the radial line so that it's orientation to it remains constant. This feature enables the network to differentiate between inner and outer edges. An alternative input vector configuration was also tried with the radial angle coded as 7 additional identical elements (total of 56). The image window pixel grey scale values of 0-128 and the radial angle number of 1-60 were both normalised into the range -1 to 1.

The training set consisted of 360 vectors (120 of each class) and the test set consisted of 240 vectors (80 of each class). The network parameters used for training were momentum \( \mu = 0.2 \), and learning coefficient = 0.7.

Four different network topologies - 49-3-3, 49-5-3, 56-3-3, 56-5-3 - were trained and tested. The results shown in tables 2-3 are the ensemble average results from 5 runs of each topology. The classification of outer and inner edges used a threshold of 80%, however the classification of an input vector as background relied on the vector being neither an inner nor an outer edge.

<table>
<thead>
<tr>
<th>Vector Type</th>
<th>Classification - 49-3-3</th>
<th>Classification - 49-3-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outer</td>
<td>59 0 21</td>
<td>59 0 21</td>
</tr>
<tr>
<td>Inner</td>
<td>0 70 13</td>
<td>0 70 13</td>
</tr>
<tr>
<td>Bkgd</td>
<td>8 72 8</td>
<td>8 72 8</td>
</tr>
</tbody>
</table>

Table 2: Test results for 7x7 network with 3 and 5 elements in hidden layer:

<table>
<thead>
<tr>
<th>Vector Type</th>
<th>Classification - 56-3-3</th>
<th>Classification - 56-3-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outer</td>
<td>63 0 17</td>
<td>59 0 21</td>
</tr>
<tr>
<td>Inner</td>
<td>0 66 14</td>
<td>0 65 15</td>
</tr>
<tr>
<td>Bkgd</td>
<td>10 69 10</td>
<td>8 71 8</td>
</tr>
</tbody>
</table>

Table 3: Test results for 7x8 network with 3 and 5 elements in hidden layer

The results show that the best combined classification of inner and outer edges was given by the 7x8 network incorporating the radial angle with three elements in it's hidden layer (56-3-3 network).

Where the network detects the same type of edge on a number of adjacent pixels on a radial line then the pixel with the highest probability is selected as the edge point. These possible edge points for each image are stored in terms of their polar coordinates in a distance-angle plot, with an inner edge stored as a 1 and an outer edge as a -1.

### 4 Contour Extraction

A "snake" is an energy minimising spline or deformable contour which is attracted by the image features it is searching for, while being subject to internal forces which regulate its development in terms of its curvature. A snake model represents the contour as a series of connected expanding or contracting inhomogeneous membranes/thin plates. The energy of the snake is given by [6]:

\[
E_{snake} = \int_0^1 E_{int}(\mathbf{v}(s)) + E_{ext}(\mathbf{v}(s))ds
\]
\[ E_{\text{int}}(v(s)) = \int_{0}^{1} \alpha(s)|v'(s)|^2 + \beta(s)|v''(s)|^2\,ds \quad (2) \]

\( E_{\text{ext}} \) represents the image energy and \( v(s) \) is a parameterisation of the contour. The internal energy, \( E_{\text{int}} \), of the snake model is a function of \( v'(s) \) and \( v''(s) \). These two terms give the stretching energy of a membrane and the bending energy of a thin plate respectively and they are weighted by the parameters \( \alpha(s) \) and \( \beta(s) \). \( E_{\text{ext}} \), represents the energy of the image which is used to attract the contour to salient features in the image. This is achieved by processing the image in such a way that the desired features in the image produce energy wells.

The epicardial and endocardial boundaries are both represented by snakes. These attempts to link correct boundary points in the polar matrix with smooth contours which will interpolate between the detected edge points, in the regions in the image where no edges were found.

The snake's total energy is minimised using the discrete dynamic programming (DP) algorithm of Amini et al [7]. This carries out a comprehensive search of \( m \) possible locations at each of \( n \) different points in the 2nd order snake to find the global energy minimum with that search area.

The snake is used as a simplification of the general model in that the points in the snake are constrained to move in only one dimension along their radial line and that each radial line must have only one point in the snake with the radii being wrapped round at the edges.

An energy map of the edges, \( I(r(n), n) \), is created by smoothing the polar matrix with a 5x5 Gaussian mask with \( \sigma = 1 \) [8]. \( I(r(n), n) \) gives the image energy acting on a point in the snake at a radius \( r(n) \) on radial line \( n \). This allows neighbouring edge points to reinforce each other. The external energy of the snake which will drive the snake towards the detected edge points is given by:

\[ E_{\text{ext}} = \gamma \sum_{n=1}^{N} \delta_n I(r(n), n) \quad (3) \]

where \( \delta_n \) is a term used to select the type of edge being looked for on each of the \( N \) radial lines and \( \gamma \) a weighting parameter.

If an outer edge is being looked for then \( \delta_n = 1 \) and if an inner edge \( \delta_n = -1 \). This term causes the desired edge type to form an energy well and the other an energy peak. \( \delta_n \) can be defined using a priori knowledge of the physiology of the LV, the endocardial boundary is composed only of inner edges whereas the epicardial boundary is mainly inner edges but has outer edges in the region of the IntraVentricular Septum (IVS).

The initialisation of the snakes to the approximate boundary points is essential to allow them to converge to the optimum boundary. The endocardial boundary is initialised to a circle whose radius is given by the radius of the innermost edge points at the top of the LV and the epicardial to one given by the outer edges on the IVS. The centre of these circles was that determined by the centre detection process. The DP minimisation for each boundary starts from the radial line each was initialised from. It also incorporates a hard constraint to keep the two boundaries at least 5 pixels apart [7].

One of the problems that the snake has to solve is to create a contour which will interpolate between edge points over several radial lines with no edge points. This will do successfully where the radii of the edge points to be interpolated between are similar. If they are of significantly different radii then the snake has trouble linking these points with a straight line. This is because the snake only tries to minimise it's energy over 3 adjacent radii (the 2nd order energy term), not the total length of the gap (figure 3).

![Figure 3: Minimum energy state of the snake when interpolating between edge points of significantly different radii](image)

This means that the snake is good at selecting the correct edge points to link when they are present, but it cannot always interpolate in the required manner. The solution to this problem was to define any edge point that the snake passed within 2 pixels of, as being a point on the contour. The full closed contour is then obtained by using simple linear interpolation to link these selected contour points if necessary, so that each of the 60 radii has one point on the contour. This closed contour is then median filtered with window width 3.

To minimise the computational requirements of searching a region covering 6 pixels on either side of the snake position (\( m_i = 13 \), a two stage process similar in concept to scale space is used. Initially the dynamic programming searches every second pixel on each radii to find an estimate of the minimum contour (\( m_1 = 7 \), then a finer scale search is carried out (\( m_2 = 3 \)) to get the final contour. This is possible because the polar edge matrix is smoothed by a Gaussian filter to create a continuous energy map. With the DP being of order \( O(m^3) \) this 'scale space' approach reduces computation time by a factor of 6.
5 Results

The detected boundaries shown here were created using snake parameters $\alpha = 0.4$, $\beta = 0.2$ and $\gamma = 1$. The boundaries in figure 4 each took 6 seconds to calculate on a Sun Sparc 10.

Figure 4: Results of boundary detection in 4 different images

6 Discussion

The detection of the LV centre using scale-space techniques and neural networks provides a robust and sufficiently accurate method of centre location. This is a vital component of any fully automatic boundary detection system.

The use of the Neural Network as the edge detector allowed both weak and strong edges to be classified, this is important with echocardiographic images with their large variations in edge strength. This gives all edge points in the energy map the same magnitude, whether weak or strong, which is a help with the snake energy minimisation as the strength of the attraction of the edge point to the snake is not therefore dependant upon the magnitude of it’s gradient. The energy map with it’s inner and outer edges improves the final contour produced as the snake is both forced away from the wrong type of edge point as well as being attracted to the correct type. It also allowed a priori physiological knowledge to be applied to define the type of edge points expected on each radii for both contours.

The detection algorithm has been developed for short-axis mid-papillary muscle images, but it will be equally applicable to other views with the only difference being in the model used to initialise the snake.

Acknowledgements

This work has been supported by Science and Engineering Research Council, UK. Studentship No 92314900.

References


Conformation Radiotherapy Optimisation
Using Distributed Simulated Annealing

Susan J. SCARTH, William A. SANDHAM, Rosemary WILKINSON

Dept. of Medical Physics and Medical Eng., University of Edinburgh, Edinburgh EH4 2XU, Scotland, U.K.,
Tel/Fax: +44 [31] 332 2525 / 315 2919, E-Mail: sscarth@festival.edinburgh.ac.uk
1Signal Processing Division, Dept. of Electronic and Electrical Eng., University of Strathclyde, Glasgow G1 1XW, Scotland, U.K.,
Tel/Fax: +44 [41] 552 4400 / 552 2487, E-Mail: was@spd.eee.strath.ac.uk
1Bioengineering Unit, University of Strathclyde, Glasgow G4 ONW, Scotland, U.K.,
Tel/Fax: +44 [41] 552 4400 / 552 6088

Abstract. A major technique for the treatment of cancerous tissue is external beam radiotherapy, in which a lethal dose of radiation is supplied from a number of photon beam sources, in order to destroy the tumour cells. Conformation radiotherapy is a relatively new technique in which the distribution of an absorbed radiation dose within a patient is optimised to conform to the target region. Amongst the available optimisation methods using this new approach, simulated annealing was chosen in the present work to determine the optimal beam profiles. This technique is very computationally intensive, and therefore parallel processing technology was investigated in order to reduce run times to a level acceptable for clinical use. A treatment planning system was written to run sequentially on a Sun Sparcstation as a benchmark. It was then implemented on a Transputer system and was successful in producing almost identical treatment plans in a much shorter time. Speed-up factors of up to four were obtained using nine Transputers compared to a Sun Sparcstation. This work has demonstrated that parallel processing may be used to reduce run times for conformal radiotherapy planning, whilst still maintaining high quality results.

1 Introduction

Radiotherapy Treatment Planning (RTP) is the procedure carried out before treatment commences, in which the distribution of an absorbed dose within a patient is predicted and optimised. The objective is to determine a beam configuration that delivers a lethal dose to the tumour while not exceeding a limited dose to the surrounding healthy tissue. There are two limitations placed on the models if they are to be used clinically. First, the model must give accurate dose distributions and second, the implementation must be fast and interactive. Conformation or conformal RTP is a technique which tries to match exactly the high dose region produced by a set of external beams with a prescribed target volume. The beam shapes are said to "conform" to the projection of the target at each orientation. The technique is especially useful for concave regions and areas where it is important not to irradiate surrounding delicate tissue such as the spinal cord.

In conventional RTP, a visual optimisation method is usually employed. The patient data (CT scans, body profiles etc) is obtained and a set of beam data is selected. The resultant dose contours are then calculated. If these contours are not acceptable, the configuration of the beams is modified slightly and a new set of dose contours are calculated. This is repeated until an acceptable set of dose contours are obtained. Conventional RTP, which is a solution to a forward problem, is essentially a trial and error approach, and the optimal dose distribution is not usually computed in clinically acceptable times.

Conformation RTP is a deterministic technique in which the starting point is the ideal or prescribed dose distribution. From this distribution the beam profiles are numerically derived. The technique addresses the inverse problem. At each gantry angle the beam portal can be shaped to geometrically conform to the target region. Multi-leaf collimators or elementary beam scanning techniques can be used effectively to shape the beam portal. Brahme [1] first introduced the idea of varying the radiation intensity across the beam profile as well as the portal shape at every angle. This produces improved correlation between the high dose region and the target volume. In practice, beam compensators can be used to modify the intensity profile.

2 Simulated Annealing

An optimal treatment plan can be determined using a systematic approach in which the whole solution space is searched, or by a stochastic approach such as simulated annealing. This is an optimisation technique which enables local minima or maxima to be transcended, by simulating a thermal system slowly being cooled from above its freezing point until it achieves its ground state. The basis of the technique is the Metropolis algorithm [2] which simulates a collection of atoms at a given temperature. The system is modified slightly and the change in energy, \( \Delta E \), is calculated. If \( \Delta E \leq 0 \) then the modification is accepted and the new state is used in the next calculation. However, if the modification causes the system to have a higher energy than the previous state, then it is not necessarily rejected but may be accepted with a probability that is dependent on the current temperature. The probability of acceptance is given by:

\[
e^{-\frac{\Delta E}{kT}}
\]

where \( t \) is the temperature and \( k \) is Boltzmann's constant. The acceptance of worst states prevents the solution from becoming trapped in local minima and allows the true global minimum to be found.

This mode can be adapted for a variety of optimisation problems. A cost function for the system is defined which is a measure of "success" of the system and is the equivalent of the energy change in the thermal model. The optimisation procedure aims to minimize this cost function. Random changes are made to the system and
each state is accepted or rejected according to the above equation. A
temperature variable is assigned which determines the probability of
acceptance. At the start of the optimisation procedure the probability of
accepting worse states is quite high, but as the temperature is
reduced this probability is reduced also. The effect of this is to enable a
course search through the solution space at high temperatures, with a
finer search near the optimal solution.

It has been shown by Kirkpatrick et al [3] that this method is much
more computationally effective than performing a global search of
all the solution space.

Webb [4, 5] has used this method extensively in conformation
RTT.

3 Implementation

In this project only a two-dimensional treatment plan was consid-
ered. This assumes that the treatment volume is uniform in the axial
direction. It is also assumed that the tissue is homogeneous, which
is acceptable if the cross sections are taken from the lower abdomen
where there are no significant inhomogeneities. However, it would
not be an acceptable assumption in the chest area due to the presence
of lung tissue. The optimisation of the beam profiles was performed
using the simulated annealing technique.

The software was written initially in ANSI C to run sequentially
on a Sun Sparcstation 1+ (typical clock rate 25 MHz). A parallel
version of the code was then written and implemented on a Melko
computing surface using up to nine Transputers with each Trans-
puter having a typical clock rate of 15 MHz. The Communicating
Sequential (CS) Tools environment was used.

3.1 Dose Calculation

The geometry of the problem is illustrated in Figure 1. The dose
prescription \( D_{i,j} \) is the desired dose to the body where \( i \) and \( j \) define
a unique position along the \( x \) and \( y \) axes.

![Figure 1: Body cross-sectional geometry in two dimensions.](image)

The beams are equi-spaced between 0 - 2\( \pi \). Let \( B_{k,m} \) represent
the weighting for the \( k^{th} \) element in the \( m^{th} \) beam orientation. The
dose \( D_{i,j} \) delivered to a point \( (i,j) \) is dependent on the distance
\( d_{i,j,m} \) from the point to the skin surface along the line of the beam
element. The dose \( D_{i,j} \) consists of the primary dose \( D_{i,j}^p \), and the
scattered dose \( D_{i,j}^s \).

\[
D_{i,j} = D_{i,j}^p + D_{i,j}^s
\]

For the purpose of demonstrating the optimisation technique, scatter-
ing effects and inverse square law effects were ignored, and only
attenuation of the primary beam was considered i.e.,

\[
D_{i,j} \approx B_{k,m} e^{-\mu d_{i,j,m}}
\]

where \( \mu \) is the narrow beam linear attenuation coefficient for x-
rays. \( \mu \) is dependent on the energy of the beam, and for 6 MeV
x-radiation \( \mu = 0.4 m m^{-1} \). The depth of the pixel from the skin
surface along a particular beam element is given by \( d_{i,j,m} \).

A beam element will deliver a dose to all the pixels in the body
that geometrically connect to that beam element. The total dose to
a point in the body will then be the sum of all the dose contributions
from each beam. If the total number of beams is \( NBEAMS \) then the
total dose delivered to one pixel is

\[
DT_{i,j} = \sum_{m=1}^{NBEAMS} B_{k,m} e^{-\mu d_{i,j,m}}
\]

For clinical applications this model would have to be extended. The
dose calculation is performed only once in the program and then
the dose for each point is stored. In the optimisation procedure, the
dose for each pixel that connects to a beam element is multiplied by
the beam weighting factor \( B_{k,m} \). Hence, to extend the model, for
example to include inhomogeneities, would not result in a significant
increase in overall runtime.

3.2 Algorithm

Patient data was obtained from a CT scan, consisting of a target area
with a required dose prescription of 100 units, two critical organs
which can have a maximum dose of 15 units, and normal tissue
which can have a maximum dose of 30 units. Simulated annealing
tries to match these values exactly rather than treating them as the
upper limits.

Each beam consists of a number of parallel beam elements with an
associated weight which represents the radiation delivered by that
element. The beam weights are initially set to zero. The beam
elements are then grown by perturbing them with a discrete amount,
known as a radiation grain.

Optimisation is carried out using the following procedure based
on simulated annealing. At each iteration a beam and element are
chosen randomly using a random number generator. A grain of
radiation is added to this element. The grain can be either positive
or negative. However, the weighting of a beam element cannot be
negative, and so if the addition of a grain were to make the overall
weighting negative, the grain is not added.

As each grain is added, a new dose array is calculated. This
enables a cost function to be determined. The cost function is the
modulus of the difference between the prescribed dose and the actual
dose in the tumour and organ regions. The treatment planner is
primarily concerned about the dosage to the tumour region and any
critical organs. However, overdosing to the normal tissue should
also be avoided. This is accounted for by including the square
root of the difference between the prescribed and the actual dose
in the cost function. The revised cost function is then compared to
its previous value, and the revision is accepted according to the
Metropolis algorithm.
As the temperature is reduced, so is the grain size. This is in order that the system rapidly approaches a rough solution at high temperatures, whilst refinements to the solution at low temperatures are made. The starting temperature was such that $kT = 10^5$, which is sufficient for half the positive changes in the cost function to be accepted. It was important to ensure that an adequate number of iterations occurred at each temperature step, so that each beam element had a significant chance of being sampled. At each temperature a check was carried out to determine whether the optimisation should terminate. This will occur if the predefined maximum number of temperature steps is exceeded or if the system is in the ground state i.e. if the potential does not change beyond a certain tolerance level.

3.3 Parallel Version

The computation time for each iteration is a small fraction of the time for the total optimisation, since a single iteration is small and unevenly spread throughout the data set. The complete optimisation is computationally expensive because of the number of iterations required to produce the optimal result. Since there are many iterations, any parallel implementation should avoid communication between processors at each iteration. There are two reasons why the conventional grid decomposition method is undesirable. First, it results in poor load balancing and second, it results in a large number of global communications and hence increases run times.

It was therefore decided to parallelise the optimisation itself by reducing the number of iterations that have to be performed sequentially. The basic idea is to allocate a processor to a beam (or a section of a beam) and each processor works on optimising the profile for its own beam. After a certain number of iterations, the beam information for each processor is communicated to all the other processors, which are arranged in a ring. The optimisation proceeds using this up to date information as a starting point for the next set of calculations.

Each processor is provided with a different seed for the random generator. Since the processor always updates its own beam there is not the random selection of a beam as in the sequential version, however there is still a random selection of an element within the beam.

For the present work, an identical number of processors and beams have been used. This is not necessary of course, since all of the elements in a beam are independent of one another, and a number of beams or a portion of one beam can be allocated to one processor, depending on how many processors are available.

4 Results

Figure 2 shows the beam profiles and dose distribution for a treatment plan involving 9 beams with 60 iterations per beam. The beam profiles are shown orientated around the body, and the critical organs are shown in white. The edge of the tumour is shown in white but the centre has been left clear so the dose inside can be seen clearly. The profiles produce a plan which minimizes critical organ dosage as much as possible, whilst maintaining a uniform dose within the tumour itself.

The timings of the parallel and sequential versions for a total number of 315 iterations per temperature are shown in Table 1. It should be noted that the speed-up increases as the number of beams (and processors) increases. However, the efficiency of the implementation decreases as the ratio of the speed-up to the number of processors decreases.

![Figure 2: Beam profiles and associated doses with 60 iterations per beam: 9 Beams.](image)

<table>
<thead>
<tr>
<th>Number of Beams</th>
<th>Parallel Time(min)</th>
<th>Serial Time(min)</th>
<th>Speed-up</th>
<th>Speed-up Beams</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>53.7</td>
<td>120.6</td>
<td>2.25</td>
<td>0.75</td>
</tr>
<tr>
<td>5</td>
<td>45.3</td>
<td>126.8</td>
<td>2.80</td>
<td>0.56</td>
</tr>
<tr>
<td>7</td>
<td>37.8</td>
<td>125.8</td>
<td>3.33</td>
<td>0.47</td>
</tr>
<tr>
<td>9</td>
<td>39.6</td>
<td>188.4</td>
<td>4.76</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table 1: Sequential and parallel timings.

For a fixed number of iterations, there is a linear speed up for three to seven beams but a much greater speed up for nine beams. The main reason for this is not the efficiency of the parallel implementation, but rather the length of time taken by the sequential version for nine beams. Since a constant number of iterations was being run it is expected that the sequential times should remain more or less constant. For seven beams or less this was the case, but for nine beams there was a large increase in run time. This is thought to be due to insufficient memory and could be avoided if upgraded hardware was used.

It should also be noted that in the parallel implementation, the program runs much slower on 9 processors, due to the increase in communication overheads. This increase in run time for nine beams using the parallel version, is offset by the time that the sequential version takes for nine beams, so there is still a significant speed-up factor.

5 Discussion

5.1 Dose Calculation Model

The model used in the dose calculation was greatly simplified and would require some additions if it were to be used clinically. It would be possible to run the dose calculations for a particular patient independently from the rest of the program, and write the results to a file which could then be retrieved whenever a plan was required for the patient. This would be useful in a clinical situation if the plan was modified on a regular basis, or if the clinician was not satisfied with the original plan.

If scattering effects are included in the model, the calculation
of the dose distribution would become more time-consuming and it may be more efficient to parallelise the calculation at each point rather than parallelising the optimisation. Thus a grid decomposition method could be adopted or a combination of both techniques, task decomposition and grid decomposition, could be employed. One advantage of having the Transputers arranged in a ring is that computation and communication can take place simultaneously, thus making efficient use of the hardware.

5.2 Number of Beams

To produce treatment plans which satisfy clinical criteria, at least nine beams are required. However, increasing the number of beams beyond eleven did not produce a significantly improved plan, provided an odd number of beams was used so as to avoid parallel opposed pairs of beams. This is consistent with the results of other workers. Bortfeld et al. [7] found that seven or nine beams were necessary for a satisfactory result, whereas Webb [4] used a minimum of 32 beams. In the present work, convergence was always achieved by 30,000 iterations compared to the 5 million iterations required by Webb. The number of beams must be reduced further in order to reduce set-up and treatment time in a clinical situation. To achieve this, the dose restrictions on normal tissue could be relaxed by up to 50% or 60%.

5.3 Parallel Implementation

The algorithms for the parallel and sequential versions are different in a number of respects. In the sequential version, an element is randomly selected from all the possible beam elements, and the effect of a change of the weighting of this beam element is calculated. This is the true implementation of simulated annealing. In the parallel version, an element is selected from a set of the total number of elements i.e. the elements in one beam. Therefore a number of changes being made simultaneously which may cause problems in the algorithm. For example, there is a possibility that beam profiles may oscillate between over and under-dosed values. However, in the test runs the treatment plans produced by the parallel version of the code were almost identical to those produced by running the program sequentially. Hence the parallelising of the code did not significantly affect the quality of the result. This is probably due to the grain size being reduced as the temperature is reduced. If problems had occurred, then the number of iterations between updates would have to be reduced as the temperature was reduced.

6 Conclusions

A distributed simulated annealing approach for conformation RTP has been formulated and demonstrated successfully on a difficult clinical treatment scenario. It has been shown that increasing the number of beams to nine enables a high correlation to be achieved between prescribed and actual dose distributions. It has also been shown that the simulated annealing technique provides optimal beam profiles that produce a uniform dose to the tumour with minimum dosage to critical organs.

By implementing the code on a Transputer-based parallel computer, run times were reduced whilst retaining high quality results. Using 9 transputers, speed-up factors of four were obtained.

Acknowledgements

The authors would like to acknowledge useful discussions with Dr J. Mills (Walsgrave Hospital, Coventry, U.K.), Mr C. Dechan (Beatson Oncology Centre, Glasgow, U.K.) and Dr A.T. Redpath (Western General Hospital, Edinburgh, U.K.).

References

An Enhanced Adaptive Codebook for a CELP Coder

Yasheng QIAN\(^1\), Gebræl CHAHINE and Peter KABAL

Dept. of Electrical Eng., McGill University, Montreal, Quebec H3A 2A7
Tel/Fax: 514 398 7130/ 398 4470, E-Mail:kabal@tsp.ee.mcgill.ca
\(^1\)Dept. of Electronic Eng., Tsinghua University, Beijing, China, 100084
Tel/Fax: 861 259 4690/ 256 4176, E-Mail:qian@tsp.ee.mcgill.ca

Abstract. An adaptive codebook for a one-tap pitch filter has been used for determining the pitch filter using an analysis-by-synthesis procedure in CELP coders. In this paper, we first present the formulations for designing an enhanced adaptive codebook for a pseudo-three-tap pitch synthesis filter which gives a better performance than a conventional one-tap pitch filter. Then, we focus on the stability analysis of the pseudo-three-tap pitch filters. We propose a sufficient test condition with a relaxed stability, which gives a better performance than a strict stability check. We have employed the enhanced adaptive codebook based on a pseudo-three-tap pitch filter with fractional pitch lags for a 4.8 kb/s CELP speech coder. Both objective and subjective quality have been improved with the enhanced adaptive codebook.

1. Introduction

An adaptive codebook for representing a one-tap pitch filter has been successfully used for determining the pitch filter using an analysis-by-synthesis procedure in CELP coders [1], [2]. This analysis-by-synthesis approach provides a better reconstructed speech quality than if the pitch synthesis filter parameters are determined from the input speech. The pitch filter in a low-bit rate speech coder has a strong impact on the performance. Earlier we have reported that a pseudo-three-tap pitch prediction filter is an efficient way to characterize the periodicity in a speech signal [3]. It gives a higher prediction gain and a more appropriate frequency response than a conventional one-tap pitch filter. In contrast to this pitch prediction filter used for speech analysis, a pitch synthesis filter, which is the inverse filter of the pitch prediction filter, is used in speech coders. The adaptive codebook for a pseudo-three-tap pitch synthesis filter is referred to as an enhanced adaptive codebook.

Stability was studied as an important issue for pitch synthesis filters determined by analyzing the input speech [4]. An unstable pitch filter enhances the coding noise. For the analysis-by-synthesis configurations, it has been argued that stability is not important since the choice of filter parameters is based on the reconstructed speech which includes the effect of noise enhancement. Our experimental results, however, show that even in analysis-by-synthesis configurations, stability remains an issue that must be considered.

In our experimental work with unquantized pitch gains, we have seen the pitch coefficients rise to values as high as 800 in transition regions (unvoiced to voiced). In one utterance we saw the average SNR for a CELP coder using an adaptive codebook with unquantized pitch coefficients drop from 7.83 dB for a one-tap filter to 3.89 dB for a threetape filter. The resulting speech contained annoying pops, clicks and a more dominant background noise. Because we have imposed constraints on the prediction coefficients of the pseudo-three-tap pitch filter, the stability conditions and stabilization procedure can be simplified.

We first describe the enhanced adaptive codebook for a pseudo-three-tap pitch synthesis filter. Then, we focus on the stability analysis for the pitch filter. We present a stabilization procedure with a relaxed stability check, which is better than a strict stability check. Finally, performances of the enhanced adaptive codebook in a 4.8 kb/s CELP coder are given.

2. An enhanced adaptive codebook

We employ an enhanced adaptive codebook to determine the pitch lag and prediction coefficients of the pseudo-three-tap pitch synthesis filter in a closed-loop search procedure, as shown in Figure 1.

A pseudo-three-tap pitch synthesis filter is a three-tap pitch filter, which has certain constraints on the pitch coefficients, as shown in Figure 2. Let the three non-zero coefficients of the pitch filter be \( \beta_1, \beta_2 \) and \( \beta_3 \). We can restrict this filter with a symmetrical set of coefficients, by assigning
\[
\beta_1 = \beta_2 = \alpha \beta, \quad \beta_2 = \beta.
\]

Both \( \beta \) and \( \alpha \) are optimized for best performance. This filter has two degrees of freedom. We can further restrict the pseudo-three-tap filter to one degree of freedom by fixing the value of \( \alpha \).

The notation for pseudo-three-tap pitch filters \( nTmDF \), means \( n \)-taps, \( m \) degrees of freedom. Thus, a pseudo-three-tap pitch filter with one degree of freedom is denoted as 3T1DF. Conventional one-tap and three-tap pitch filters are denoted as 1T1DF and 3T3DF, respectively.

The enhanced adaptive codebook corresponds to the set of the pseudo-three-tap pitch filter outputs. Normally, the

---

This research was supported by a grant from the Canadian Institute for Telecommunications Research under the NCE program of the Government of Canada.
The i-th codeword vector of N elements, $D_i$, is equal to the $d(n)$ corresponding the given $M$, $\beta$ and $\alpha$. If the subframe length of $N$ is shorter than the delay of $M$, the output of the pitch filter, $d(n)$, is obtained by

$$d(n) = \beta \alpha d(n-M-1) + \beta d(n-M) + \beta \alpha d(n-M+1);$$

$$1 \leq n \leq N.$$

The optimal codeword of the enhanced adaptive codebook is determined by minimizing the perceptual weighted mean square errors (MSE) during the closed-loop search. The error between the input speech and reconstructed speech is

$$e(n) = x(n) - \sum_{k=0}^{n} d(k) h(n-k)$$

where $h(n)$ is the impulse response of the FMA filter; $d(n)$ corresponds to the codeword, $D_i$. The perceptual weighted error $e_p(n)$ is the convolution of the error $e(n)$ and the impulse response of the perceptual weighted filter $h_p(n)$. There are 128 codewords with integer pitch lags and 128 codewords with non-integer lags, as defined in the Federal Standard 1016 [2].

3. Stability
We first explore the stability of the pitch synthesis filter, as determined by an analysis-by-synthesis search procedure. The output of the pitch filter depends on the output of the previous subframe. We can decompose the output into two components: one excited by an ideal prediction residual (at analysis stage), and a quantization noise output.

For the prediction residual, stability is not a problem because of pole/zero cancellation in the synthesis phase. However, the quantization noise passes through only the unstable synthesis filter. We model the quantization noise to be an additive noise (possibly correlated with the prediction residual). An unstable filter can result in a large boost in the output noise energy. Therefore, the augmented noise can result in pitch filter parameter errors during searching of the adaptive codebook. Furthermore, the noise can be amplified for the consequent subframes, because the adaptive codebook is updated with the accumulated noise of an unstable pitch filter.

The average SNR of the conventional (unquantized pitch coefficient) 3T3DF pitch filter (one testing sentence), using the closed-loop search, drops down to 3.89 dB, comparing to 9.0 dB for a 1T1DF, and produces annoying pops, clicks and dominant background noise. The waveform of the reconstructed speech with an adaptive codebook for a 3T3DF is shown in Figure 3. Comparing to the original speech waveform in Figure 4, we find that the unstable 3T3DF filter severely impacts the reconstructed speech. In order to alleviate the unstable problem, two stability sufficient test formulas and stabilization techniques have been proposed to efficiently reduce the effect of an unstable pitch filter in [4]. The simple sufficient stability conditions are:

$$|\beta| < 1, \quad 1T1DF$$

$$|\beta_1| + |\beta_2| + |\beta_3| < 1. \quad 3T3DF$$

Let $a = \beta_1 + \beta_3$ and $b = \beta_1 - \beta_3$. The tight sufficient stability conditions for a 3T3DF pitch filter are [4]:

1. if $|a| \geq |b|$, $|\beta_1| + |\beta_2| + |\beta_3| < 1.$
2. if $|a| < |b|$, and $|\beta_2| + |a| < 1$;

$$b^2 \leq a, \quad b^2 - (1-b^2)(b^2-a^2) < 0.$$
\( b = 0 \) and \(|a| > 0 \). The 3T1DF pitch filter gives a better stability performance than a conventional 3T2DF filter, since we constrain the side prediction coefficients \( \beta_1 = \beta_2 \) to be a small proportion of the center coefficient \( \beta_3 \). Let \( \alpha = \beta_1 / \beta_3 \). Therefore, the 3T1DF filter meets the sufficient condition for the simplest stability test in (5)

\[
|\beta_3| < \frac{1}{1+2|\alpha|}
\]

For a 3T2DF pitch filter with \( \beta_1 = \beta_2 = \gamma \), the simplest sufficient condition is

\[
2|\gamma| + |\beta_3| < 1
\]

A simple stabilization method of the scaled-down coefficients is utilized to stabilize the pitch synthesis filter, if unstable. We scale down the pitch coefficients by multiplying a factor \( c \),

\[
c = \frac{1}{(1+|\gamma|+|\beta_3|)}, \quad \text{if} \quad (|\beta_1| + |\beta_2| + |\beta_3|) > T_h.
\]

The threshold \( T_h \) is an experimentally determined threshold. With \( T_h = \infty \) no stabilization method is used. With \( T_h = 1 \), a strict stability condition is imposed. Figure 5 is the waveform for a simple stabilized 3T3DF filter. Comparing to the impaired waveform of the unstable 3T3DF filter (Figure 3), we find that the improvement with the stabilization method is very good.

4. Performance of the enhanced adaptive codebook

The enhanced adaptive codebook for pseudo-three-tap pitch filters, 3T1DF and 3T2DF pitch filters with unquantized coefficients were incorporated into a FS1016 CELP coder. The block diagram of the improved CELP speech coder is depicted in Figure 6. The conventional adaptive codebook is replaced by the enhanced adaptive codebook. Other blocks in the speech coder are the same as the FS1016 standard. We employ three performance measures: the average SNR–signal-to-noise ratio, the SEGSNR–segmental signal-to-noise ratio (average of log SNR’s evaluated for 16 ms segments and the SFG–synthesis-filter-gain. We define the SFG as the ratio of the energy of the original speech signal and the energy of error between the original speech and the reconstructed speech signal using only the adaptive codebook excitation for the formant synthesis filter. A high value of the SFG indicates that the pitch filter is contributing a large part of the reconstructed signal, while the stochastic codebook is contributing a relatively small part.

Table 1 shows these performance measures for two male and two female test sentences. For comparison, a conventional one-tap filter (1T1DF) and a three-tap filter (3T3DF) are also included. The coefficients are unquantized and the pitch lags are integers, but stabilization as described above is applied. The adaptive codebook for the 3T3DF filter obtains a significant increment of SNR gain of 1.16 dB over 1T1DF filter. The stability threshold \( T_h \) is set to be 1.0, 1.10, 1.15 and 1.20 for comparisons. The threshold \( T_h \) is denoted in the subscript of the type of the pitch filter. For example, The 3T1DF filter uses the \( T_h = \infty \). It means that the pitch filter is not stabilized.

The results show that the stabilization actually improves the performance. Moreover, a relaxed stability constraint is better than a strict stability constraint. The reason is that the increasing pitch pulse amplitudes are better to model the fast growing voicing onset. The SNR for 3T1DF increases by 1.15 dB. The SNR difference between the 3T1DF and the 3T3DF filter is small (0.32 dB).

We have also applied quantization to the 3T1DF pitch filter coefficients. The quantization table is defined in the FS1016 CELP coder specification. Notice that the quantization is in effect present, since the largest quantized value for \( |\beta_3| \) is 1.961. Therefore, the maximum sum of \( |\beta_3| (1+2|\alpha|) \) = 2.53, because we select \( \alpha = 0.135 \). With quantization, the SNR for the 3T1DF configuration drops by only 0.13 dB.

Finally, we have evaluated the SNR and SEGSNR for the 3T1DF pitch filter with fractional pitch lags \( [k, \delta] \) and pitch quantizer (FS1016 CELP coder). The results show that the SNR and SEGSNR increase by 0.44 dB and 0.05 dB, respectively, over those of the integer pitch filter. The SNR and SEGSNR are higher than standard FS1016 coder by 0.46 dB and 0.1 dB. An informal listening test show that the improved CELP coder with 3T1DF pitch filter is better than the original FS1016 CELP coder.

5. Conclusions

The enhanced adaptive codebook for pseudo-three-tap pitch synthesis filters can be incorporated in a CELP coder to improve the speech quality. A scaled-down pitch coefficients technique with a relaxed sufficient constraints to obtain a stable pitch synthesis filter can track fast changing segments during a unvoicing to voicing onset. The performance of the improved 4.8 kb/s CELP coder with the pseudo-three-tap pitch filter is better than the FS1016 coder with a one-tap pitch filter.

References


Figure 3  Reconstructed waveforms with an unstable 3T3DF pitch synthesis filter

Figure 4  Original speech waveforms

Figure 5  Reconstructed waveforms with a stabilized 3T3DF$_{61.0}$ pitch filter

<table>
<thead>
<tr>
<th>Type</th>
<th>SNR (dB)</th>
<th>SEGSNR (dB)</th>
<th>SFQ (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3T1DF$_{60}$</td>
<td>7.80</td>
<td>7.77</td>
<td>5.52</td>
</tr>
<tr>
<td>1T1DF$_{81.0}$</td>
<td>7.13</td>
<td>7.74</td>
<td>5.33</td>
</tr>
<tr>
<td>1T1DF$_{81.1}$</td>
<td>7.85</td>
<td>7.80</td>
<td>5.29</td>
</tr>
<tr>
<td>1T1DF$_{81.15}$</td>
<td>7.81</td>
<td>7.73</td>
<td>5.27</td>
</tr>
<tr>
<td>1T1DF$_{82.0}$</td>
<td>7.89</td>
<td>7.88</td>
<td>5.27</td>
</tr>
<tr>
<td>3T1DF$_{60}$</td>
<td>6.77</td>
<td>7.89</td>
<td>5.66</td>
</tr>
<tr>
<td>3T1DF$_{81.0}$</td>
<td>7.72</td>
<td>7.88</td>
<td>5.17</td>
</tr>
<tr>
<td>3T1DF$_{81.1}$</td>
<td>8.11</td>
<td>7.97</td>
<td>5.40</td>
</tr>
<tr>
<td>3T1DF$_{81.15}$</td>
<td>8.26</td>
<td>8.02</td>
<td>5.42</td>
</tr>
<tr>
<td>3T1DF$_{82.0}$</td>
<td>8.29</td>
<td>8.00</td>
<td>5.59</td>
</tr>
<tr>
<td>3T2DF$_{80}$</td>
<td>4.60</td>
<td>8.03</td>
<td>5.78</td>
</tr>
<tr>
<td>3T2DF$_{81.0}$</td>
<td>6.89</td>
<td>7.19</td>
<td>4.85</td>
</tr>
<tr>
<td>3T2DF$_{81.1}$</td>
<td>7.28</td>
<td>7.32</td>
<td>5.09</td>
</tr>
<tr>
<td>3T2DF$_{81.15}$</td>
<td>7.43</td>
<td>7.64</td>
<td>5.36</td>
</tr>
<tr>
<td>3T2DF$_{82.0}$</td>
<td>8.30</td>
<td>8.18</td>
<td>5.68</td>
</tr>
<tr>
<td>3T3DF$_{80}$</td>
<td>3.89</td>
<td>8.27</td>
<td>3.98</td>
</tr>
<tr>
<td>3T3DF$_{81.0}$</td>
<td>7.37</td>
<td>7.58</td>
<td>4.75</td>
</tr>
<tr>
<td>3T3DF$_{81.1}$</td>
<td>7.78</td>
<td>7.94</td>
<td>5.65</td>
</tr>
<tr>
<td>3T3DF$_{82.0}$</td>
<td>8.01</td>
<td>8.32</td>
<td>5.91</td>
</tr>
</tbody>
</table>

Table 1  SNR (dB) comparisons for different pitch synthesis filters in a CELP speech coder

Figure 6  Block diagram of the improved CELP speech coder
On Improving Wideband CELP Speech Coders

Ciarán McELROY, Brian P. MURRAY, Anthony D. FAGAN

DSP Research Group,
Dept. of Electronic and Electrical Engineering,
University College, Dublin,
Ireland.
Tel. +353-1-7061964, Fax. +353-1-2830921
E-Mail : ctaranmc@maxwell.wcd.ie

Abstract. This paper examines the problem of achieving near transparent speech coding at medium bit rates for wideband speech (0-7 kHz). Two methods of improving the quality of CELP coders are considered. One involves the parameterisation of the prediction residual through the use of multiple codebooks and pulses coupled with adaptive bit allocation. The other uses double-sided linear prediction to improve the vocal tract modelling. The multiple codebook approach produces coders operating at 16 kb/s and 24 kb/s that provide comparable speech quality to that produced by the CCIT G.722 coder operating at 64 kb/s. However the double-sided linear prediction fails to produce the improvements in speech quality that would be expected from its improved prediction gain.

1. Introduction

The speech production models used in conventional CELP coders [1] usually consist of two main sections, namely the vocal tract model and its excitation. The excitation is a combination of a weighted noise sequence and previous excitation. The vocal tract model is a one-sided linear predictor (i.e. it assumes that the present speech sample is a linear combination of past samples). Using the above models it is possible to obtain reasonably high quality speech coding at moderate bit rates. However, an audible and, sometimes annoying, distortion is present. This distortion remains even if the coder operates at high bit rates, indicating that it is due to limitations in the model of the speech production process used in the coder.

Often this distortion is acceptable in narrowband coders (0-4 kHz), where the emphasis is on achieving the lowest possible bit rate. In wideband coders the emphasis is usually on achieving the highest possible speech quality at a moderate bit rate, thus the distortion is unacceptable in wideband coders.

2. Improving the Excitation Model

Conventional CELP coders use a combination of a weighted white noise sequence and previous excitations as the excitation for the synthesis filter. This model is not suited to accurately modelling unvoiced to voiced transitions [2] and so a distinctive warble is introduced into the speech. Also the white nature of the excitation fails to take account of the relative perceptual importance of certain frequency ranges. Only 10-20 % of the information contained in a speech signal is contained in the frequencies above 4 kHz.

A considerable amount of work has been carried out in low bit rate narrowband speech coding towards solving the onset problem [2]. This work indicates that the problem is due to the adaptive codebook of past excitations being incapable of introducing the required pitch pulses at the onset. The adaptive codebook works well in steady state voiced speech however it fails during the transitions. If the adaptive codebook misses the pitch pulse then it must be modelled by the fixed codebook. Since this codebook is stochastic it cannot do this without distorting the excitation around the pulse. If an impulse was used instead of the stochastic codebook then this problem would be minimised. However the stochastic codebook would have to be retained to code the rest of the speech. The improvement due to the inclusion of the impulse is clearly seen in Figure 1. The accuracy of the excitation coding around the pitch pulse is greatly improved.

![Figure 1: Excitation signals, (a) the original prediction residual, (b) the excitation using a white codebook, (c) the excitation when using a white codebook and pulses.](image-url)
As stated earlier only 10–20% of the information in a speech signal is contained in the frequencies above 4 kHz. When a white codebook is used all frequencies are given equal importance. This is inefficient. The situation can be improved by using bandlimited codebooks to provide greater emphasis on the perceptually important frequencies.

As indicated above the excitation signal is not accurately represented by just one type of signal. Therefore we need a number of different excitation to be available in each frame of speech. This can be achieved by searching the codebooks sequentially and adding their contributions in the same way as the adaptive codebook's effect is added. The situation is complicated by the fact that the order in which the codebooks are searched affects the resulting speech quality. For example, when an onset is present the impulse is important and the white codebook is less so but during a fricative the white codebook is more important than the impulse. This means that the order in which the codebooks are searched must change with the characteristics of the speech. In the coder suggested here all the codebooks are searched to find the codeword which will produce the highest SNR. This codebook is then excluded and the rest of the codebooks are searched, allowing for the effect of the previously selected codeword.

The pitch codebook is not included in the adaptive ordering as this was found to reduce speech quality, instead the pitch search is carried out first. This is most likely due to the pitch codebook improving the periodicity of the speech even if it does not produce the highest SNR. Periodicity in voiced speech is critical to the speech quality [3].

The adaptive ordering of the codebook searches produces an audible improvement in speech quality and also facilitates the introduction of adaptive bit allocation. The excitation is the sum of all the selected codewords so the order in which they were searched is irrelevant to the synthesizer. Likewise the presence or absence of certain codewords is unimportant. Adaptive bit allocation is achieved by only using the first few codebooks that are selected by the adaptive ordering algorithm. An index indicating which codebooks have been used is then transmitted as side information. If bandlimited codebooks are used some frequency bands may be temporarily ignored by the fixed codebook searches. These gaps can be adequately filled by the pitch codebook.

The coders tested in this work used six codebooks, an impulse codebook, a stochastic codebook and four bandlimited codebooks with the following bandwidths 0–1 kHz, 1–2.5 kHz, 2.5–5 kHz and 5–8 kHz. The bandwidths were chosen to increase roughly logarithmically with frequency. Of these six codebooks only three are used at one time, requiring 5 bits to describe the allocation. The codebooks contain 2048 codewords and the gains are quantised to 5 bits, implying 16 bits per codeword. The impulse codebook contains the same number of codewords as its vector dimension but the adaptive bit allocation scheme insists that 16 bits be set aside for it. The impulse codebook is used in 10–20% of subframes, so this wastage of bits is insignificant. The pitch codebook contains 256 codewords and so uses a total of 13 bits. Therefore each subframe requires 66 bits. A twentieth order single-sided predictor is used and its coefficients are updated once per frame, using 70 bits. A frame length of 25 ms was used and the number of subframes per frame dictated the final bit rate. The 24 kb/s coder used 8 subframes per frame and the 16 kb/s coder used 5.

3. Improving the Vocal Tract Model

Conventional CELP coders perform significantly better for male speech than for female speech. This difference is partly due to the model for the vocal tract used. The prediction gain achieved by the traditional single-sided vocal tract model is several dB lower for high pitch voices than it is for low pitch voices, see Figure 2. This means that the starting point for the excitation coder is significantly worse for high pitch voices. The implication of this is that if the vocal tract model can somehow be improved then it may be possible to achieve transparent speech coding for all speech types.

![Figure 2: A typical plot of prediction gain against predictor order for male and female speech.](image)

The conventional model assumes that the present speech sample is a linear combination of past samples (single-sided prediction). This model can be improved by assuming that the present sample is a linear combination of past and future samples (double-sided prediction) [4]. The prediction gain achieved by this model is significantly higher than that achieved by the single sided predictor, see Figure 3.

Determining the optimum coefficients for a double-sided predictor involves only a slight extension to the method for designing a single-sided predictor and is summarised below.
Assume

$$\hat{x}_n = -\sum_{m=0}^{\infty} a_m x_{n-m}$$  \hspace{1cm} (1)

where $x_n$ is a speech sample and $\hat{x}_n$ is its predicted value. We want to minimise the error:

$$E = \sum_n [x_n - \hat{x}_n]^2 = \sum_n \left[ x_n + \sum_{m=p}^{\infty} a_m x_{n-m} \right]^2$$  \hspace{1cm} (2)

The solution of which yields the following equations:

$$\sum_{m=p}^{\infty} a_m x_{n-m} = -\sum_n x_n x_{n-k} - p \leq k \leq p, k \neq 0$$  \hspace{1cm} (3)

This system of equations can be solved using Cholesky decomposition or other standard methods.

However, there are several major problems associated with double-sided prediction. The first is that the predictor is almost always a non-minimum phase system, yielding an unstable synthesis filter. This problem can be solved by reflecting any zeros of the predictor that are outside the unit circle in the z-plane to their conjugate reciprocal locations. The resulting system is minimum phase but a number of zeros will be very close to the unit circle, producing a marginally unstable synthesis filter. A bandwidth expansion factor can be used to radially translate all the zeros towards the origin.

The next problem is that the double-sided predictor is not a noise whitening filter [5]. Many false peaks appear in its frequency response. Since the excitation model assumes that the residual signal has a roughly flat spectrum we get tonal artefacts introduced into the synthesised speech. One solution to this problem is to use a cascade of a low order double-sided predictor and a large order single-sided predictor. The double-sided predictor provides a large prediction gain. While the single-sided predictor whitens the residual as well as providing some additional prediction gain. Experimental results suggest that this cascade is capable of producing a prediction gain similar to that achieved by a large order double-sided predictor. Once the cascade is designed a delay is introduced to make the system causal and its impulse response is determined. As far as the excitation coder is concerned this looks exactly like the one-sided predictor, just with a different set of coefficients.

When the single-sided vocal tract model is replaced by the cascade system only a slight increase in SNR is observed. Furthermore the subjective quality of the synthesised speech is significantly reduced. The reason for this is not fully understood. If the SNR is looked at as a function of frequency then it is clear that the cascade performs better than the single-sided model only at low frequencies. The increased distortion at high frequencies may be what the ear is objecting to.

4. Results

The quality of the speech produced by the coders using double-sided prediction was obviously inferior to that produced by those using single-sided prediction. For this reason they were not included in the subjective testing. Informal subjective tests were carried out on the multi-codebook coder (MC-CELP) operating at 24 kbps and 16 kbps. These coders were compared with speech produced by the G.722 coder operating at 64, 56 and 48 kbps. Four speakers were used, two male and two female, each speaking different sentences for ten seconds. The tests were carried out by ten untrained listeners. Tables 1 and 2 show the preferences expressed in these tests.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>G.722</th>
<th>MC-CELP</th>
<th>Neither</th>
</tr>
</thead>
<tbody>
<tr>
<td>64 vs. 24</td>
<td>80 %</td>
<td>5 %</td>
<td>15 %</td>
</tr>
<tr>
<td>56 vs. 24</td>
<td>70 %</td>
<td>20 %</td>
<td>10 %</td>
</tr>
<tr>
<td>48 vs. 24</td>
<td>10 %</td>
<td>90 %</td>
<td>0 %</td>
</tr>
<tr>
<td>64 vs. 16</td>
<td>90 %</td>
<td>5 %</td>
<td>5 %</td>
</tr>
<tr>
<td>56 vs. 16</td>
<td>85 %</td>
<td>10 %</td>
<td>5 %</td>
</tr>
<tr>
<td>48 vs. 16</td>
<td>50 %</td>
<td>45 %</td>
<td>5 %</td>
</tr>
</tbody>
</table>

Table 1: Subjective test results for female speech.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>G.722</th>
<th>MC-CELP</th>
<th>Neither</th>
</tr>
</thead>
<tbody>
<tr>
<td>64 vs. 24</td>
<td>20 %</td>
<td>30 %</td>
<td>50 %</td>
</tr>
<tr>
<td>56 vs. 24</td>
<td>0 %</td>
<td>55 %</td>
<td>45 %</td>
</tr>
<tr>
<td>48 vs. 24</td>
<td>5 %</td>
<td>80 %</td>
<td>15 %</td>
</tr>
<tr>
<td>64 vs. 16</td>
<td>65 %</td>
<td>20 %</td>
<td>15 %</td>
</tr>
<tr>
<td>56 vs. 16</td>
<td>40 %</td>
<td>35 %</td>
<td>25 %</td>
</tr>
<tr>
<td>48 vs. 16</td>
<td>30 %</td>
<td>60 %</td>
<td>10 %</td>
</tr>
</tbody>
</table>

Table 2: Subjective test results for male speech.

The results indicate that the multi-codebook coder performs significantly better for low pitch voices. The 24 kbps coder is as good as G.722 at 64 kbps for male speech while the quality for female speech is somewhere
in between that for G.722 at 56 and 48 kb/s. At 16 kb/s we get a coder comparable to G.722 at 56 kb/s for male speakers but comparable to G.722 at 48 kb/s for female speakers.

The coders were also compared with another multi-codebook coder with three stochastic codebooks (i.e. no dynamic ordering and no adaptive bit allocation) and a coder using just one stochastic codebook. As expected the MC-CELP coder is significantly better than the coder using just one codebook. The performance of the coder using three codebooks was speaker specific. For some speakers it was virtually indistinguishable from the MC-CELP while for others there was a significant noise component present.

5. Conclusions

The subjective tests clearly show that the use of multiple codebooks and pulses coupled with adaptive bit allocation in modelling the excitation can produce very high quality speech at bit rates as low as 24 kb/s and 16 kb/s. This gain in quality is achieved at the expense of a significant increase in the computational complexity of the coder over conventional coders. However, there is still an audible distortion present, especially with high pitch voices. This distortion is not completely removed by the inclusion of the impulse codebook. The situation may be improved by using glottal impulses instead [6].

Another method of reducing this distortion would be to improve the vocal tract model. It has been noted the prediction gain achieved when coding high pitch voices is significantly lower than that for low pitch voices. A double-sided linear predictor was suggested to rectify this problem. The double-sided predictor increases prediction gain by approximately 50%. However the SNR of the resulting speech increases only marginally and the subjective quality is significantly worse than that from the single-sided predictor. This problem has not been satisfactorily explained and further work is needed.

Acknowledgements

The authors would like to acknowledge the assistance provided by Teltec Ireland who sponsored the work.

References

CELP Coding with Data Rates Below 4 kbit/s Using a Phonetically Oriented Excitation

Ulrich BALSS, Ulrich KIPPER, Herbert REININGER and Dietrich WOLF

Institut für Angewandte Physik der J.W. Goethe-Universität Frankfurt
Robert-Mayer-Straße 2-4, D-60054 Frankfurt a.M., FRG
Tel.: +49 69 798 3490, Fax: +49 69 798 8510, E-Mail: herbrein@apx00.physik.uni-frankfurt.de

Abstract. Recently proposed CELP schemes incorporate phonetic knowledge in order to lower the data rate [1, 2]. Speech segments belonging to different phonetic categories are encoded at different rates which results in a variable data rate. Here we describe a CELP scheme with phonetically oriented excitation based on pulse-grids. With a different bit allocation between synthesis filter and excitation the different characteristics of voiced and unvoiced speech segments are taken into account during the optimization of the excitation vectors. An explicit phonetic classification of speech segments is avoided by selecting an optimum excitation vector in accordance with the minimum coding error. The proposed coding scheme operates at a constant data rate and achieves the speech quality of a 4.8 kbit/s CELP scheme at a data rate of 3.8 kbit/s.

1. Introduction
In order to achieve a good speech quality with CELP schemes at data rates below 4.8 kbit/s the different signal characteristics of speech segments belonging to different phonetic categories can be taken into account. Recently proposed CELP schemes [1, 2] incorporate phonetic knowledge by encoding speech segments from different phonetic categories at different data rates. The available bit rate is distributed between the synthesis filter and the excitation according to the perceptually significant speech segment. Furthermore, different kinds of excitation are used for generating speech segments belonging to different categories. In these approaches the phonetic classification of a speech segment is based on criteria different from the coding error which is used for choosing the optimum excitation. Here we describe a CELP scheme, called ACELP, with a phonetically oriented excitation consisting of excitation vectors especially designed for generating voiced or unvoiced speech segments. The excitation consists of pulse-grids with different number of pulses and pulse amplitudes which are adapted to the actual speech signal. By increasing the quantization accuracy of the pulse amplitudes the synthesized speech signal is getting closer to the original speech signal. Therefore, with a variable bit allocation between the synthesis filter and the amplitude codebooks excitation vectors can be designed which are appropriate for generating speech segments with different phonetic characteristics.

2. CELP with Adaptive Excitation
In ACELP [3] the fixed stochastic codebook is replaced by a codebook containing excitation vectors with parameters adapted to an actual speech frame. The excitation vectors of the adaptive codebook (AC) are generated as shown in Figure 1 - from regularly spaced pulses, denoted as pulse-grids. The pulse-grids are grouped into G different grid-classes each comprising all grids with equal number $N_F$ of pulses and equal distance $\Delta$ of the pulses but different positions of the first pulses. Assuming that every position of an excitation vector with $N_S$ elements occurs as a pulse position in a grid-class, the number $N_G$ of pulse-grids within a class is defined by $N_S$ and $N_F$. Then a pulse-grid of a grid-class is uniquely defined by the offset $\phi \in [0, N_F - 1]$ of the first pulse in a grid.

A pulse-grid is adapted to an actual speech frame by calculating (A) pulse amplitudes such that the squared error between the speech signal $s(n)$ and the speech signal synthesized with this pulse-grid is minimum. The vector of optimum pulse amplitudes is quantized using a gain-shape vector quantizer (VQG) [4], i.e. an amplitude vector is first normalized from which a gain factor $g$ and a so-called shape vector results. The gain factor is quantized with a scalar quantizer and the shape vector with a conventional vector quantizer. The vector dimension and the number of codebook vectors depend on the grid-class. The number of codebook vectors is chosen such that a constant data rate results for
each grid-class. Therefore, the quantization accuracy of a pulse amplitude decreases with increasing $N_p$.

The speech adapted pulse-grids with quantized amplitudes $c_i(n)$ constitute the adaptive excitation codebook. From this codebook the optimum excitation vector is determined using the conventional analysis-by-synthesis method as in a CELP scheme with stochastic excitation. An optimum excitation vector is uniquely defined by its grid-class, offset, gain factor, and shape vector. These parameters can efficiently be encoded at very low bit rate.

### 3. Performance Analysis of the Synthesis Filter and the Excitation

Using several pulse-grids for encoding consecutive speech frames, excitation sequences with variable pulse rates are realized. Comparative listening tests showed that ACELP achieves a substantially higher speech quality than CELP at a data rate of 4.8 kbit/s. At data rates lower than 4.8 kbit/s quantization noise, mainly due to the small amount of bit rate available for quantizing the pulse amplitudes, is audible in the coded speech. One way to provide more bits for the quantization of the pulse amplitudes, without increasing the overall bit rate, is to drop the fixed bit allocation between the synthesis filter and the excitation to the generation of speech segments belonging to different phonetic categories.

In order to optimize the bit allocation with respect to the phonetic characteristics of the speech segments, the performance of the synthesis filter was analyzed by measuring the prediction gains of the short-term predictor (STP) and the long-term predictor (LTP) during the encoding several phonemes. As can be seen from Table 1, the prediction gains of STP and LTP are extremely high for vowels. Also in other voiced speech segments the predictors operate quite well. For the unvoiced fricatives the prediction gain of the STP as well as of the LTP is relatively low. At the beginning of voiced speech segments following unvoiced or pause segments the prediction gains are also low because the filter memories contain only inadequate samples. In particular the LTP is very inefficient due to its long memory. The synthesized speech is in this cases almost identical with the excitation signal. Therefore, the more accurate the amplitudes of the pulse vectors of ACELP are quantized the better the synthesized speech signal resembles the original speech signal.

### 4. ACELP with Phonetically Oriented Excitation Vectors

In order to provide grid-classes suited for generating phonetically different speech segments, all possible grid-classes were considered with and without LTP. The bit rate available for the encoding of the excitation is in one case distributed between the parameters of the LTP and the excitation parameters. In the other case, the bit rate is exclusively used for the excitation parameters. In the process of selecting an optimum excitation vector both excitation types are considered. Therefore, an optimum excitation vector is chosen consistently with the objective of the coding procedure which is the minimization of the coding error.

The performance of this approach was investigated using an ACELP codec designed for a data rate of about 4.0 kbit/s. The frame length for adaptation of the STP was set to 256 samples. The parameters of

### Table 1: SNR and prediction gain (PG) of short-term predictor (STP) and of long-term predictor (LTP) for encoding speech signals belonging to isolated phonemes

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>$SNR$ (dB)</th>
<th>$PG_{STP}$ (dB)</th>
<th>$PG_{LTP}$ (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>11.266</td>
<td>16.668</td>
<td>7.708</td>
</tr>
<tr>
<td>/e/</td>
<td>12.080</td>
<td>16.475</td>
<td>4.730</td>
</tr>
<tr>
<td>/i/</td>
<td>15.478</td>
<td>18.268</td>
<td>5.030</td>
</tr>
<tr>
<td>/o/</td>
<td>12.582</td>
<td>24.551</td>
<td>7.857</td>
</tr>
<tr>
<td>/u/</td>
<td>8.847</td>
<td>33.981</td>
<td>4.748</td>
</tr>
<tr>
<td>/m/</td>
<td>8.453</td>
<td>9.413</td>
<td>5.802</td>
</tr>
<tr>
<td>/n/</td>
<td>8.095</td>
<td>9.593</td>
<td>5.996</td>
</tr>
<tr>
<td>/z/</td>
<td>12.092</td>
<td>16.554</td>
<td>4.588</td>
</tr>
<tr>
<td>/æ/</td>
<td>10.014</td>
<td>9.298</td>
<td>5.662</td>
</tr>
<tr>
<td>/s/</td>
<td>1.643</td>
<td>2.197</td>
<td>0.582</td>
</tr>
<tr>
<td>/f/</td>
<td>1.463</td>
<td>1.513</td>
<td>0.516</td>
</tr>
</tbody>
</table>

917
Table 2: Configuration of grid-classes and bit rates

<table>
<thead>
<tr>
<th>G</th>
<th>Δ</th>
<th>N_P</th>
<th>φ</th>
<th>g</th>
<th>S</th>
<th>LTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>16</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>32</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>64</td>
<td></td>
<td>3</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>32</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>14</td>
<td>—</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>14</td>
<td>—</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>16</td>
<td>—</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>16</td>
<td>2</td>
<td>4</td>
<td>16</td>
<td>—</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>32</td>
<td>1</td>
<td>3</td>
<td>18</td>
<td>—</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>64</td>
<td></td>
<td>4</td>
<td>18</td>
<td>—</td>
</tr>
</tbody>
</table>

a 10th order predictor, represented by Line Spectrum Frequencies, were quantized using a 24 bit vector quantizing scheme with split codebooks [6]. This results in a data rate of 750 bit/s for transmitting the parameters of STP. The LTP and the excitation were adapted every \( N_S = 64 \) samples, so that 26 bit per subframe could be used for encoding an excitation vector and the parameters of LTP. With \( N_S = 64 \) samples, 6 grid-classes with the numbers \( N_P \) of pulses shown in Table 2 could be realized. These grid-classes occur in combination with and without LTP. Therefore, 12 grid-classes were effectively available and thus, 4 bit were required for uniquely defining a grid-class. The remaining 22 bit could be used for encoding an excitation vector if no LTP was used. In the case with LTP, 9 bit were needed for encoding the predictor coefficient and the delay coefficient and thus only 13 bit were available for an excitation vector. How the bit rate of an excitation vector is assigned to the parameters offset \( \phi \), gain \( g \), and shape vector \( S \), is also shown in Table 2.

In the mode of the coder with LTP a shape vector was quantized using tree-search vector quantizers of dimension \( N_P \). In the coder without LTP a shape vector was splitted into two vectors each of dimension \( N_P/2 \). The quantization of the vector component was performed using separate tree-search vector quantizers each with a code rate which is half of the total bit rate for the shape vector.

Figure 2 shows the histograms of the selected grid-classes occurring at the coding of isolated phonemes. As can be seen from the histograms of the grid-classes used for generating the vowels /a/ and /e/, with an
LTP quasi-periodic signals can be synthesized very efficiently. Even in the case of the nasal /m/, where the performance of STP is relatively low, the speech synthesized with an LTP is closer to the original than that originating from a more precise quantized excitation. The same is true for the voiced fricative /z/, although the histogram of the grid-classes looks quite different. In the case of unvoiced fricatives /s/ and /f/, speech synthesized without LTP often yields a smaller coding error than with LTP.

The fact that the automatically selected grid type corresponds well with the phonetic type of the speech segments confirms the consistency of the design method with the closed-loop optimization procedure.

In subjective listening tests the speech quality of ACELP with phonetically oriented excitation was compared to that of standard ACELP and CELP schemes with stochastic excitation vectors at different bit rates. At a data rate of 4.8 kbit/s the quality of the two ACELP schemes was rated higher than that of the CELP scheme but no difference of the speech processed by the ACELP schemes was noticed. For a data rate of 4.0 kbit/s ACELP with phonetically oriented excitation achieves a significantly better speech quality as the other schemes.

Figure 3 shows the histogram of selected grid-classes resulting from the encoding of a speech signal containing several sentences spoken from different speakers. As can be seen, the excitation vectors from the voiced grid-class 8 are not used very often. This is due to the relatively large number of pulses for which only a crude quantization of the amplitudes can be provided. Furthermore, can be seen that excitation vectors from the unvoiced grid classes 7 to 9 are also used very rarely. These grid-classes, containing only a small number of pulses, are not adequate for generating unvoiced speech. By discarding the grid-classes 6 to 9 the number of grid-classes can be reduced from 12 to 8. This lowers the data rate from 4.8 kbit/s to 3.8 kbit/s. The subjective speech quality was not affected by this restriction of the number of possible grid-classes.

5. Conclusions

With ACELP using an excitation consisting of regularly spaced pulses organized in grid-classes a phonetically oriented excitation can easily be realized. By different bit allocation between the parameters of the LTP and the quantizers of the pulse amplitudes, excitation vectors are designed which are especially suited for generating voiced and unvoiced speech. Using this kind of excitation the data rate of an ACELP codec can be reduced from 4.8 kbit/s to 3.8 kbit/s without a loss of speech quality.

References


AN EFFICIENT CODEBOOK STRUCTURE FOR CELP

Brian P. MURRAY and Anthony D. FAGAN

Department of Electronic and Electrical Engineering,
University College Dublin,
Ireland
Tel / Fax +353-1-7061925 / 2830921,
E-Mail bmurray@faraday.ucd.ie

Abstract. A new sparse overlapping codebook structure is proposed for a CELP coder that takes particular advantage of the architecture of a typical DSP coder implementation to reduce the complexity of the cross-correlation term and the energy term in the codebook search, without a reduction in performance. For a codebook of dimension $N$, size $L$, with 80% sparsity, and a shift of 2 samples between successive codebook vectors, the complexity of the cross-correlation term (excluding overhead) is reduced to $0.3NL$ MAC operations and the complexity of the energy term is reduced to $0.3(0.8R.L)$ MAC operations where $R$ is the length of the truncated impulse response. This is a reduction in the search complexity by a factor of 3 over the usual random sparse overlapping codebook.

1. Introduction

At low bit rates analysis by synthesis coders e.g. CELP [1] give very good reconstructed speech quality at the cost of a very high computational load. The most significant portions of the computational load are the codebook searches. Typically a CELP coder uses a short-time and a long-time linear predictor and a random codebook, the codebook is searched using a perceptually weighted error criterion [2].

The codebook search is outlined as follows, the ringing due to the memory of the linear predictors is subtracted from the weighted speech signal $s_n$ to give a target vector $t_n$, the codebook vector that minimises the weighted error signal $e_n$ is chosen as the optimum vector.

The error term for the $j$th codebook vector $c_n^j$ can be calculated using (1) [3], where $A_n$ is the zero memory impulse response of the weighted synthesis filter, whose coefficients are $a_k$.

$$R_j = \sum_{n=0}^{N-1} c_n^j ( \sum_{n=0}^{N-1} t_n A_{n,n} )$$

However, the calculation of the energy term (i.e. the denominator term) is much more intensive, this is because of the recursive filtering, thus $(p + 1)NL$ MAC operations are required. A number of methods have been proposed that significantly reduce the computational complexity of the energy term e.g. by using SVD (singular valued decomposition) or the FFT [4], however the calculation of the additional side information is high.

2. Efficient Codebooks for CELP

In recent years a number of different efficient codebook structures have been proposed that greatly reduce the complexity of the energy term, the most important of these is the overlapping codebook [5]. Typically a random overlapping codebook with a shift of 2 samples between vectors is used with CELP coders, this codebook is generated from a random sequence of numbers $v_n$ of length $2L + N - 2$.

Because the codebook is overlapping the recursive end-point correction algorithm (3).

$$E_{t_{n+1}} = E_{t_n} - c_{n}^j ( \rho_n + \sum_{i=1}^{N-1} c_{i,n}^j \rho_n ) + c_{n+1,n}^j ( c_{n+1,n}^j \rho_n + 2 \sum_{i=1}^{N-1} c_{n+1,i}^j \rho_n )$$

where $\rho_n$ is the error at $n$. The complexity of the final recursion is approximately $0.3N^2 L$ MAC operations.
can be used to reduce the complexity of calculating the energy term to 4RL MAC operations [3], where the impulse response of the weighted filter has been truncated to length \( R \) and \( \rho_n \) is the autocorrelation of the truncated impulse response.

Equation (3) calculates the change in the energy term for successive vectors due to element \( c_{j-1} \) being dropped and element \( c_{j+1} \) being added. For a shift of 2 samples between successive codebook vectors equation (3) is used twice. The complexity of the energy term can be reduced further by using a random sparse overlapping codebook, typically a codebook with 80% sparsity is used. Hence if either \( c_{j-1} \) or \( c_{j+1} \) are equal to zero, the corresponding terms inside the brackets do not need to be calculated, and thus the complexity is reduced to 0.8RL MAC operations plus overhead for the test operation. A further reduction in storage (but not complexity\(^1\)) can be achieved for a VLSI implementation by using a ternary valued codebook (-1, 0, +1).

These codebook yields a great reduction in storage and complexity compared to a random codebook without reducing the synthesised speech quality [1, 6]. For a CELP coder using \( N = 60 \) and \( R = 30 \) the complexity of the energy term is now reduced to half the complexity of the cross-correlation term.

3. Further Complexity Reduction

Though great success has been achieved with these codebooks in reducing the complexity of the energy term, no reduction has been achieved in the complexity of the cross-correlation term. An examination of equations (2) and (3) show that the calculation of the these terms requires the correlation of a random sparse vector with different fixed sequences. It is clear that there are a large number of multiply by zero and accumulate operations used in each of these correlations, these are obviously superfluous.

It has sometimes been claimed that advantage can be taken of the presence of a large number of zeroes to reduce the complexity of these correlations [7], however with a DSP implementation this is not efficient. This is because a DSP architecture is optimised for correlation and convolution operations, usually two address increments are implemented in parallel with pipelined data moves and a MAC operation. A test/skip operation to identify a zero would require more operations that the MAC operation itself [8, 9].

If the position of each non-zero element is stored, it is then only necessary to perform the MAC operations for those elements. However, because of the pipeline delays in loading up addressing pointers and data moves, this method is only efficient if there are only one or two non-zero elements per vector. Unfortunately if the sparseness of the codebook is reduced to only one or two non-zero elements per vector (i.e. > 95% sparsity) the speech quality is reduced. With the usual 12 or so non-zero elements per vector this method is not feasible.

The crux of the problem is how can you feed the appropriate data (i.e. the non-zero elements) to the MAC unit as part of an efficient data pipeline. The key to this is to take full advantage of the possibilities allowed in a DSP architecture.

4. Typical DSP Architecture

A DSP architecture can implement a MAC operation in parallel with two address pointer increments and two data moves all in a single cycle. In addition, zero overhead looping and modulo addressing is provided. These are particularly useful for filter or correlation operations where at the end of the loop, the end-of-loop test is performed in hardware and the coefficient pointer automatically wraps back around to the first coefficient again. This allows the \( N \) point correlations to be implemented in \( N \) cycles with only a very small overhead. An address increment of 1 is used for each address pointer and modulo \( N \) or \( R \) addressing is used for the fixed sequence.

Most DSP architectures have a number of different address pointers (usually 8 or more) and provide a very rich set of address update operations. As well as supporting pre and post increment and decrement by 1, the address generation hardware will usually also allow an address increment or decrement by fixed offset \( p \). By taking advantage of the these features it is possible to design a random like sparse codebook where the non-zero elements can be efficiently addressed.

5. New Efficient Codebook Structure

A new efficient sparse overlapping ternary codebook is now proposed. This codebook has a shift of 2 samples between vectors and is generated using (4) from a sequence of numbers \( v_n \) of length \( 2L+N-2 \) which have a particular structure:

\[
c_{j}'_{n} = v_{n+2j} \quad 0 \leq j \leq L - 1
\]

(4)

Every \( p^{th}, q^{th} \) and \( r^{th} \) element of the sequence \( v_n \) has one of three values, -1, 0, +1, (with a high percentage of these being non-zero), where \( p, q \) and \( r \) are prime numbers. All other elements of \( v_n \) are zero. This sequence of numbers has the appearance of randomness but has an underlying structure that can be efficiently exploited using a DSP architecture.

With this codebook there are no more than \( m \) non-zero elements per vector where \( m = Np + Nq + Nr + 3 \).

---

\(^1\)A MAC unit is already required for the adaptive codebook search and other algorithms so there is no advantage in reducing the multiply to an add/subtract.
Because it is possible to increment an address pointer by a fixed amount, just as efficiently as an increment by 1, the cross-correlation term can be calculated with just \( m' \) MAC operations per codebook vector. Likewise each correlation in the energy term, equation (3), can be calculated with just \( m' \) MAC operations per codebook vector where \( m' = Rl/p + Rl/q + Rl/r + 3 \).

The codebook is stored as three compact sequences of numbers, \( v^p_n \), \( v^q_n \) and \( v^r_n \), i.e. every \( p^{th}, q^{th} \) and \( r^{th} \) element of \( v_n \). These sequences are of length \( N/p+1 \), \( N/q+1 \) and \( N/r+1 \) respectively. Thus

\[
v_n = v^p_{np} + v^q_{nq} + v^r_{nr}
\]

(5)

where \( v^p_{np} \) if \( n/p \) is not an integer. A restriction must be placed on the non-zero values of the compact sequences to ensure that no elements of \( v_n \) have values other than -1, 0 or +1, this can happen for elements such that more than one of \( p, q \) or \( r \) are factors of \( n \).

5.1 Calculation of the Cross-correlation Term

The cross-correlation term can be efficiently calculated as follows,

\[
R^t = \sum_{i=0}^{N/p-1} x_{p+i} v_{n[p+i]}^p + \sum_{i=0}^{N/q-1} x_{q+i} v_{n[q+i]}^q + \sum_{i=0}^{N/r-1} x_{r+i} v_{n[r+i]}^r
\]

(6)

where \( \lfloor 2j/p \rfloor \) is the highest integer \( < 2j/p \), \( x_n \) is the backward filtered target vector and is given by,

\[
x_n = \sum_{i=0}^{N-1-n} A_{i+n}, \quad 0 \leq n \leq N-1
\]

(7)

Now \( o_p^t \) is an offset that is decremented by 2 between successive vectors and is given by

\[
o_p^t = p - 1 - 2j_{\text{MOD}(p)}
\]

(8)

and likewise for \( q \) and \( r \).

In practice the calculation of pointers is much simpler than initially appears, the codebook has been deliberately constructed so that the pointer calculations required in the codebook search, can be efficiently implemented with a DSP architecture.

Each of the sequences, e.g. \( v^p_n \), is accessed using linear addressing with an increment of 1 and a decrement of \( p \) at the end of the each summation to return the pointer back to the start for the next summation. The sequence \( x_n \) is accessed using modulo \( M^p \) addressing where,

\[
M^p = \lfloor N/p+1 \rfloor p + 2
\]

(9)

and a fixed increment of \( p \) is used each cycle. With this arrangement the offset is automatically calculated, in fact, all of these address calculation can be calculated in parallel with the MAC operations with no overhead (or additional instructions). One operation is required to store the result of the summation. Two out of every \( p \) times an additional increment at the end of the summation is required for each of the pointers, this happens each time a non-zero element is dropped due to the shift of 2. This occurrence is regular for each sequence. The correlation can be calculated as three partial sums, i.e. first calculate the contribution to each \( R^t \) due to \( v_n^p \), and then the contribution due to \( v_n^q \) and \( v_n^r \). This can be efficiently coded, though an additional operation is required to read the previous partial sums.

5.2 Calculation of the Energy Term

The recursive end-point correction algorithm (3) is used to calculate the energy term for the structured codebook. Each of the correlations in (3) can be efficiently calculated in a very similar way to the calculation of the cross-correlation calculation. About 80% of the time the correlation calculation does not need to be performed because the element being added or dropped is zero.

As with the cross-correlation term each correlation is calculated as the sum of three correlations, i.e. \( R^t \) with \( v_n^p \), \( v_n^q \) and \( v_n^r \). The recursive end point correction algorithm can thus be broken down into the sum of 6 partial sums. The most efficient way to test if a non-zero element has been added or dropped depends on the exact details of the architecture, the choice is between testing the sequence \( v_n \) or the pointers to \( v_n^p \), \( v_n^q \) and \( v_n^r \), for zeros. If a non-zero element has not been added or dropped then the only operations required is the incrementing and decrementing of the pointers, because there are more pointers required with a structured codebook there is a greater overhead than with a random codebook.

5.3 Computational Estimates

With a random 80% sparse overlapping codebook of dimension \( N \) and size \( L \) the calculation of the cross-correlation term requires \( N.L \) MAC operations and the calculation of the energy term requires \( 0.8.R.L \) MAC operations. There are additional calculations required to search for the maximum value, and there is also overhead for various for load and store operations and pointer initialisation. However the MAC operations represent about 70-80% of the computational load for the codebook search, and in turn the codebook search is about 2/3 of the total (full duplex) computation load [6].

For values of \( p = 7, q = 11 \) and \( r = 13 \) with 80% sparsity (i.e. about 71% of the elements of the compact sequences are non-zero), and with \( N = 60 \) and \( R = 30 \) there are 20 MAC operations required for the cross-correlation
term (6), and 11 MAC operations for each correlation in the energy term. Hence the calculation of the cross-correlation term requires $0.33NL$ MAC operations and the calculation of the energy term requires $0.36(0.8R)N$ MAC operations. This is a reduction by a factor of about 3. For $L = 1024$ the number of MIPS required for the calculation of the energy term and the cross-correlation term is about 12 MIPS with the random codebook and about 4 MIPS with the structured codebook. Using the figures given in [6] as an estimate for the computational load of the rest of a CELP coder, the structured codebook yields a reduction in the overall computational complexity of about 30% and 40% for a codebooks of size 512 and 1024 respectively.

6. Results

A CELP coder has been implemented using a $12^{th}$ order linear predictor with a frame size of 30 ms and either 4 sub-frames of 7.5 ms each, or 6 sub-frames of 5 ms each. Closed loop pitch prediction with a minimum pitch delay of 20 and a maximum delay of 147 is used, with a fixed codebook of size 1024. The codebook is a structured sparse ternary overlapping codebook, with $p = 7$, $q = 11$ and $r = 13$ and 80% sparsity. The predictor coefficients are quantized to 42 bits, the pitch coefficient to 6 bits, and the gain coefficient to 6 bits, this corresponds to 5.3 kb/s and 7.2 kb/s for a sub-frames size of 7.5 ms and 5 ms respectively. These coders were compared with identical coders that used a random sparse ternary overlapping codebook with 80% sparsity.

Informal subjective tests were carried out using one male speakers and two female speakers, with speech segments of length 20 seconds each. For both the 5.3 kb/s and 7.2 kb/s coders it was very difficult if at all possible to distinguish between the speech quality of the sparse random codebook and the structured codebook. In addition, for every case, the weighted segmental SNR's of the coders were within 0.1 dB of each other (assuming the same bit rate and speech segment). These coders were also compared with a coder using a 95% sparse random codebook, there was a noticeable reduction in speech quality for the 5.3 kb/s coders as would be expected.

There are obviously a number of other variations of the above structured sparse codebook, e.g. a codebook with $p = 11$, $q = 13$ and $r = 17$ and 80% sparsity yields a complexity reduction by a factor of 4, or a codebook with $p = 5$ and $q = 7$, (with no third sequence) and 80% sparsity, also yields a complexity reduction by a factor of 3 but with less overhead. Some tests were conducted with these codebooks, once again it is difficult to distinguish between the coders, but further study is required.

7. Conclusion

A new efficient codebook structure for a CELP coder has been presented which reduces the codebook search complexity by a factor of 3, by taking advantage of the architecture of a modern DSP chip. The complexity of both the cross-correlation term and the energy term have been reduced. This has been achieved without compromising the speech quality.

References


Improving CELP Voice Quality by Weighting the Error Criterion

Miguel A. FERRER-BALLESTER†, Aníbal R. FIGUEIRAS-VIDAL††

† ETSI Telecommunicación, Univ. Las Palmas, Campus de Tajira, 35017 Las Palmas, Spain.
Tel/Fax: +34 28 451250 / 451243 E-mail: mferrer@ei.teleco.ulpgc.es
†† GPSS-DSSR, ETSI Telecom-UPM, Ciudad Universitaria, 28040 Madrid, Spain.
Tel/ Fax: +34.1.3367226 / 3367350 E-mail: wernaga@gits.upm.es

Abstract. It is well-known the limited quality of the reconstructed speech by Code-Excited Linear Prediction (CELP) speech coder at low bit-rate (2400-8000 bps). The major source of audible distortion has been attributed to an inaccurate degree of periodicity of the voiced speech signal. In the present paper we alleviate this drawback by modifying the MSE distance measure, which is not able to capture the periodicity adequately. The new proposed distance measure is an MSE distance weighted with a function of the LP residue energy. The aim of the proposed measure is reducing the maxima of the codification-error energy of the CELP-coded residue in the pitch-pulses neighbourhood. The improvement of the quality of the speech reconstructed by the CELP with weighted MSE distance measure has been checked by subjective A-B test. This result emphasizes the perceptual importance of an adequate description of the pitch-pulse waveform of the original (uncoded) LP residue by the CELP-coded one.

1. Introduction

Speech coding algorithms operating at rates between 2.4 and 8.0 Kb/s have been significantly improved during the last decade [1]. Among the current coders operating at these rates, this paper pays attention to the Code-Excited Linear Prediction (CELP) [2], a type of linear-prediction (LP) speech coding algorithm based on analysis-by-synthesis techniques.

Briefly, in CELP the excitation for the all-pole synthesis filter is modelled as a sum of two gain-scaled vectors [2]: the first vector is obtained from an adaptive codebook that contains the past excitation [3]; the second vector is obtained from a fixed stochastic codebook. The two excitation vectors are selected by minimizing a perceptually-weighted error between the original and the reconstructed speech [2]. A more detailed description (with a block diagram) of this algorithm could be found in Appendix A at the end of this paper.

The intelligibility of the speech reconstructed by this type of coders is high; however, a drawback for many practical applications is the increasing unnatural speech quality with decreasing bit rate.

This unnatural speech quality is mainly caused by three types of distortion, which are usually described as noise, reverberation, and tonal artifacts [4]. All three are closely connected with the periodicity of the voiced speech signal. Then, it can be said that the major source of audible distortion is mainly caused by an inaccurate degree of periodicity of the voiced speech signal [4].

Some solutions, which change the CELP-coder structure, have been proposed in order to allow obtaining a highly periodic excitation; for instance, Grazow et al. have proposed in [5] to synthesize periodic speech using single-pulse excitation, and Wang and Gerstho [6] have coded each frame taking into account the local phonetic content. Other approaches to avoid the problem are based on a sinusoidal representation of the speech signal [7]. All the above schemes include voiced/unvoiced decision.

In this paper, in order to improve the speech periodicity, we pay attention to the traditional objective measure: the signal-to-noise ratio (SNR). It is well known that the SNR is not able to capture the periodicity adequately; in fact, even in waveform coders such as CELP it has been found advantageous to modify the reconstructed voiced speech signal to have a higher level of periodicity, although this results in a lower instantaneous SNR [8].

Considering the above reason, we have modified the CELP error criterion by weighting the MSE measure with a function of the LP residue energy. The proposed weighting function reduces the codification-error energy at the pitch-pulses neighbourhood of the CELP-coded residue, improving the CELP-speech periodicity. This approach has the advantage of improving the quality of the final CELP-speech neither changing the CELP-coder structure or including the voice/unvoiced decision. Additionally, the computational load is not significantly increased.

The organization of this paper is as follows: in the next Section we describe the proposed weighted error criterion. Section 3 presents graphical, subjective and objective results. After it, the conclusions in Section 4 close the paper. An appendix which describes the CELP speech coder algorithm is included.

2. The weighted MSE distance measure

The limited quality of the MSE criterion in the reconstruction of the periodic signal could be seen from the function of the codification-error energy of the CELP-coded residue, which is defined in eq.(1) as $E_{cod}(n)$ and depicted in Figure 1.
where $r(n)$ is the original (uncoded) LP residue, $r'(n)$ is the CELP-coded residue, $N$ the subframe length. We have found that four is a good value for $M$.

In Figure 1, we can see how the $E_{RCE}(n)$ has its maxima around the pitch-pulses of the LP residue. This difficult the possibility of reconstructing the pitch-pulses waveform in a correct way, and, therefore, of describing correctly the dynamics of the voice pitch-pulses waveform (the waveform of one pitch cycle) along time, which is very important in the description of the periodicity of the speech signal [4].

The contribution of this paper consists on a modification that allows a better consideration of voice periodicity by smoothing the $E_{RCE}(n)$ function by means of weighting the MSE criterion (WMSE) as in eq. (2):

$$WMSE = \sum_{n=1}^{N} W(n)[x(n)-y(n)]^2$$

where $x(n)$ and $y(n)$ are the original and CELP-coded perceptually-weighted residue, respectively, $W(n)$ is the MSE-weighting function, and $N$ is the CELP subframe length.

The determination of the weighting function has been made heuristically. It is based in emphasizing an error in the pitch-pulses neighbourhood of the LP residue with respect to the rest of the frame. After different tentative functions, a good result has been obtained from weighting function $W(n)$ of eq.(3):

$$E_x(n) = \sum_{m=-M}^{M} x^2(n+m); 1 \leq m \leq N$$

$$E_y(n) = \sum_{m=-M}^{M} y^2(n+m); 1 \leq n \leq N$$

$$W(n) = \frac{\max[E_x(n), E_y(n)]}{\max[E_x(m), E_y(m); 1 \leq m \leq N]} ; 1 \leq n \leq N$$

where $\max\{a,b,\ldots\}$ equals the largest value of its terms and the other variables have been already defined. We have found that four is a good value for $M$. The computation load increase of the CELP with WMSE criterion is negligible (less than 3% with respect to standard CELP).

The choice of this way of modifying the error measure has the following advantages: 1) it does not need additional bits, 2) the additional computation charge is very little, and 3) does not modify the CELP-coder structure. Also, in the unvoiced frames, we expect the quality remains unchanged since in these frames the weighting function is nearly constant. In the same way, the new distance measure does not modify the CELP robustness against additive noise.

3. Results

In this section we present results of the WMSE CELP-coder in comparison with the 4800 bps standard CELP-coder [3]. The parameters have been coded in the same way for both schemes. The results are classified in three subsections: 1) graphical views of the improvement of the WMSE CELP-coder, 2) perceptual comparisons, and 3) objective measures.

3.1 Graphical comparisons

A comparison between the $E_{RCE}(n)$ functions obtained from the standard CELP and the WMSE CELP can be seen in Figure 2a. This figure shows how the final objective has been reached: to smooth the function of the codification-error energy of the CELP-coded residue by reducing its maxima. Consistently, the WMSE CELP-coder reproduces the original LP residue better than the MSE CELP-coder, as we can see in the example of Figure 2b.

The penalty of the above improvement is a slight increase of the error of codification in some unvoiced frames (see Figure 3) of the WMSE CELP speech coder.

3.2 Subjective comparison

In order to obtain a subjective comparison, we have generated two versions of coded speech: one from a 4800 bps standard CELP, and the other from the 4800 WMSE CELP. The coding procedures have been the same for both schemes. Then, the difference between the two coded versions of speech can be attributed here to the different distance measure.

Subjective quality evaluation has been done here through informal A-B or pair comparison test [9] using 10 listeners. Six Spanish sentences (spoken by three male and three female
speakers recorded from different sources: microphone, AM and FM radio station) have been used for evaluation. Each comparison is done between the two coded versions of a sentence. All possible A-B pairs are generated and presented in a randomized order. Listener's task has been to prefer either one or the other of the two-coded versions, or to indicate no preference. Results from these informal tests have shown that 80% indicated WMSE CELP preference, 15% CELP preference, and 5% indifference. From this, we have concluded that, in this experiment, the 4800 bps WMSE CELP-coder is preferred over the 4800 bps standard CELP.

### 3.3 Objective results

The objective evaluation has been done with the segmental signal-to-noise ratio (SEGSNR) [9]. As expected, the SEGSNR obtained with the WMSE distance measure has been slightly less (about 0.5 dB) that the SEGSNR offered by MSE standard CELP (about 8.0 dB). A reason of the above is that the WMSE CELP does not search to minimize the MSE distance.

### 4. Conclusions

We have introduced a modification, by weighting, of the MSE distance measure for CELP speech coder, to improve the degree of periodicity in the reconstructed CELP speech signal. The weighting function task is to smooth the energy of the error of codification of the CELP-coded residue, which has maxima in pitch-pulse environments.

The subjective improvement of the reconstructed WMSE CELP-coder speech quality has been checked by A-B test. This result emphasizes the importance of the correct description of the original LP-residual pitch-pulse waveform by the CELP-coder. This improvement is due to the better selection (in a perceptual sense) of the codevectors from the stochastic and adaptive codebooks that the WMSE criterion allows. This advantage do not require additional bits nor a significant increase of the computational load.

### References


### Appendix A

The CELP speech coder description

In this appendix we follow the notation of [8]. The CELP-coder procedure (a block diagram is depicted in Figure 3), as is follows: speech signal $s(n)$ is divided into frames of length $L$ and subframes of length $N$ (usually $L/N=4$). Each subframe is filtered by a pole-zero, noise-weighting linear filter to obtain $X(z)=S(z)A(z)/W(z)$, where $x(n)$ is the target signal for the coding process, $A(z)=1+a(1)z^{-1}+...+a(p)z^{-p}$ (with $p=10$) is the standard LP polynomial corresponding to the current frame, and $W(z)=1+a(1)z^{-1}+...+a(p)z^{-p}$ is a modified polynomial obtained from $A(z)$ by shifting the zeroes towards the origin in the $z$-plane, i.e., by using the coefficients $a(i)=a(i)'$ with $0<\gamma<1$ (typical value: $\gamma=0.8$).

The coder attempts to synthesize, at every frame, a signal $y(n)$ as close to the target signal $x(n)$ as possible, using a Mean-Square-Error (MSE) sense. The synthesis algorithm is based on the following equations.

$$r'(n,T)=r'(n-kT) ; k=0,1,2,...,N \quad 1 \leq n \leq N$$

$$y(n)=y_0(n)+\beta r'(n,T)xh(n)+g(n)\beta xh(n) ; 1 \leq n \leq N$$  \hspace{1cm} (5)

where $*$ denotes the convolution operation, $h(n)$ is the impulse response of the $W(z)$ filter in the range [0,...,N-1], $y_0(n)$ is the initial state response (without any input), and $r'(n)$ is the past coded LP excitation.

The other variables are the parameters that code the LP residue, and are divided into two groups. The first group, called adaptive contribution, is formed by pitch gain $\beta$ and pitch delay $T$. Note that $r'(n,T)$ is a codebook of vectors (as many as possible values of $T$) which is updated every subframe: it is called the adaptive codebook. The second group, called stochastic contribution, is formed by the stochastic vector gain $g$ and the stochastic codebook label $l$. Each vector $c(n,l)$ of the stochastic codebook contains a white Gaussian sequence of zero mean and unit variance. The possible values of these parameters are contained in several tables: the size of these tables determines the number of bits available to the system for synthesizing the coded voice.

A key issue in CELP coding is the strategy of selecting a good set of parameters from the various codebooks to minimize the distance between the target and the synthesized signal. An exhaustive search, although possible in principle, is prohibitively complex. Therefore, sub-optimal procedures are used. The usual strategy is to separate pitch parameters $T$ and $\beta$ from stochastic parameters $g$ and $l$, and to select the two groups independently. $T$ and $\beta$ are found first and, after it, the best $g$ and $l$ are found. Three steps are required:

1. Find best $T$ and $\beta$ according to eq.(6):

$$T^*,\beta^*=\arg\min_{T,\beta} \sum_{n=1}^{N} [x(n) - y_0(n) - \beta r'(n,T)xh(n)]^2$$ \hspace{1cm} (6)

2. Calculate the resulting error signal with eq.(7):

$$d(n)=x(n)-y_0(n)-\beta^* r'(n,T^*)xh(n) ; 1 \leq n \leq N$$ \hspace{1cm} (7)

3. Finally, the coder tries to find $g$ and $l$ which best match $d(n)$ according to eq.(8):

$$g^*,l^*=\arg\min_{g,l} \sum_{n=1}^{N} [d(n) - g \cdot c(n,l)xh(n)]^2$$ \hspace{1cm} (8)

After getting the parameters, we synthesize the CELP voice by the means of

$$r'(n)=\beta^* r'(n,T^*)+g^* c(n,l^*) \quad ; 1 \leq n \leq N$$ \hspace{1cm} (9)

$$\sum_{i=0}^{P} a_i z(n-i)=r'(n)$$ \hspace{1cm} (10)

where $z(n)$ is the CELP synthesized voice.
Combined CELP Speech Coding and Quadrature Amplitude Modulation

Arild FULDSETH

Department of Telecommunications, The Norwegian Institute of Technology, N-7034 Trondheim, Norway,
Tel/Fax: +47 73 59 26 50/47 73 59 26 40, E-Mail: arild@tele.unit.no

Abstract. In this paper, a method for optimizing the index map between the output indices of a wideband (7 kHz) CELP speech coder and the channel symbols for a power constrained channel and multilevel signaling is presented. The method is based on simulated annealing and makes use of dynamic scaling of a QAM signal constellation during the optimization procedure in order to optimize for a fixed transmitted power. The resulting system is shown to give a significant gain in terms of channel signal-to-noise ratio when compared to a reference system based on arbitrary index maps.

1. Introduction

Recently, there has been a lot of work in the area of combined source and channel coding. Most of this work has focused on the transmission of samples from a discrete source over a binary symmetric channel (BSC) without power constraints [1-4]. In [1] simulated annealing (SA) was used to find a good index map between the output indices from a vector quantizer (VQ) and the binary codewords of a BSC. In this case, the optimum index map is the map which minimizes the influence of bit errors in the reconstructed source symbols. The simulated annealing algorithm results in an index map corresponding to at least a local optimum in a finite number of iterations. In [4] the SA algorithm was used to find the index map for a CELP speech coder.

This paper extends the work in [1,4] to include channels with a power constraint. It is shown how the optimization procedure based on SA can be modified taking into account the power constraint with multilevel signaling and maximum likelihood (ML) demodulation. Given the fact that the source symbols have unequal probabilities, a good index map should also result in as low average transmitted power as possible. Equivalently, for a fixed average transmitted power, the index map should permit as large minimum distance \( d_{min} \) of the signal constellation as possible. This additional property of the index map is achieved by modifying the method in [1,4] to include dynamic scaling of the signal constellation during the optimization. The procedure is applied to the quantizer indices of a wideband CELP speech coder.

The rest of the paper is organized as follows. In Section 2, a description of the CELP coder structure as well as the transmission system to be considered is given. The proposed method for optimizing the index map between the CELP coder parameters and the channel symbols is described in detail in Section 3 followed by a description of the experiments and the results in Section 4. Finally, the conclusions are given in Section 5.

2. System Description

2.1. The CELP Coder Structure

The CELP coder which operates at a sampling rate of 16 kHz with a bit rate of 16 kbit/s is described in [5]. However, in order to fit into a quadrature amplitude modulation (QAM) signaling scheme, a few changes in the parameter representation and the bitallocation are made. A 20th order short-term filter is updated every 20 ms and the filter parameters are encoded by use of line spectrum pair (LSP) differences [6] with 4 bits per parameter. For the long-term prediction (LTP) filter we use 8 bits for the lag value and 4 bits for the filter coefficient. Note that the LTP lag is interpreted as an index of an adaptive codebook. Finally, we use 8 and 4 bits for the stochastic index and gain, respectively. The LTP parameters as well as the stochastic codebook index and gain are updated every 2 ms. The coder configuration is shown in Table 1.

2.2. The Transmission System

Consider the transmission system of Figure 1. An \( L \)-dimensional amplitude continuous vector \( x \) is mapped to an index \( i \) belonging to an index set \( I = \{0, \ldots, N-1 \} \) by use of a quantizer \( Q \) of size \( N \). The vector \( x \) could represent any of the unquantized CELP coder parameters such as the gain factors, the LSP differences (\( L=1 \)), or any combination of these. Alternatively, the quantizer could represent the adaptive codebook or the stochastic codebook (\( L=32 \)). Next, the index \( i \) is mapped to an index \( k = m(i) \) corresponding to a \( K \)-dimensional (\( K \) even) channel space vector \( c_k \) of a signal constellation, \( C = \{c_0, \ldots, c_{N-1} \} \). We will refer to \( m(i) \) as the index map. Note that the case \( K = 2 \) corresponds to one QAM symbol, while for \( K > 2 \) the signal constellation can be realized by considering several consecutive QAM symbols in time as one entity.
### Table 1 CEP coder configuration

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Frame length [ms]</th>
<th>Frame rate</th>
<th>No. of bits/frame</th>
<th>Bit rate [bits/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSP (20 parameters)</td>
<td>20</td>
<td>50</td>
<td>80 (20 x 4)</td>
<td>4000</td>
</tr>
<tr>
<td>LTP gain</td>
<td>2</td>
<td>500</td>
<td>4</td>
<td>2000</td>
</tr>
<tr>
<td>LTP index</td>
<td>2</td>
<td>500</td>
<td>8</td>
<td>4000</td>
</tr>
<tr>
<td>Stochastic codebook gain</td>
<td>2</td>
<td>500</td>
<td>4</td>
<td>2000</td>
</tr>
<tr>
<td>Stochastic codebook index</td>
<td>2</td>
<td>500</td>
<td>8</td>
<td>4000</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>16000</td>
</tr>
</tbody>
</table>

### Figure 1 General transmission system block diagram.

The signal constellation $C$ is normalized to unit minimum distance $d_{\text{min,0}} = 1$, and the desired average transmitted power is obtained by use of a scale factor $\alpha(m)$. Assuming a given average transmitted energy $\varepsilon$ per $K$-dimensional channel symbol, the scale factor $\alpha(m)$ is given by

$$\alpha(m) = \frac{\varepsilon}{\sum_{i=0}^{N-1} P(i) \epsilon^T m(i) e m(i)}$$  

where $P(i)$ is the probability of index $i$. Note that for a given value of the transmitted power $\varepsilon$, the scale factor $\alpha(m)$ depends on the index map $m(i)$. Furthermore, the vectors $\alpha(m) e_k$, $k=0,\ldots,N-1$ define a signal constellation $C'$ with minimum distance $d_{\text{min}} = \alpha(m)d_{\text{min,0}}$.

The transmitted vector $y = \alpha(m)e_k$ is corrupted by an AWGN channel which means that a $K$-dimensional Gaussian noise vector $q$ with zero mean and variance per component $\sigma_q^2$, is added to the transmitted vector $y$. Thus, the received vector $y'$ is given by $y' = y + q$. At the receiver side, an ML demodulator is used to obtain a channel index $k'$. The ML demodulator is followed by an inverse index map $m^{-1}$ to determine the quantizer index $i$. Finally, the reconstructed vector $x' = s_{i'}$ is chosen from the set of quantizer reconstruction symbols $S = \{s_0,s_1,\ldots,s_{N-1}\}$ by the inverse quantizer $Q^{-1}$. Note that we without loss of generality, can assume that the indexing of the channel space vectors $c_k$ is fixed.

### 3. Optimizing the Index Maps

Given the CEP coder configuration of Table 1, and the transmission system of Figure 1, the goal is to design an index map for each parameter such that the influence of channel errors in the reconstructed speech signal is as small as possible. To achieve this, it is obvious that the desired index maps should have the following two properties:

- To minimize the effects of transmission errors, symbols that are close in the source space should be close in the channel space.
- To maximize the minimum distance $d_{\text{min}}$ of the signal constellation $C'$ for a given transmitted energy, the most probable source symbols should be mapped to the channel symbols having the smallest amplitude.

The first requirement is also encountered when considering discrete channels without a power constraint [1,4]. Following a similar approach, we choose to minimize the following error criterion

$$\epsilon(m) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} Y(m(j) m(i)) D(s_i, s_j)$$  

929
where $Y(m(j)|m(i))$ is the conditional probability of receiving symbol $m(j)$ given that the symbol $m(i)$ was transmitted and where $D(s_i, s_j)$ is some distortion measure between the quantizer reconstruction vectors $s_i$ and $s_j$. In this work, we have chosen to use the squared error distortion measure

$$D(s_i, s_j) = \frac{1}{L} (s_i - s_j)^T (s_i - s_j)$$

(3)

An exception is the adaptive codebook where the reconstruction vectors $s_i$ vary with time. In this case, we use the expected value of the squared error

$$D(s_i, s_j) = \frac{1}{L} E\left[(s_i - s_j)^T (s_i - s_j)\right]$$

(4)

The expectation value is estimated by averaging over a finite length speech material.

Note that in [4], a more sophisticated distortion measure based on the original and the reconstructed speech signals is used. However, since the main purpose of this paper is to extend the method to channels with a power constraint, we have chosen to use the distortion measures of eq. (3) and eq. (4).

The SA algorithm is used to find an index map corresponding to at least a local minimum of the error criterion in eq. (2). The algorithm is based on iterative perturbations of the index map $m_i$, evaluating the error criterion in eq. (2) at each iteration. Depending on the change in $e(m_i)$ the new index map is accepted or rejected at each iterative stage. One of the advantages with the SA algorithm is the possibility of accepting an index map corresponding to an increase in the error criterion $e(m)$. This property reduces the probability of getting trapped in a poor local minimum.

The method described above cannot be applied directly to a channel with a power constraint and multilevel signaling. This is due to the fact that for a fixed average transmitted power and unequal symbol probabilities the gain factor $\alpha(m)$ as well as the channel transition probabilities $Y(m(i)|m(j))$ depend on the particular index map. Thus, we have

$$e(m) = \sum_{i=0}^{N-1} P(i) \sum_{j=0}^{N-1} Y(m(j)|m(i)) D(s_i, s_j)$$

(5)

Ideally one would like to estimate the true transition probabilities $Y_{\text{true}}(m(j)|m(i))$ by Monte Carlo simulation at each iteration of the SA algorithm. This would automatically result in an index map with the best trade-off between the two requirements mentioned above. However, in order to reduce the computational complexity, we propose a suboptimal, but computational efficient solution to the problem.

Using the fact that $D(s_i, s_j) = 0$, the approximate solution is obtained by rewriting eq. (5) as

$$e(m) = \sum_{i=0}^{N-1} P(i) \sum_{j=0}^{N-1} \beta_{i,j} Y_0(m(j)|m(i)) D(s_i, s_j)$$

(6)

where

$$\beta_{i,j} = \frac{Y_0(m(j)|m(i))}{Y_0(m(i)|m(i))}$$

(7)

and where $Y_0(m(j)|m(i))$ are the transition probabilities corresponding to a fixed gain factor $\alpha_0$ defined as

$$\alpha_0 = \sqrt{\sum_{i=0}^{N-1} \frac{1}{N} \sum_{i=0}^{N-1} e^T m(i) e m(i)}$$

(8)

Next, $\beta_{i,j}$ is approximated by

$$\beta_{i,j} = \gamma = \frac{Q(\alpha_0/2\sigma_q)}{Q(\alpha_0/2\sigma_q)}$$

(9)

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

(10)

Thus, we obtain a final expression for our approximation

$$e(m) = \gamma \sum_{i=0}^{N-1} P(i) \sum_{j=0}^{N-1} Y_0(m(j)|m(i)) D(s_i, s_j)$$

(11)

The computation of a new value of the scale factor $\gamma$ at each iteration of the SA algorithm corresponds to a dynamic scaling of the signal constellation $C'$.

4. Experiments and Results

The method described above was evaluated for both 16-QAM and for 256-QAM signal constellations. For 16-QAM, the LSP differences as well as the gain factors used one QAM symbol each, while each of the codebook indices were mapped to two consecutive QAM symbols. For 256-QAM, the codebook indices were mapped to one QAM symbol each, while the LSP differences were grouped into pairs each, being mapped to one QAM symbol. Finally, the two gain factors were mapped into the same QAM symbol. The configurations for 16-QAM and 256-QAM are shown in Table 2.

When optimizing the index maps the probabilities $P(i)$ as well as the distortion measure of eq. (4) were estimated from 192 sec. of speech from 6 male and 6 female speakers.

In Figure 2 and Figure 3 we show the segmental signal-to-noise ratio (SNR) as a function of the channel signal-to-noise ratio (CSNR). For evaluation purposes we used 180 sec. of speech from 2 male and 2 female speakers. As a reference we have

930
Table 2 Configuration for 16 QAM and 256 QAM

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Frame length [ms]</th>
<th>Frame rate</th>
<th>No. of QAM Symbols/frame</th>
<th>Symbol rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSP (20 parameters)</td>
<td>20</td>
<td>50</td>
<td>20 (20x1)</td>
<td>1000 500</td>
</tr>
<tr>
<td>LTP gain and stochastic gain</td>
<td>2</td>
<td>500</td>
<td>2</td>
<td>1 1000 500</td>
</tr>
<tr>
<td>LTP index</td>
<td>2</td>
<td>500</td>
<td>2</td>
<td>1 1000 500</td>
</tr>
<tr>
<td>Stochastic codebook index</td>
<td>2</td>
<td>500</td>
<td>2</td>
<td>1 1000 500</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>4000 500</td>
</tr>
</tbody>
</table>

Figure 2 SEGSNR vs. CSNR for 16-QAM, random index maps (—), optimized index maps (— — —).

Figure 3 SEGSNR vs. CSNR for 256-QAM, random index maps (— — —), optimized index maps (— — —).

included the performance for the average of 10 randomly chosen mappings. From the figures, we observe a gain in terms of CSNR in the order of 1-2 dB for high values of SNRSEG.

Furthermore, for optimized index maps, we observe more graceful degradation for 256-QAM than for 16-QAM. A possible explanation is that for 256-QAM, the dimension of the source space and the channel space is the same for several parameters (LSPs and gain factors) making the index map trivial.

5. Conclusion

A method for optimizing the source- to channel space index map for a channel with a power constraint and multilevel signaling has been presented. In particular, an efficient method for dynamic scaling of the signal constellation during the optimization procedure has been described. The method is applied to a wideband CELP speech coder for 16-QAM and 256-QAM signaling. For both cases, significant gains in terms of CSNR were achieved.

References


931
CEL P coder: Vector Quantization of Pitch Predictor Parameters Determined in Open-Loop Configuration

Ira S. KONVALINKA, Borislav M. LUKOVIĆ and Milan Ž. MARKOVIĆ

Institute of Applied Mathematics and Electronics, Kneza Miloša 37, 11000 Belgrade, Yugoslavia

Abstract. In this work, a partial problem referred to the vector quantization of pitch predictor parameters in a specific realization of CELP 4800 bit/s algorithm based on the pitch predictor parameters determination in an open-loop configuration is considered. The solution for vector quantization of pitch predictor parameters using the well-known LBG algorithm for a codebook synthesis is presented. The influence of such quantization on the quality of synthesized speech is analyzed, and it has been observed that quality is not degraded.

1. Introduction

CELP (Code-Excited Linear Predictive) coder was first introduced by B.S. Atal and M.A. Schroeder [1,2]. Since then, a great number of the researchers have been making efforts to achieve a compromise between an adequate quality of CELP synthesized speech and a possibility of a CELP realization on a single digital signal processor. In this sense, the best achievement is U. S. Federal Standard CELP 4800 bit/s speech coder, FED STD 1016 [4]. The main characteristics of this CELP coder are the application of 10th order autocorrelation method with 15 Hz bandwidth expansion as LPC analysis method, LSP for the formant predictor information transmission, an adaptive codebook with integer and non-integer delays as a pitch predictor, and a special stochastic codebook suitable for fast search procedures. Listening and subjective (DRT, DAM) tests show that FED STD 1016 coder outperforms all standard speech coders operating at 16 kbit/s and below, and is comparable to 32 kbit/s CVSD.

Recently, the authors have developed an original CELP 4800 bit/s coding algorithm, called KLMCELP. The basic characteristics of KLMCELP coding algorithm are:

- Pitch predictor determination in an open-loop configuration
- Vector quantization of pitch predictor parameters
- Modified covariance method for LPC parameter estimation.

The remaining characteristics of KLMCELP coder (a stochastic codebook form, bit rate, bit allocation, LSP quantization of LPC parameters, a frames and subframes form, and a LSP interpolation procedure) are taken from FED STD 1016 CELP 4800 bit/s coding algorithm. Informal listening tests show that the quality of KLMCELP synthesized speech is near to FED STD 1016 quality at the same bit rate, while the KLMCELP numerical complexity is by 20 % lower than FED STD 1016 complexity.

A partial problem referred to the vector quantization of pitch predictor parameters in KLMCELP coding algorithm is considered in the paper. A new solution for vector quantization of pitch predictor parameters using the LBG algorithm [7] for codebook synthesis is presented. The influence of the proposed pitch predictor parameters quantization to the quality of synthesized speech is experimentally evaluated. It has been observed that the quality is not degraded.

The paper is organized in a following way. The determination of the pitch predictor in KLMCELP coder in the open-loop configuration is described in Section 2. The applied vector quantizer of the pitch predictor parameters based on the LBG algorithm is considered in Section 3. Experimental analysis is presented in Section 4. Conclusion and summary are provided in Section 5.

2. Determination of pitch predictor in open-loop configuration

Pitch predictor of the basic CELP coding algorithm [1,2] (and KLMCELP) has a form:
\[ \frac{1}{P(z)} = \frac{1}{1 - \sum_{i=1}^{q} b_i z^{-D+i-1}} \]

where \( D \) is a pitch period estimate expressed in number of samples, \( \{b_i\}, i=1,...,q \); are pitch predictor coefficients, and \( q \) is typically 1, 2, or 3. In this paper, the open-loop determination of the pitch predictor parameters is suggested instead of the closed-loop determination and adaptive codebook modeling of pitch predictor in FED STD 1016 algorithm. This approach is motivated by the possibility of considerable savings in numerical computations with a very small degradation of the synthesized speech quality compared to the FED STD 1016. The open-loop configuration means that the pitch predictor parameters are determined on the basis of a residual speech signal (for each subframe of 7.5 ms) obtained by inverse LPC filtering. Thus, in the open-loop pitch predictor determination, the importance of the LPC predictor is emphasized. Namely, in order to apply the fast suboptimal pitch predictor parameters determination algorithm [5], the off-diagonal elements of a residual covariance matrix should be as small as possible. In the other words, the proposed concept of the pitch predictor parameters determination is based on the assumption that LPC residual signal, produced by the inverse filtering, is "as white as possible". In this sense, the authors are oriented to the determination of the LPC predictor using the modified covariance method due to its well-known good spectral characteristics instead of the autocorrelation method used in FED STD 1016. To justify this orientation, a comparative experimental analysis referred to the application of different standard LPC analysis methods in the KLMCELP coder is presented in [8]. It is concluded that the best results are obtained using the KLMCELP coder with the modified covariance method as the LPC analysis method.

3. Vector quantizer of the pitch predictor parameters

According to the KLMCELP 4800 bit/s bit allocation, six bits are provided for the transmission of the pitch predictor parameters \( \{b_i\}, i=1,...,q \); for a subframe of length \( L=60 \) speech samples. This fact suggests the vector quantization of the KLMCELP pitch predictor parameters.

An \( N \)-level \( K \)-dimensional vector quantizer is the mapping, \( q \), that assigns to each input vector, \( x = (x_0, ..., x_K) \), a reproduction vector, \( x^* = q(x) \), drawn from a finite reproduction alphabet, \( A \), a reproduction alphabet, \( A = \{y_i; i=1,...,N\} \). The quantizer \( q \) is completely described by the reproduction alphabet (or codebook) \( A \) together with the partitioning, \( \{S_i; i=1,...,N\}, \) of the input vector space into the sets \( S_i=q(x); q(x)=y_ij \) of the input vectors mapped into the \( i \)-th reproduction vector (or codeword). The most popular algorithm for designing the vector quantizer \( q \) is LBG algorithm [7]. This algorithm could be described by the following:

(1) **Initialization:** Given \( N \) number of levels, distortion threshold \( \epsilon \equiv \alpha \), an initial \( N \)-level reproduction alphabet \( A_0 \), and a training sequence \( \{x_j; j=0,...,n-1\} \). Set \( m=0 \) and \( D_{m+1} = \infty \).

(2) Given \( A_m = \{y_i; i=1,...,N\} \), find the minimum distortion partition \( P(A_m) = S_i; i=1,...,N \) of the training sequence: \( x_j \in S_i \) if \( d(x_j,y_i) \leq d(x_j,y_j) \) for all \( i \). Compute the average distortion:

\[ D_m = \sum_{j=0}^{n-1} \min_{y_j \in A_m} d(x_j,y_j) \]

(3) If \( (D_{m+1} - D_m) / D_m \leq \epsilon \), halt with \( A_m \) final reproduction alphabet. Otherwise continue.

(4) Find the optimal reproduction alphabet \( x(P(A_m)) = \{x(S_i); i=1,...,N\} \) for \( P(A_m) \). Set \( A_{m+1} = x(P(A_m)) \). Replace \( m \) by \( m+1 \) and go to (2).

A distance measure and the initial reproduction alphabet, \( A_0 \), must be defined in advance. A nature and small number \( q \) of the pitch predictor parameters suggest the use of the Euclidean distance as a distance measure. There are several ways to choose the initial reproduction alphabet required by the algorithm. In this paper, we choose the 'splitting' procedure as the most adequate method [7]. This procedure requires that the number of vector quantizer levels, \( N \), is a multiple of \( 2^k \), and could be described as follows:

(1) **Initialization:** Set \( K=1 \) and define \( A_d(l)=x(A), l \) the centroid of the entire training sequence.

(2) Given the reproduction alphabet \( A_d(K) \) containing \( K \) vectors \( y_l; i=1,...,K \), "split" each vector \( y_l \) into two close vectors \( y_l+e \) and \( y_l-e \), where \( e \) is a fixed perturbation vector. The collection \( A_d \) of \( y_l+e, y_l-e \) \( i=1,...,K \) has \( 2K \) vectors. Replace \( K \) by \( 2K \).

(3) If \( K = N \) then set \( A_0 = A_d(K) \) and halt. \( A_0 \) is the initial reproduction alphabet for the \( N \)-level quantization algorithm. Else, run the algorithm for an \( K \)-level quantizer on \( A_d(K) \) to produce a good reproduction alphabet, \( A_d(K) \), and then return to step (2).

Using the splitting algorithm on a training sequence, one starts with a one-level quantizer consisting of the centroid of the training sequence. The vector is then split into two vectors and the two-level quantizer algorithm is run on this pair to obtain a two-level quantizer. This procedure is to be repeated \( M \) times, e.g. until \( K = N \).

In the next Section, the chosen method for the vector quantization of the KLMCELP pitch predictor parameters by using the LBG algorithm is experimentally evaluated.

4. Experimental analysis

The experimental analysis is performed using the KLMCELP software realization on PC 486 DX. Stochastic codebook is taken from FED STD 1016 and has 512
codewords with 60 samples length. The full search of stochastic codebook is done for each subframe (7.5 ms). A
10th order LPC predictor is determined by using the modified covariance method over the frames with 240
samples length (30 ms). The difference between the original and the synthesized speech is filtered by the
perceptually weighting filter with a weighting factor, \( q = 0.8 \). The root mean square value of this filter output
(PRMS - Perceptive RMS) represents an objective criterion for this experimental analysis.

The test speech base consists of 4 files, each containing 10 isolated digits, spoken by three male speakers (R1, R4,
and R9) and a female speaker (R14). The pitch predictor parameters vector quantization is performed using the LBG
algorithm with Euclidean distance and \( N = 64 \) levels (e.g. 6 bits are provided for index). The codebooks for \( q = 1, 2, \) and \( q = 3 \) are used in the analysis. The codebooks are
obtained by using a training data set of about 8000 \( q \)-dimensional vectors of the pitch predictor parameters. The
training data set is obtained by analyzing of the connected speech with male and female speakers using the pitch
predictor determination algorithm with the stability testing and correcting procedure [5,6]. In the other words, the
training data set consists of either stable or stabilized pitch predictor parameters vectors. The test speech base files,
R1, R4, R9, and R14, are not included in the training data set.

Table 1 presents the PRMS criterion values obtained by using the KLMCELP 4800 bit/s speech coding algorithm
with and without vector quantization of the pitch predictor parameters for \( q = 1, 2, \) and \( 3 \).

<table>
<thead>
<tr>
<th>Test file</th>
<th>R1</th>
<th>R4</th>
<th>R9</th>
<th>R14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>64000</td>
<td>68000</td>
<td>72000</td>
<td>77000</td>
</tr>
<tr>
<td>( q = 1 )</td>
<td>not VQ</td>
<td>57.135</td>
<td>83.625</td>
<td>70.529</td>
</tr>
<tr>
<td></td>
<td>VQ</td>
<td>57.143</td>
<td>84.002</td>
<td>70.422</td>
</tr>
<tr>
<td>( q = 2 )</td>
<td>not VQ</td>
<td>53.782</td>
<td>75.970</td>
<td>67.403</td>
</tr>
<tr>
<td></td>
<td>VQ</td>
<td>54.657</td>
<td>76.158</td>
<td>68.869</td>
</tr>
<tr>
<td>( q = 3 )</td>
<td>not VQ</td>
<td>55.317</td>
<td>72.977</td>
<td>65.682</td>
</tr>
<tr>
<td></td>
<td>VQ</td>
<td>54.038</td>
<td>74.021</td>
<td>64.415</td>
</tr>
</tbody>
</table>

The experimental results, presented in Table 1, justify the choice of the LBG algorithm for vector quantization of
the open-loop determined pitch predictor parameters in the KLMCELP algorithm. Namely, the objectively measured
quality of the synthesized speech after using the KLMCELP 4800 bit/s coder with the vector quantized pitch predictor parameters is not degraded. This fact is especially important in the cases of \( q = 2 \) and \( q = 3 \) where the
PRMS values are substantially smaller than the PRMS values in the case of \( q = 1 \).

5. Conclusion
In the paper, the solution for the vector quantization of the pitch predictor parameters determined in the open-loop
configuration in the original CELP 4800 bit/s speech coding algorithm, called KLMCELP, is considered and
experimentally evaluated. The presented experimental results justify the use of the well-known LBG algorithm
with Euclidean distance for the vector quantizer design. The quality of the synthesized speech obtained by using
the KLMCELP algorithm with the vector quantized pitch predictor parameters is not degraded.

Acknowledgements
This work was supported by Ministry of Sciences and Technology of Republic Serbia, project No. 0403, through
Institute of Mathematics SANU, Belgrade.

References

IEEE Int. Conf. on Communications, pp. 48.1, May 1984.
1985.
XXXVIII Conf. ETAN, Belgrade, September 1993.
Interpolation of Autoregressive Processes at Discontinuities: Application to LPC based Speech Coding

Jan S. ERKELENS and Piet M.T. BROERSEN

Delft University of Technology, Department of Applied Physics
P.O. Box 3046, 2600 GA Delft, The Netherlands
Tel +31 (0)15-781823 / +31 (0)15-786419
Fax +31 (0)15-6081
E-Mail: erkelens@dutnch.tn.tudelft.nl / broersen@dutnch.tn.tudelft.nl

Abstract. Most low bit rate speech coders use autoregressive modeling to describe the speech spectrum. The model is updated in frames of a fixed length. In transition regions large changes in the spectral properties can occur across frame boundaries. Interpolation of the spectral model is used to cope with this problem, improving the quality of the coders. Several methods exist to interpolate the spectral model. Former studies compare these methods by means of experiments on speech data. We have followed a different approach. By means of theoretical arguments we show that some of the representations investigated in literature are not suitable for interpolation of the spectral model. Moreover, the best interpolation procedure is interpolation of the normalized autocorrelation coefficients of a model, weighted with the energy of the analysis frames.

1. Introduction
In speech coding, speech signals are assumed to be stationary over relatively short time intervals of about 20 ms. In such intervals, LPC based speech coders describe the speech spectrum with an autoregressive model. In transition segments large changes in energy and spectral characteristics can occur in short time intervals. To cope with this problem, one could update the LPC model more frequently, for example every 5 ms or so. The increase in bitrate in estimating the parameters more frequently can be avoided by interpolating the spectral models of two consecutive analysis frames. The best interpolation method is the one giving the best approximation to the models that would be obtained from a more frequent update of the LPC parameters.

Interpolation can be performed with different representations, such as the Autocorrelation Function (ACF), Reflection Coefficients (RC), Log Area Ratios (LAR), Arcsine of the Reflection Coefficients (ASRC) and Line Spectrum Pairs (LSP) [1-3]. These representations are one-to-one related with the LPC parameters and ensure stability of the model when interpolated. Former studies [4,5] compared the interpolation behaviour of different representations by means of experimentally results on speech data.

In a previous paper [6] we showed by using autoregressive theory that averaging of the sample autocorrelation functions of two consecutive frames yields the best model for the total of the two frames. This is used in section 4 of this paper to develop an optimal interpolation scheme which is theoretically better than other methods studied in literature.

It is known [7,8] that estimated reflection coefficients can have much larger variances than parameters if the last reflection coefficient of a process is close to plus or minus unity. This fact is used in section 3 to show that RC, LAR and ASRC are not suitable for interpolation of the spectral model.

2. Autoregressive Theory
Consider a stationary autoregressive process, described by:

\[ x(n) = - \sum_{i=1}^{K} \alpha_i x(n-i) + \varepsilon(n) \]  

(1)

The coefficients \( \alpha_i \) are the autoregressive parameters, \( K \) is the order of the process and \( \varepsilon(n) \) is an independent identically distributed zero mean Gaussian process with variance \( \sigma^2 \). LPC based speech coders use an autoregressive model to describe the short term speech spectrum. The parameters are estimated by minimizing the residual energy in some way. For the autocorrelation method, the covariance method and the modified covariance method the parameter estimation procedure comes down to minimizing the following expression:

\[ a^T R a = s^2 \]  

(2)

where \( a \) is the parameter vector to be estimated, \( R \) is the sample autocovariance matrix of the data record from which the parameters are to be estimated and \( s^2 \) is an estimate of the variance of the innovation process. Estimation methods differ in the way \( R \) is computed. For the autocorrelation method for example, \( R \) is a positive definite, symmetric Toeplitz matrix, which ensures a stable model. The elements of \( R \) are the elements of an autocorrelation function \( \hat{R}(k) \) of the signal \( x(n) \):

\[ \hat{R}(k) = \sum_n x(n)x(n+k) \]  

(3)

The summation ranges over the frame length.

The expectation of \( s^2 \), as defined in eq (2) is slightly less than \( \sigma^2 \) and equals \((1-p)/N\sigma^2\) in asymptotic theory, provided the order of the model \( p \) is greater than or equal to the true process order \( K \). \( N \) is the number of observations of \( x(n) \) used to estimate the
parameters. If we substitute for $R$ in eq.(2) the theoretical covariance matrix of the process, the expectation of $s^2$ is somewhat greater than $\sigma^2$ and equals $(1+p/N)\sigma^2$. For short data records, results depend on the estimation method [9]. More generally, the expression $a^2 R_a$, with an arbitrary parameter vector and $R$ the theoretical covariance matrix of any autoregressive process is a measure of how well the model fits to the process. It is called Prediction Error (PE) and it also has a frequency domain interpretation, given by:

$$\frac{\text{PE}}{\sigma^2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{A(e^{i\omega})}{A(e^{i\omega})} - \hat{A}(e^{i\omega}) \right|^2 d\omega,$$

(4)

where $A(z)$ is the autoregressive polynomial of the process and $\hat{A}(z)$ is the autoregressive polynomial of the model.

From eq.(3) it is readily seen that the sample correlation function of two consecutive frames of a signal is almost equal to the average of the sample correlation functions of the two frames. Autoregressive estimation methods use their own estimates of the autocorrelation function to estimate the parameters. The model obtained when averaging the autocorrelation functions is close to the model obtained when estimating parameters from the total of the two frames. The model obtained when averaging the autocorrelation functions gives a global minimization of the residual energy over the two frames together.

Averaging or interpolation reduces the variance of the parameters for stationary processes because the number of observations used to estimate the parameters is increased. In several simulation experiments with different processes all representations gave almost the same result for the variance of the residuals, i.e. $(1+p/2N)\sigma^2$. This agrees with the above reasoning. For some processes however interpolation of reflection coefficients, LAR and ASRC gives inferior results due to the statistical properties of these coefficients. This is the subject of the next section.

3. Reflection Coefficients

Parameters can be transformed to reflection coefficients. The reflection coefficients describe the same model, however, the statistical properties of reflection coefficients differ from those of LPC parameters. The theoretical covariance matrix $C_a$ of estimated parameters of a third order process can be written as:

$$C_a = \frac{1}{N} \begin{bmatrix}
1 & a_1 & a_2 & a_3 \\
a_1 & 1 & a_2 & a_3 \\
a_2 & a_1 & 1 & a_3 \\
a_3 & a_2 & a_1 & 1
\end{bmatrix}.$$

(5)

The theoretical covariance matrix of estimated reflection coefficients $C_k$ can be computed with a recursive procedure [7]. It is shown below for a third order process:

$$C_k = \frac{1}{N} \begin{bmatrix}
(1-k_1^2)(1-k_2^2)(1-k_3^2) & -2k_1(1-k_1^2)(1-k_2^2)(1-k_3^2) & 0 \\
(1-k_1^2)(1-k_2^2) & 1 & -2k_2(1-k_1^2)(1-k_2^2)(1-k_3^2) & 0 \\
-2k_1(1-k_1^2)(1-k_2^2)(1-k_3^2) & (1-k_1^2)(1-k_2^2) & 1 & -2k_3(1-k_1^2)(1-k_2^2)(1-k_3^2) & 0 \\
0 & 0 & (1-k_1^2)(1-k_2^2)
\end{bmatrix}.$$

(6)

$N$ is the number of observations used to estimate the parameters. The third estimated reflection coefficient is uncorrelated with the first and second. This is a general property of estimated reflection coefficients: estimated reflection coefficients $k_i$ of a model, like any other reflection coefficients of the model. Moreover it can be seen from eq.(6) that the variances and covariances can become very large if the third reflection coefficient is close to plus or minus unity. More generally, if the last reflection coefficient of a process is close to plus or minus unity, the lower estimated reflection coefficients can have very large variances and covariances. This is not true for the LPC parameters. The largest element in eq.(5) is theoretically limited for stationary processes to $4/N$, in eq.(6) it can become unbounded. Due to these large variances and covariances of estimated reflection coefficients they can differ much from their theoretical values, but still yield good models in terms of Prediction Error if the quotient $p/N$ is small.

Consider the third order process with reflection coefficients $[0.0.9.0.9]$. The variances of the first and second estimated reflection coefficients are 12.02 and the covariance between these is 11.98; they are very strongly correlated. The variance of the third estimated reflection coefficient is 0.15. The estimated reflection coefficients fill up a long, stretched banana-shaped volume in RC space. The models in this volume have a small Prediction Error. The prediction error $a^2 R_a$ of a model $[k_1 k_2 k_3]$ with respect to this process can be written out in terms of $k_1$, $k_2$ and $k_3$. Figure 1 gives the value of $k_2$ which minimizes the Prediction Error as a function of $k_1$; $k_3$ is kept fixed to its theoretical value -0.92. The maximum Prediction Error on this curve is 1.02$\sigma^2$. The crosses (*) lying on the curve in Figure 1 show two models which could be estimated from realizations of this process. The models have reflection coefficients $[0.36 -0.36 -0.92]$ and $[0.31 0.45 -0.92]$. The Prediction Errors associated with these models are 1.01$\sigma^2$ and 1.02$\sigma^2$ respectively. Averaging of these models leads to the model with reflection coefficients $[0.03 -0.09 -0.92]$. This model is denoted by 'b'. It does not lie on the "minimum Prediction Error curve". The prediction Error associated with this averaged model is 1.17$\sigma^2$ and is larger than the Prediction Errors of both models it was obtained from.

As an other example, consider the process with reflection
coefficient vector \([-0.95, 0.5, 0.9, 0.1]^{T}\). \(C_k\) for this process is given by:

\[
C_k = \frac{1}{N} \begin{bmatrix}
+0.06 & -1.34 & -0.46 & 0 \\
-1.34 & +30.00 & +10.30 & 0 \\
-0.46 & +10.30 & +3.59 & 0 \\
0 & 0 & 0 & +0.19
\end{bmatrix}.
\]

As can be seen from this matrix, the first three estimated reflection coefficients are all strongly correlated. The second and third reflection coefficients have large variances. However, averaging of reflection coefficients can lead to models which violate the statistical properties of estimated reflection coefficients, in other words, averaging can lead to models which are very unlikely to be estimated and yield a large Prediction Error. In simulations, models have been estimated from 40,000 frames of length \(N=200\) of this process and averaged. In 43\% of the cases the Prediction Error associated with the averaged model was larger than the Prediction Errors associated with the models it was obtained from. To illustrate this point further, a typical example is given. In one frame, the estimated reflection coefficient vector was \([-0.96, 0.78, 0.58, 0.91]^{T}\). The Prediction Error associated with this vector is \(1.00\sigma^2\). The next frame, the estimated reflection coefficient was \([-0.89, -0.06, 0.17, 0.89]^{T}\), with a Prediction Error of \(1.06\sigma^2\). The average of these vectors is \([-0.93, 0.36, 0.37, 0.90]^{T}\), with a Prediction Error of \(1.38\sigma^2\). For the process considered here, the standard deviation of the Prediction Error is experimentally found to be \(0.50\sigma^2\). The Prediction Error associated with the averaged model differs about 7 standard deviations from the expected value for 200 observations.

In conclusion, estimated reflection coefficients may have large variances and covariances if the last reflection coefficient is close to plus or minus unity. Estimated models from such processes lie in long curved areas in RC space, having small Prediction Errors. Averaging of reflection coefficients may result in models laying outside this area. These models are not in agreement with the statistical properties of the estimated reflection coefficients and have a large Prediction Error. The LAR and ASRC coefficients are transformations of the corresponding reflection coefficients and the same arguments are valid. So reflection coefficients, LAR and ASRC are not suitable for interpolation.

4. Interpolation

Autoregressive estimation methods use the autocorrelation function of the data to estimate the parameters. In this section an optimal interpolation method for the spectral model is developed. Consider Figure 2a where the spectral model is estimated in non-overlapping analysis frames of a fixed length. Suppose we want to have a model of the signal at twice the rate. This can be accomplished by letting analysis frames have 50\% overlap. Half of the analysis frames will be the original frames, the other half will be centered around the boundary of the original frames as in Figure 2b. The bitrate for encoding the LPC parameters will double. The purpose of interpolation of the spectral model is to avoid the increase in bitrate resulting from the increased updating of the LPC parameters.

For the moment, assume that the speech signal is stationary in an analysis frame. Any change in characteristics is assumed to occur only at a frame boundary in Figure 4a. From section 2 it is known that averaging of the sample autocorrelation functions of two consecutive frames will give a model close to the model we would get if we estimated parameters from the total of the two frames. The averaged sample autocorrelation function of frames 1 and 3 in Figure 2b will be the best estimate of the sample autocorrelation function of frame 2. If frame 2 in Figure 2b would be shifted to contain 75\% of frame 1 and 25\% of frame 3, the best approximation of its autocorrelation function would be 3/4 times the autocorrelation function of frame 1 and 1/4 times the autocorrelation function of frame 3.

In speech coding we are restricted to interpolation methods which use information available in the decoder. The quantized LPC parameters are available at the decoder but they themselves cannot be used for interpolation because stability is not ensured. They have to be transformed to other representations. It seems that of these representations ACF is optimal for approximating the sample autocovariance function. This, however, is not true. For, in the computation of the ACF from the LPC parameters, the residual variance is assumed to be the same for both frames. This assumption is not true always. For example, in unvoiced/voiced and voiced/unvoiced transitions there is a large difference in the residual energy due to the pitch pulses in the residual of the voiced segment. The optimal interpolation can be performed by weighting the normalized autocorrelation function with the frame energy (EWNACF) and interpolating this function. In speech coders the frame energy is usually transmitted to the decoder side and therefore this interpolation can be performed without an increase in bitrate.

Weighting the normalized ACF with the frame energy takes both the power gain of the autoregressive transfer function and the energy of the excitation into account, in contrast with other representations.

4.1 Discussion

The above statements have been made under the assumption that any change in signal characteristics occurs at a frame boundary. What if a change in energy of the signal does occur somewhere else in frame 1 or frame 3 in Figure 2b? Without loss of generality we consider a low to high energy transition. If the transition occurs somewhere in the first half of frame 1, the best approximation of the autocorrelation function of frame 2 would be the autocorrelation function of frame 3. However, averaging the energy weighted normalized autocorrelation functions (EWNACF) of frame 1 and 3 is still better than interpolation of any of the other existing methods. Interpolation of EWNACF is better also if the change in energy occurs somewhere in the second half of frame 1. If the change in energy occurs in the first half of frame 3, it depends on the difference in energies and spectral characteristics whether EWNACF or ACF interpolation will give best approximation.

Finally, if the transition is in the second half of frame 3
interpolation of EWNACF will lose from ACF. However, EWNACF will only lose from ACF when approximating the autocorrelation function of a low energy frame. Generally, high energy segments are more important to model accurately than low energy segments, and interpolation of EWNACF is better in high energy segments.

4.2 Example:
Figure 3 presents a speech segment of length 320 samples (40 ms). This segment contains an unvoiced/voiced transition. The segment is subdivided into two frames of 160 samples each. Parameters are estimated from both segments with the Burg method. The model order was 10. The different representations are averaged, and also EWNACF. The resulting models are compared with the model obtained from the segment of length 160 centered at the frame boundary. The Prediction Error is used to compare the models. Results are given in Table 1 below.

Table 1. Comparison of interpolation methods for the signal of Figure 3.

<table>
<thead>
<tr>
<th>method</th>
<th>PE</th>
<th>method</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NACF</td>
<td>2.04</td>
<td>RC</td>
<td>2.79</td>
</tr>
<tr>
<td>ACF</td>
<td>1.39</td>
<td>LAR</td>
<td>1.89</td>
</tr>
<tr>
<td>EWNACF</td>
<td>1.04</td>
<td>ASRC</td>
<td>2.31</td>
</tr>
<tr>
<td>LSP</td>
<td>1.90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It is clear that EWNACF performs best. Interpolation of ACF is second best. EWNACF outperforms all other representations because it takes into account the difference in excitation energy of both frames, whereas other representations don’t.

5. Summary and Conclusions
We have analyzed the problem of spectral interpolation, by means of autoregressive theory. Simulation experiments on stationary data show only minor differences in the interpolation performance of different LPC representations. For stationary processes, an important effect of interpolation is reducing the variance of the parameters, since the effective number of observations used to estimate the parameters is increased.

For some processes, interpolation of RC, LAR and ASRC parameters gives inferior results, because the statistical properties of estimated reflection coefficients may be violated, leading to a model with a high Prediction Error. RC, LAR and ASRC coefficients are therefore not suitable for interpolation.

The best approximation to the correlation function of a segment somewhere between two analysis frames, is linear interpolation of the sample correlation functions of these analysis frames. This can be implemented in speech coders by interpolating EWNACF, which takes into account both power gain of the autoregressive transfer functions and excitation energies.

Acknowledgement
This research is supported by the Technology Foundation (STW) under project no. DTN11.2436.

References

Figure 3. Speech segment containing an unvoiced/voiced transition.
Efficient Quantization of LPC Parameters Using a Mixed LSP/PARCOR Representation

Cheung-Fat CHAN

Department of Electronic Engineering, City Polytechnic of Hong Kong
83, Tat Chee Avenue, Kowloon, HONG KONG
Fax: (852) 7887791, E-mail: EECFCHAN@CPhKVA.BITNET

Abstract. This paper proposes a mixed LSP/PARCOR representation scheme for efficient quantization of LPC parameters. In the proposed scheme, the PARCOR coefficients of an LPC system are split into two sets. The first set is transformed to LSP domain for quantization. The second set is quantized using a residual minimization scheme in the PARCOR domain. It is well known that the first few PARCOR coefficients of an LPC system are more dominant in terms of overall quantization performance than the last few PARCOR coefficients. The main advantage of the mixed LSP/PARCOR scheme is that efficient LSP quantization techniques can be applied on the first set of PARCOR coefficients while the second set can be quantized more coarsely, thereby facilitating an optimum allocation of available resource (bits) for the quantization process. Simulation results show that, in average, 2-bit gain in quantization efficiency can be achieved by using a mixed representation.

1. Introduction
Partial Correlation (PARCOR) coefficients are widely used parameters for linear predictive coding (LPC) of speech. The advantages of using PARCOR coefficients are that they are numerically bounded to unity allowing them to be computed entirely on fixed-point arithmetic[1]. In recent years, line spectrum pair (LSP) representation has become even more popular in speech coding applications superseding PARCOR representation. It has been shown that LSP parameters are distributed locally in frequency domain, and, hence they can be quantized efficiently using split vector quantization (SVQ) technique[2]. Recently, a split residual VQ (SRVQ) scheme based on the partitioning of the PARCOR coefficients was proposed[3]. The SRVQ scheme combines the LSP analysis and quantization in a single step. The VQ encoding is carried out in a sequential manner where both the forward and backward prediction residuals are coupled across VQ stages. The SRVQ scheme was shown to achieve better performance in terms of spectral distortion than SVQ scheme using LSP representation, however, this is achieved if the number of VQ stages is less than three[3].

This paper proposes a mixed LSP/PARCOR representation for efficient quantization of LPC parameters. The advantages of this mixed representation scheme are that both the useful properties of LSP and PARCOR parameters can be exploited to develop a yet more efficient quantization scheme. This paper will first point out some interesting relationships between PARCOR coefficients and line spectrum frequencies which may be exploited to reduce the quantization distortion. This paper will also prove that the inmutance spectral pairs[4] of an order-m system are actually the line spectral pairs of an order-(m-1) system. Based on these properties, several mixed LSP/PARCOR quantization schemes are proposed. Simulation results will then be given to show that the proposed schemes achieve lower quantization distortion than conventional schemes using pure LSP or PARCOR representation.

2. An Efficient Algorithm for Converting PARCOR to LSP
It is well known that the polynomial $A_m(z)$ associated with an order-m LPC system satisfies the recurrence relationship[5];

$$A_m(z) - A_{m-1}(z) + k_m z^{-m} A_{m-1}(z^{-1})$$

where $k_m$ is PARCOR coefficient. By setting two extreme conditions of $k_{m+1} = \pm 1$, a symmetric polynomial and an anti-symmetric polynomial are derived;

$$P_m(z) = A_m(z) - z^{-m} A_m(z^{-1}) = 1 + p_1 z^{-1} + p_2 z^{-2} + \ldots + p_m z^{-m} + z^{-m+1}$$

(2)

$$Q_m(z) = A_m(z) + z^{-m} A_m(z^{-1}) = 1 + q_1 z^{-1} + q_2 z^{-2} + \ldots + q_m z^{-m} - z^{-m+1}$$

(3)

It had been shown that, if the system is stable, the roots of $P_m(z)$ and $Q_m(z)$ must alternate each other on the unit circle[6]. Hence, we have

$$P_m(z) = \prod_{i=1, odd}^{m} (1 + c_i z^{-1} + z^{-2}) \quad m \text{ even}$$

(4)

$$Q_m(z) = \prod_{i=1, even}^{m} (1 + c_i z^{-1} + z^{-2}) \quad m \text{ even}$$

(5)

where $c_i = -2 \cos \Omega_i$ with $\Omega_i$ being the line spectral frequency of an order-m system. If we define $C_m(z) = z^{m} A_m(z)$, then from (1), an alternate set of recursive formula is derived

$$C_m(z) = z^{\frac{1}{2}} [C_{m-1}(z) + k_m C_{m-1}(z^{-1})]$$

(6)

By evaluating $C_m(z)$ on the unit circle and organizing the real and imaginary parts of $C_m(z)$ into a two-dimensional vector as $C_m(z) = [R[C_m(z)] \quad I[C_m(z)]]^T$ and noting that $C_m(z)$ is a real-coefficient polynomial, we immediately derive...
\[ C_m(\theta) = \varphi \left( \frac{\theta}{2} \right) [I + kG] C_{m-1}(\theta) \]  
(7)

where \( \varphi(\theta) = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \), \( I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \), and \( G = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \).

By using similar definition, and noting the fact that
\[ P_{m+1}(\theta) = \varphi \left( \frac{m+2\theta}{2} \right) C_{m+1}(\theta) \]
and
\[ Q_{m+1}(\theta) = \varphi \left( \frac{m+2\theta}{2} \right) \]
we obtain
\[ P_{m+1}(\theta) = \varphi \left( \frac{m+2\theta}{2} \right) \]
(8)

Similarly, we also have
\[ Q_{m+1}(\theta) = \varphi \left( \frac{m+2\theta}{2} \right) \]
(9)

From (8) and (9), we have
\[ P_{m+1}(\theta) = 2\Re \left( C_m(\theta^\pi) \right) \]
and
\[ Q_{m+1}(\theta) = 2\Im \left( C_m(\theta^\pi) \right) \].
Therefore, the odd line spectral frequencies are the frequencies at which \( \Re \left( C_m(\theta^\pi) \right) = 0 \) and the even line spectral frequencies are the frequencies at which \( \Im \left( C_m(\theta^\pi) \right) = 0 \). In other words, the even and odd line spectrum frequencies are the zero-cross points on the real and imaginary axis of the complex plane defined by \( C_m(\theta^\pi) \), respectively. Since the line spectral frequencies satisfy the ordering properties, i.e.,
\[ \theta_1^\pi < \theta_2^\pi < \cdots < \theta_m^\pi, \]
the zero-cross points on the real and imaginary axis of \( C_m(\theta^\pi) \) are in alternate order. The ordering of zero-cross points leads to a very simple algorithm to locate the line spectral frequencies[7].

3. Some Useful Relationships Between PARCOR and LSP Parameters

By evaluating \( P_{m+1}(\theta) \) and \( Q_{m+1}(\theta) \) on the unit circle and rearranging the result using (8) and (9), we obtain
\[ C_m(\theta) = \begin{cases} \frac{1}{2} \sum_{i=2, odd}^{m} \left( \cos \theta - \cos \theta_i^\pi \right) & m \text{ even} \\ \frac{m+1}{2} \sum_{i=2, even}^{m} \left( \cos \theta - \cos \theta_i^\pi \right) & m \text{ odd} \end{cases} \]
(10)

By multiplying the product terms in (10), and comparing terms in (2) and (3), and using the fact that \( k_m = \frac{1}{2} (p_m + q_m) \), we have
\[ k_m = \begin{cases} \frac{1}{2} \sum_{i=2, even}^{m} \cos \theta_i^\pi - \sum_{i=1, odd}^{m} \cos \theta_i^m & m \text{ even} \\ \sum_{i=2, even}^{m} \cos \theta_i^\pi - \sum_{i=1, odd}^{m} \cos \theta_i^m & m \text{ odd} \end{cases} \]
(11)

Also, we can derive
\[ \sum_{i=1}^{m} \cos \theta_i^\pi = \sum_{i=1}^{m} \cos \theta_i^m, \quad k_m = 1 \]
(12)
Similarly, if we evaluate the terms \( p_2 \) and \( q_2 \), and calculate \( k_{m-1} \), using Levinson step-down procedure, we achieve
\[ k_{m-1} = \frac{1}{1 - k_m^2} \begin{cases} \sum_{i=2, even}^{m} \cos \theta_i^\pi - \sum_{i=1, odd}^{m} \cos \theta_i^m & m \text{ even} \\ \sum_{i=2, even}^{m} \cos \theta_i^\pi - \sum_{i=1, odd}^{m} \cos \theta_i^m & m \text{ odd} \end{cases} \]
(13)

The result in (11) indicates that when all \( \theta_i^\pi \) are closely placed in pairs, \( k_m \) is approaching 1 or -1. On the other hand, when all \( \theta_i^\pi \) are evenly distributed in frequency domain, \( k_m \) is approaching zero. It is well known that the LSPs are quite often distributed evenly and hence the last PARCOR coefficient usually has a small absolute value.

3.1. Sequence-Reversed System

If PARCOR coefficients of an order-m system are modified as \( \tilde{k}_i - k_{m-i+1}, i \leq m \), then the odd line spectrum frequencies of the modified system remain unchanged.

Proof: The spectrum of a sequence-reversed system is obtained as
\[ C_{m}(\theta) = \varphi \left( \frac{\theta}{2} \right) [I + kG] \cdots [I + kG] \varphi \left( \frac{\theta}{2} \right) \]
Rewrite this equation in reversed order to achieve
\[ \begin{pmatrix} \prod_{i=1}^{m} (1 - k_i^2) & J \varphi \left( \frac{\theta}{2} \right) [I + kG] \cdots [I + kG] \varphi \left( \frac{\theta}{2} \right) J C_m(\theta) \\ 0 \end{pmatrix} = 0 \]
where \( J \) is an exchange matrix. For \( \theta = \theta_i^\pi \), \( i \text{ odd} \), i.e., the odd line spectrum frequencies of the sequence-reversed system, we have
\[ \begin{pmatrix} \prod_{i=1}^{m} (1 - k_i^2) & \Re \left( C_m(e^{2\pi \theta_j}) \right) J C_m(\theta_j) \\ 0 \end{pmatrix} = 0 \]
Therefore, \( \Re \left( C_m(e^{2\pi \theta_j}) \right) = 0 \); to satisfy this we must have \( \theta_j^\pi = \theta_i^\pi \), \( i \text{ odd} \).

3.2. Sign-Inverted System

The even line spectrum frequencies of an order-m system whose PARCOR coefficients are sign-inverted, i.e., \( \tilde{k}_i = -k_i, i \leq m \), and the even line spectrum frequencies of the order-reversed system are the same.

Proof: The spectrum of a sign-inverted system is generated as
\[ \tilde{C}_{m}(\theta) = \varphi \left( \frac{\theta}{2} \right) [I - kG] \cdots [I - kG] \varphi \left( \frac{\theta}{2} \right) \]
Similarly, rewrite the equation in reversed order and substitute \( \theta = \theta_i^\pi \), \( i \text{ even} \), we have
\[ \begin{pmatrix} \prod_{i=1}^{m} (1 - k_i^2) & \Re \left( \tilde{C}_m(e^{2\pi \theta_j}) \right) \tilde{C}_m(\theta_j) \\ 0 \end{pmatrix} = 0 \]
Therefore, \( \Re \left( \tilde{C}_m(e^{2\pi \theta_j}) \right) = 0 \); this is again satisfied if \( \theta_j^\pi = \theta_i^\pi \), \( i \text{ even} \). Interestingly, properties 3.1 and 3.2 may allow
us to modify PARCOR coefficients in such a way that the LSP parameters of the modified system can be quantized more efficiently. In fact extension of 3.1 and 3.2 leads to a duality relationship shown in Fig. 1.

![Diagram](image_url)

Fig. 1 Duality of LSP and PARCOR

3.3. Immittance Spectrum Pair
The first m-1 immittance spectrum frequencies of an order-m system are actually the line spectrum frequencies of an order-(m-1) system.

Proof: From the definition of immittance spectrum pair (ISP) in [4], the ISPs are the roots of polynomials

\[ X_m(z) = A_m(z) + z^{-m}A_m(z^{-1}) \]

and

\[ Y_m(z) = A_m(z) - z^{-m}A_m(z^{-1}) \]

Using the definition in previous sections to obtain

\[ X_m(\theta) = \varphi^T(\theta)C_m(\theta) + \varphi(\theta)C_m(-\theta) \]

\[ - \varphi^T(\theta)2(1+k_e)S_{m-1}(e^{\theta}) \]

\[ \text{and} \]

\[ Y_m(\theta) = \varphi^T(\theta)C_m(\theta) - \varphi(\theta)C_m(-\theta) \]

\[ - \varphi^T(\theta)2(1-k_e)S_{m-1}(e^{\theta}) \]

Therefore, we can see that the first m-1 ISPs of an order-m system are actually the LSPs of an order-(m-1) system (note that the last ISP actually equal to k_m).

4. Efficient Quantization Schemes Using Mixed LSP/PARCOR Representations
It has been shown in [4] that utilization of ISP representation for quantization can save at least 1 bit when compared to the LSP counterpart. However, we have shown in previous section that ISP is actually the LSP of a lower order system. Obviously, the attribute to the gain in quantization efficiency is mainly due to the last ISP coefficient, i.e., the last PARCOR coefficient. It is well known that the last PARCOR coefficient contributes relatively less in the quantization error than the first PARCOR coefficient. Normally, it can be quantized to 2 bits as compared to 4 bits for a LSP parameter. Therefore, we can expect that a mixed LSP/PARCOR representation for quantization would be more efficient. A block diagram of the proposed mixed LSP/PARCOR quantization scheme is depicted in Fig. 2. In the proposed scheme, m PARCOR coefficients, i.e., k_1 to k_m, of an order-m system are firstly obtained by LPC analysis. The first n PARCOR coefficients, where n < m, are transformed to LSP domain for quantization (in this case, the set of n LSPs represents an order-n system). The last m-n PARCOR coefficients are quantized as they would be in the PARCOR domain. In general, scalar and vector quantization may be used in both domains. However, the quantization must be carried out in a sequential manner. In other words, quantization of the LSP parameters is firstly carried out and the quantized LSP parameters are then transformed back to PARCOR domain. Lattice transformation of the autocorrelation of the input speech using the quantized parameters is then performed to generate a set of auto-and cross-correlation of forward and backward residuals[3] which are then utilized for the quantization of the last m-n PARCOR coefficients.

![Diagram](image_url)

Fig. 2 Block Diagram of the Mixed LSP/PARCOR Quantization Scheme

In this work, both scalar and vector quantization schemes are used for testing the mixed representation scheme. Three schemes had been tested using scalar quantization. The combinations of LSP and PARCOR parameters are: \((\theta_1 - \theta_6, k_7 - k_{10})\), \((\theta_1 - \theta_7, k_8 - k_{10})\) and \((\theta_1 - \theta_8, k_9 - k_{10})\) and they are called MixSQ(6,4), MixSQ(7,3) and MixSQ(8,3), respectively. All LSPs are calculated from PARCOR coefficients using the algorithm described in the previous section. Two additional scalar quantization schemes using pure LSP and pure PARCOR representations are also tested for comparison purposes. The bit allocations for all these schemes were empirically determined so that the resulting quantizers achieve the lowest spectral distortion. In case of vector quantization, we utilized a three-stage split residual vector quantization technique. In the first two stages, 7 LSPs of an order-7 system are split into two sets of dimensions 4 and 3 for vector quantization. In the third stage, the last three PARCOR coefficients of an order-10 system are grouped for vector quantization. This scheme is denoted as MixSRVQ(7,3). Note that the quantization must be carried out in a sequential manner, that means, after finishing the VQ encoding at one stage, the quantized coefficients are utilized to generate the forward and backward residuals for the next stage. In case of the mixed quantization scheme, the LSP vectors in the first two stages are quantized independently, but, after the quantization, the quantized LSPs are transformed back to the PARCOR domain for generating the residuals for the third stage VQ which uses PARCOR coefficients as vector for quantization. For comparison purposes, two split VQ schemes using pure LSP and log area ratio (LAR) features were also tested. The SRVQ scheme proposed in [3] was also tested. The VQ dimensions and bit allocations for codebook indexes were also empirically determined to give the optimum performance.
5. Simulation Results
A speech database contributed by 4 female and 4 males speakers was used in these tests. Speech signals were bandlimited to 3.4 kHz and sampled at 8 kHz using a 16-bit A/D converter. The autocorrelation method for an order-10 LPC analysis using a Hamming window with a size of 32 ms was employed. The LPC frame update rate was 50 frames per second. The quantization tables for the LSP parameters were trained using the generalized Lloyd algorithm with weighted mean squared error as distortion measure [2]. The quantization tables for the PARCOR coefficients were trained using the method described in [3]. There were totally 25,000 LPC frames for training. A separate set of 2,000 LPC frames was allocated for evaluation. Tables I shows the results for the average spectral distortion using scalar quantization. From these results, all mixed quantization schemes perform better than other schemes using pure LSP and PARCOR representations. The MixSQ(7,3) scheme achieves, in average, the lowest spectral distortion in which transparent quantization is obtained at 32 bits; 2 bits less than pure LSP scheme. Fig. 3 illustrates the plots of average spectral distortion against bit rate for the scalar quantization schemes.

<table>
<thead>
<tr>
<th>Bits</th>
<th>PureSQ-K</th>
<th>PureSQ-LSP</th>
<th>MixSQ(6,4)</th>
<th>MixSQ(7,3)</th>
<th>MixSQ(8,3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>2.015</td>
<td>1.812</td>
<td>1.745</td>
<td>1.720</td>
<td>1.813</td>
</tr>
<tr>
<td>23</td>
<td>1.812</td>
<td>1.561</td>
<td>1.319</td>
<td>1.403</td>
<td>1.491</td>
</tr>
<tr>
<td>30</td>
<td>1.523</td>
<td>1.292</td>
<td>1.322</td>
<td>1.211</td>
<td>1.212</td>
</tr>
<tr>
<td>32</td>
<td>1.295</td>
<td>1.083</td>
<td>1.130</td>
<td>1.026</td>
<td>1.094</td>
</tr>
<tr>
<td>34</td>
<td>1.109</td>
<td>0.991</td>
<td>0.961</td>
<td>0.933</td>
<td>0.976</td>
</tr>
<tr>
<td>36</td>
<td>0.911</td>
<td>0.922</td>
<td>0.906</td>
<td>0.873</td>
<td>0.865</td>
</tr>
<tr>
<td>33</td>
<td>0.793</td>
<td>0.831</td>
<td>0.831</td>
<td>0.815</td>
<td>0.812</td>
</tr>
<tr>
<td>40</td>
<td>0.722</td>
<td>0.694</td>
<td>0.767</td>
<td>0.718</td>
<td>0.740</td>
</tr>
</tbody>
</table>

Table I Spectral Distortion (Scalar Quantization)

![Fig. 3 Spectral Distortion vs Bit Rate (SQ Schemes)](image)

2.00
1.80
1.60
1.40
1.20
1.00
0.80
0.60

Bits per frame

Table II Spectral Distortion (Vector Quantization)

<table>
<thead>
<tr>
<th>3-stage VQ</th>
<th>SVQ-LAR</th>
<th>SRVQ-K</th>
<th>SVQ-LSP</th>
<th>MixSRVQ(7,3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4,3,3)</td>
<td>(24dBi)</td>
<td>(4dBi)</td>
<td>(24dBi)</td>
<td>(24dBi)</td>
</tr>
<tr>
<td>(Av.SD)</td>
<td>(Av.SD)</td>
<td>(Av.SD)</td>
<td>(Av.SD)</td>
<td>(Av.SD)</td>
</tr>
<tr>
<td>20</td>
<td>1.913</td>
<td>1.565</td>
<td>1.681</td>
<td>1.583</td>
</tr>
<tr>
<td>39.5</td>
<td>1.2</td>
<td>0.32</td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td>22</td>
<td>1.624</td>
<td>1.323</td>
<td>1.392</td>
<td>1.504</td>
</tr>
<tr>
<td>19.8</td>
<td>1.2</td>
<td>0.19</td>
<td>0.19</td>
<td>0.13</td>
</tr>
<tr>
<td>24</td>
<td>1.446</td>
<td>1.172</td>
<td>1.211</td>
<td>1.171</td>
</tr>
<tr>
<td>16.5</td>
<td>0.38</td>
<td>0.01</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>26</td>
<td>1.281</td>
<td>1.064</td>
<td>1.095</td>
<td>1.022</td>
</tr>
<tr>
<td>9.65</td>
<td>0.16</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>28</td>
<td>1.126</td>
<td>0.992</td>
<td>1.004</td>
<td>0.962</td>
</tr>
<tr>
<td>4.57</td>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>30</td>
<td>0.987</td>
<td>0.945</td>
<td>0.933</td>
<td>0.917</td>
</tr>
<tr>
<td>2.42</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>11 11 11</td>
<td>1.1</td>
<td>1.28</td>
<td>1.03</td>
<td></td>
</tr>
<tr>
<td>1.11 11 11</td>
<td>1.1</td>
<td>1.28</td>
<td>1.03</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4 Spectral Distortion vs Bit Rate (VQ Schemes)

6. Conclusion
A mixed LSP/PARCOR representation for efficient quantization of LPC parameters was proposed. The performance of the mixed quantization scheme was shown to be better than other LPC quantization schemes using pure LSP or pure PARCOR

References
Enhanced Multiband Excitation Coding of Speech at 2.4 kb/s with Phonetic Classification and Variable Dimension VQ

Amitava DAS and Allen GERSHO

Center for Information Processing Research, Dept. of Electrical & Computer Engineering, University of California, Santa Barbara, CA 93106. Tel/Fax: 805 893-2037 / 893-3262, E-Mail: das@kane.ece.ucsb.edu, gersho@ece.ucsb.edu

Abstract. We present a new 2.4 kb/s speech coding algorithm, called enhanced multiband excitation (EMBE) coding. Subjective tests indicate that its quality is as good as two higher rate standard coders, CELP at 4.8 kb/s and IMBE at 4.15 kb/s. The algorithm enhances the basic multiband excitation vocoder scheme by adding several important new features including phonetic classification and a novel spectral quantization technique called variable dimension vector quantization (VDVQ). The phonetic classifier allows the bit allocation and coding strategy for each frame to match the current acoustic-phonetic character of the speech, resulting in improved quality and robustness. The VDVQ scheme represents the log-spectrum with relatively few bits while preserving perceptually important features.

1. Introduction

Speech compression with high quality at 2.4 kb/s is a very challenging task. The quest for an effective solution forces us to reach beyond the capabilities of the current generation of waveform coders as well as the prevailing parametric vocoders. A solution based on waveform coding methods, e.g., CELP, seems unlikely, considering the vast amount of research directed so far towards improving and expanding CELP to rates below 4 kb/s. We believe that there is greater promise for a solution based on a high quality parametric coding structure combined with efficient quantization techniques. Harmonic coders, such as the multiband excitation (MBE) coder [1], offer an effective model for speech representation by preserving the perceptually important features of the short term spectra of the speech signal with relatively few parameters. The MBE spectral model coupled with sinusoidal and noise synthesis from the model parameters provides very high quality synthetic speech when the parameters are unquantized. The model handles the mixed voicing condition where certain bands of the short term spectrum exhibit distinct pitch harmonics while other bands have a random-like unvoiced character. Another advantage of this model is its robustness to background noise. Recently, several researchers have begun to study low bit rate speech coding using variations of the MBE algorithm [3-8].

In our search for an efficient, high-quality, low bit-rate speech coding algorithm, we have selected MBB as a starting point and introduced phonetic classification as a key component to enhance the MBE modeling. Since speech is composed of distinct phonetic segments as well as background noise and silence, different phonetic classes need different time-frequency resolutions as well as different degrees of coding accuracy. We have found that class-based processing facilitates both modeling and quantization, improving the overall quality.

The MBE model offers the potential of good speech quality with a parameter set that is very compact compared to the much larger set of signal entities that needs to be quantized in CELP. Therefore, it is potentially well-suited for a low bit rate implementation. However, efficient quantization of the spectral samples in MBE has been a formidable obstacle to achieving at 2.4 kb/s the quality of unquantized MBE. Since the spectral shape is represented by a set of spectral magnitude samples taken at the harmonics of the estimated pitch, the vector formed by the spectral samples has a variable dimension. In this paper, we will call this vector the spectral shape vector or SSV. Since the pitch varies from frame to frame as well as from speaker to speaker, the dimension of the SSV varies, thereby prohibiting the use of conventional vector quantization (VQ) techniques to encode the SSV.

Several indirect methods [2-8] have been studied in the past to circumvent the problem of variable dimensionality. A few coders [3, 5, 6, 7] convert the variable dimension VQ problem to a fixed dimension VQ problem by approximating the spectral shape with an all-pole model and then vector quantizing the parameter set of the fixed dimension model. This approach incurs the extra penalty of modeling error, often leading to audible distortions such as reverberations. The Inmanat-M IMBE standard algorithm [2] uses the DCT with scalar quantization of the DCT coefficients, allocating bits and stop sizes according to the dimension. The resulting highly complex scheme requires a large number of bits and does not offer the advantage of vector quantization over scalar quantization. Two other coders [4, 8] convert the SSVs to fixed dimension vectors using sampling rate conversion followed by standard VQ methods. The error due to dimension conversion is also quite significant. For harmonic coders, the overall perceptual quality heavily depends on the quality of the spectral quantization. At low bit rates, it is sufficiently challenging to limit quantization errors to tolerable levels; there is no margin to tolerate any added distortion due to modeling or dimension-altering methods.

In this paper, we present a 2.4 kb/s enhanced multiband excitation (EMBE) algorithm which integrates phonetic classification into
the MBE spectral model. Class-dependent processing enables EMBE to deliver coding resolution effectively among various phonetic classes and to improve the modeling of unvoiced segments. A novel quantization scheme, called variable dimension vector quantization (VDVQ), directly and efficiently quantizes the variable dimension SSVs. Subjective quality comparisons with the 4.3 kbps U.S. Federal Standard CELP 1016 [9] coder and the 4.15 kbps Inmarsat-M MBE coder [2] demonstrate that the EMBE coder delivers speech quality comparable to or better than these two standard coders.

The system overview of the EMBE algorithm is presented in Figure 1. First, each frame of speech is classified by the phonetic classifier. Based on the class information, a spectral model is chosen and the estimated spectral parameters are quantized by class-based quantization schemes. Finally, the output speech frame is synthesized using the class information and the decoded spectral parameters.

![Image of the EMBE coder - system overview](image)

**Figure 1. The EMBE coder - system overview**

2. The EMBE Model

As in MBE [1, 2], EMBE models the short term spectrum by specifying 1) the pitch or fundamental frequency (F₀), 2) the frequency-domain multiband voicing/unvoicing decisions (VUV), and 3) the spectral shape vector. Two key problems of the MBE model (as in [2]) are errors due to incorrect pitch estimations and incorrect modeling of high-energy unvoiced frames. The first problem causes distinctly audible artifacts, especially when the spectral quantization noise is high. The second problem adds undue periodicity in high-energy unvoiced segments (see Figure 2) as well as in noisy speech, creating annoying audible artifacts.

In the MBE algorithm, the pitch is estimated during unvoiced segments as well. Thus, often during unvoiced segments, the pitch tracker goes away. The heuristics used to bring the pitch back to the correct track often do not work. As a result, during the perceptually crucial unvoiced-to-voiced transitions, often the pitch-estimator takes a few frames to recover from pitch-errors. Since the spectral magnitude shape is sampled at harmonics of the estimated "pitch", the random nature of the pitch estimator during unvoiced segments may lead to another vulnerability. The estimated pitch frequency could be too high, thereby producing an SSV containing only a few samples, too inadequate to represent the acoustical signature in the spectral shape. As a result, the synthetic unvoiced sound could be muffled and non-crisp.

Integration of an external phonetic classifier in the EMBE model mitigates all of the above problems and enhances the quality and robustness of the model itself. As we will see later, class-based processing enables EMBE to distribute higher coding resolutions to some classes than does the standard MBE approach. Class-based quantization also delivers higher quantization efficiency since we can use custom-designed codebooks -- one for each class.

![Image of processing of high energy unvoiced segments](image)

**Figure 2. Processing of high energy unvoiced segments**

2.1 Phonetic Classification of EMBE

The voice activity detector [10] classifies each 20 ms input speech frame either as active speech or noise (N). All non-noise frames are processed by the active speech classifier which labels them as either unvoiced (UV) or voiced. The voiced/unvoiced decision is made by extracting and evaluating a number of parameters including energy, spectral tilt, low-band energy, and periodicity. For a voiced frame, the pitch F₀ and the binary-valued voicing/unvoicing vector VUV are estimated. If all components of VUV are one then the voiced frame is classified as fully voiced (FV). Otherwise, the voiced frame is classified as mixed voiced (MV).

2.2 Class-based Modeling of EMBE

In the EMBE coder, pitch is estimated over the full range only during an unvoiced to voiced frame (MV or FV) transition. During a voiced-to-voiced frame transition, pitch is "tracked" within a window around the past pitch value. During unvoiced frames, pitch is not estimated and a fixed number of spectral magnitude samples are taken to form the SSV. The resulting class-based pitch estimator delivers better accuracy at significantly less complexity than the MBE pitch tracker. The fixed-rate spectral sampling during unvoiced frames guarantees an adequate representation of spectral shape leading to better quality. Further-
more, a fixed dimension allows the use of conventional VQ
techniques for the unvoiced SSVs. External classification also
eliminates the artifacts of MBB during high-energy unvoiced
segments (Figure 2).

For an N or UV frame, the spectrum is represented only by a fixed
dimension SSV. For an FV frame, the spectrum is represented by
Fo and the variable dimension SSV. For MV type frames, the
spectrum is represented by all three: Fo, UBV and SSV. For N
and UV type frames, all components of UBV are forced to zero.

The EMBE synthesis is almost identical to the MBB synthesis.
Minor modifications have been made to take care of the fixed-rate
sampling of the spectrum during N or UV type frames.

3. Quantization of EMBE Parameters

The pitch Fo is quantized with 8 bits. The UBV information is
itself binary and can have a maximum of 12 bits (see [2] for
details). It is encoded with four numbers, n1, n2, n3, and n4, using
8 bits. The approximation of UBV is formed by n1 zeros, then n2
ones, then n3 zeros, then n4 ones, the rest (if any) is zero padded.
One bit is given to n4, 3 to n3, 2 to n2, and 2 to n1. Observations
indicate that in most cases the UBV pattern can be perfectly rep-
resented using this scheme. For the very few cases of mismatch,
the resulting error was perceptually insignificant.

The SSV, denoted S, is quantized as follows. First it is converted
to the log domain and mean-removed vector quantization
(MRVQ) is applied as shown in Figure 3.

![Figure 3. Spectral Shape Vector encoding/decoding](image)

Note that MRVQ in the log domain is actually gain-shape VQ in
the spectral domain and R represents the logarithm of the normal-
ized spectral shape. For UV and N frames, the VDOQ block is
replaced by a split VQ module. Table 1 presents the bit allocation
of the EMBE coder. Note that, phonetic classification enables
EMBE to allocate extra bits to spectral shape quantization for cer-
tain classes (16 for UV and N and 8 for FV). A larger bit alloca-
tion for the spectral shape leads to higher quality.

<table>
<thead>
<tr>
<th>Table 1: Bit allocations of the EMBE coder</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHONETIC CLASS</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Class</td>
</tr>
<tr>
<td>Pitch</td>
</tr>
<tr>
<td>VBV</td>
</tr>
<tr>
<td>Spectral Gain</td>
</tr>
<tr>
<td>Spectral Shape</td>
</tr>
<tr>
<td>Total Bits</td>
</tr>
</tbody>
</table>

The basic principle of VDOQ is presented next.

3.1 Variable Dimension Vector Quantization

In our approach to spectral quantization, we view the SSV as a
subsampled version of the overall spectral shape of the frame.
During different frames, the spectral shapes are represented by a
different number of samples, taken at a set of frequency locations
dictated by the estimated pitch for each frame. The vocal tract
largely determines the spectral shape, whereas the pitch is deter-
mined by glottal vibrations. The spectral shape and the pitch
(which dictates the sampling) can be assumed to be statistically
independent (a reason-able approximation justified by the physi-
ology of human speech production). Hence, any particular pho-
notome will exhibit roughly the same spectral shape when uttered
by speakers with differing pitch. However, the SSVs will be
formed differently and will have different dimensions when the
same spectral shape is sampled with different Fo values. Thus,
female speech, will generally produce lower dimension SSVs,
since a higher pitch means fewer harmonics cover the voice band.
For male speech, lower pitch values will generate higher dimen-
sional SSVs. Therefore, even though SSVs of varying dimension
are generated due to varying pitch, the underlying task is the
quantization of the overall spectral shape which characterizes the
phonetic character of the frame.

![Figure 4. Spectral encoding using VDOQ](image)

Consequently, we construct a universal spectral shape codebook
C containing a rich collection of K dimensional code vectors Yi
representing various spectral shapes over the full spectral range
represented by K samples at the frequency resolution of the DFT.
Then, any spectral shape vector S can be well approximated by
suitably sub-sampling a code vector Ym with a matching shape
selected by searching through C, regardless of the dimensionality
of S. To find the best matching code vector, we simply compare
the components of S to the corresponding components of each Yi.
In other words, before comparison with S, each universal shape
codevector Yi is sub-sampled to form a vector Ym with the same
dimension as S, in exactly the same way as the SSV was formed from
the full-band spectrum using the pitch information. Figure 4 illus-
trates the basic VDOQ encoding principle. A more detailed presen-
tation of VDOQ can be found in [11].

Therefore, the variable dimension SSVs can be directly quantized
using VDOQ without any further modeling or dimension-altering
techniques. The resulting gain can be seen in Table 2 which com-
pare VDOQ with two other spectral quantization techniques in
terms of spectral distortion (SD) defined as:

\[
SD = \sqrt{\frac{1}{L-1} \sum_{n=1}^{L-1} \left(10\log_5 (n_{av}) - 10\log_5 (\hat{n}_{av})\right)^2}
\]

SD: original SSV of dimension L; \( \hat{S} \): quantized SSV
The average SD value is computed by processing a large number of frames from a multi-speaker database. In Table 2, LP-10 refers to a tenth order all-pole model, similar to [3], followed by split VQ of the 10 LSFs with scalar gain quantization. To encode the SSVs, 30 bits were used for both LP-10 and VDVQ and 63 for IMBE [2]. The superiority of VDVQ is quite evident in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>IMBE</th>
<th>LP-10</th>
<th>VDVQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male Speech</td>
<td>3.00</td>
<td>2.75</td>
<td>1.45</td>
</tr>
<tr>
<td>Female Speech</td>
<td>3.03</td>
<td>2.28</td>
<td>1.18</td>
</tr>
<tr>
<td>Overall</td>
<td>3.02</td>
<td>2.52</td>
<td>1.31</td>
</tr>
</tbody>
</table>

4. Performance of The EMBE Coder

The 2.4 kbps EMBE coder was compared with two higher rate standard coders (the Inmarsat-M IMBE coder at 4.15 kbps and U.S. Federal Standard 1016 CELP coder at 4.8 kbps). A-B comparison tests were conducted in which 20 listeners each compared the subjective quality of 64 sentence pairs: 16 EMBE vs. IMBE clean sentence pairs, 16 EMBE vs. CELP clean sentence pairs, 16 EMBE vs. IMBE noisy sentence pairs and 16 EMBE vs. CELP noisy sentence pairs. The noisy sentences have vehicle noise (15 dB SNR) in the background. The A-B ordering of the sentence pairs was randomized. Listeners were allowed to indicate "same" if they could not distinguish the quality of the two sentences.

![Figure 5. Percentage preference of EMBE when compared with higher rate standard coders](image)

Clean Speech Noisy Speech Clean Speech Noisy Speech

EMBE vs. CELP EMBE vs. IMBE

prefer prefer prefer prefer
EMBE EMBE EMBE EMBE
same same same same
24 78 10 12
62 132 35 177

The results (Figure 5) indicate that for clean speech, EMBE is comparable in quality to IMBE. However, the distinctly higher margin of preference of EMBE over CELP-1016 clearly demonstrates the superiority of EMBE over CELP-1016. For noisy speech, EMBE is judged to be superior in quality to both IMBE and CELP-1016. Considering that the bit rates of the CELP and IMBE reference coders are about twice the EMBE rate, the EMBE algorithm clearly demonstrates a significant improvement in rate-distortion trade-off.

5. Conclusions

We have presented a 2.4 kbps enhanced multiband excitation (EMBE) speech coding algorithm which exploits phonetic classification and a novel spectral quantization technique to deliver high speech quality. Integration of phonetic classification enables efficient distribution of coding resolutions among the various phonetic classes enhancing the quality and robustness of the model. A class-based coding approach also delivers significantly higher quantization efficiency since we can use custom-designed codebooks - one for each class. A novel direct quantization scheme (VDVQ) effectively encodes the variable dimensional spectral shape vectors with relatively few bits while preserving the perceptually important features. Subjective quality evaluations indicate that the quality of the EMBE coder is superior to the 4.8 kbps CELP-1016 coder and comparable to the 4.15 kbps IMBE coder. For noisy speech, EMBE is judged to be distinctly superior to both CELP-1016 and IMBE.

We have also extended the EMBE algorithm to obtain a variable rate coder, the variable rate spectral coder (VRSC) [12], which uses variable bit allocations and different coding strategies for different phonetic classes and operates at an average rate of 1.4 kbps while offering speech quality roughly equivalent to the 2.4 kbps EMBE coder.

6. References


Optimal Recursive Weighted Order Statistic Filters for Efficient High Speed Implementations

Lori E. LUCKE and Randall A. KRONEKKE

Dept. of Electrical Engineering
University of Minnesota
Minneapolis, MN 55455, USA
email: lucke@ee.umn.edu, rkroenke@ee.umn.edu
Tel/Fax: 612-625-4388, 612-625-4383

ABSTRACT. The median filter is a popular non-linear filter that removes impulsive noise while retaining edge information. This class of filters has been extended to weighted order statistics (WOS) filters, in order to improve the flexibility of the filtering operation. As in linear filtering, optimal filtering theory has been developed for non-recursive WOS filters. This theory is analogous to optimal FIR filtering. In this paper we develop optimal filtering theory for recursive WOS filters. A recursive WOS filter contains previous calculated outputs within the sample window and is analogous to the IIR filter. We show the optimal recursive WOS filter requires fewer sample points within the sample window compared to the optimal non-recursive WOS filter. This is important because it reduces the complexity of the corresponding implementation of the WOS filter. The disadvantage of the recursive WOS filter is that it is difficult to pipeline due to the feedback loop. We show how to generate high speed recursive WOS filters using a technique similar to lookahead which allows us to generate efficient pipelined implementations. Using a combination of lookahead and reduced sample window size, we can generate high speed implementations for recursive WOS filters with reduced complexity.

1. INTRODUCTION

The median filter is a popular non-linear filter that preserves edges and attenuates impulsive noises. The median filter chooses the median of a sample window as its output. Several classes of rank order based filters have recently been developed which improve the filtering response of the median filter class. These include stack filters [1] and weighted order statistic (WOS) filters [2]. The median filter does not account for temporal information during the filtering operation while the stack filter does. The WOS filter is a subset of the stack filter class and a superset of the median filter class. It weights each sample in the sample window thus allowing some temporal control of the filtering operation although not to the extent of the stack filter.

As in linear filtering, optimal filtering theory has been developed for the weighted order statistic filter [3, 4]. In [4] we developed optimal theory for recursive WOS filters. A recursive structure operates on both input samples and previously calculated outputs. The advantage of recursive structures is that they require fewer points within the sample window to generate similar results as the equivalent non-recursive structure. This is very important from an implementation standpoint since the complexity of the implementation is dependent on the size of the sample window. The disadvantage of the recursive structure is that they contain feedback loops which make them difficult to pipeline.

We overcome this limitation by using approximate lookahead techniques [5]. In this paper, we extend the development of optimal filter theory for recursive WOS filters to support lookahead. Recently we introduced several VLSI architectures to implement recursive weighted order statistic filters [6]. The use of lookahead allows us to pipeline these implementations and operate them at a higher speed. The disadvantage of lookahead is an increased window size. However the recursive structures still have the advantage of reduced sample size. Therefore, it is possible to offset the increase in sample window size due to lookahead, by the reduced complexity afforded due to the recursion. In this paper we extend optimal filtering theory for recursive WOS structures to support lookahead. We demonstrate the increased complexity of the lookahead structure can be reduced by using smaller sample window sizes.

2. WEIGHTED ORDER STATISTICS FILTERS

A recursive WOS filter operates on a window of sample values. The sample window is defined as

\[
\tilde{R}(n) = [\tilde{S}(n-N), \ldots, \tilde{S}(n-1), X(n), \ldots, X(n+N)]^T
\]

where the \(X\) samples are the non-recursive inputs and the \(\tilde{S}\) samples are the previously calculated outputs or recursive samples. The size of the sample window is \(Z = 2N + 1\). By comparison, the non-recursive sample window only contains the non-recursive input samples \(X(n-N)\) to \(X(n+N)\). The sample window may also be two-dimensional with size \(Z = (2N + 1)(2M + 1)\).
For a two-dimensional window, the recursive samples depend on the direction the pixels are scanned within the image. For example, with horizontal scanning, the first $N$ rows plus the first $M$ samples of the $N + 1$ row contain recursive samples. We assume horizontal scanning in this paper although all the techniques also apply for other types of two-dimensional windows. All equations are shown using one-dimensional notation but apply equally well in two dimensions.

The recursive WOS filter chooses an output based on the weighted ranks of the inputs within the sample window and can be described [3] by

$$
\hat{S}(n) = T \text{ largest of} \nonumber
\begin{align*}
&W_{-N} \odot \hat{S}(n-N), \ldots, W_{-1} \odot \hat{S}(n-1), \\
&W_0 \odot X(n), \ldots, W_N \odot X(n+N)
\end{align*}
$$

where $T$ is the rank or threshold of the filter, $\tilde{W} = [W_{-N}, \ldots, W_N]^T$ is an integer weight vector, and $\odot$ denotes duplication of the sample by the corresponding weight value.

3. SUPPORTING LOOKAHEAD

We recently introduced several architectures to implement recursive WOS filters [6]. The sample period of these recursive architectures is limited by the delay in the feedback loop. One way to speed up these architectures is with pipelining but this is difficult because of the inherent recursion. To support pipelining, the recursive algorithm must be restructured to create additional delays in the feedback paths, which can be used for pipelining. This is accomplished using lookahead [5]. There are two types of lookahead: scattered, where the algorithm is iterated to generate a function dependent on scattered past states, and clustered, where the algorithm is iterated to generate a function dependent on consecutive past states. The type of lookahead used depends on the implementation. For lookahead, the $N$th order recursive function $y(n) = F[y(n-1), y(n-2), \ldots, y(n-N)]$ is iterated $A$ times to become

$$
y(n + A) = F[y(n - 1), y(n - 1 - (B + 1)), \ldots, y(n - 1 - (N - 1)(B + 1))],
$$

where $B = A$ for scattered lookahead and $B = 0$ for clustered lookahead. The $A$ delays in the iterated function are used to pipeline the corresponding recursive realization. With $A = 1$ there is one level of pipelining and a potential speedup of 2. For $A$ levels of pipelining, the potential speedup is $A + 1$.

Lookahead techniques cannot be applied directly to the non-linear recursive WOS filters because it is not possible to exactly iterate the non-linear WOS function. Instead we train the WOS filter using the lookahead data dependencies. To support lookahead in the recursive WOS filter the sample window of Eq. 1 is replaced with a sample window with the lookahead data dependencies. For example, for clustered lookahead, the sample window becomes

$$
\tilde{R}(n) = [\hat{S}(n-N-K), \ldots, \hat{S}(n-K), \\
X(n-K+1), \ldots, X(n+N)]^T
$$

where $K$ is an integer dependent on the amount of lookahead.

For two-dimensional windows, we apply lookahead in the direction of scanning. For example, with horizontal scanning a $3 \times 3$ window, $\tilde{R}(n,m)$, with one level of lookahead becomes

$$
\hat{S}(n-1, m-2), \hat{S}(n-1, m-1), \hat{S}(n-1, m), \hat{S}(n-1, m+1), \\
\hat{S}(n, m-2), X(n, m-1), X(n, m), X(n, m+1), \\
X(n+1, m-2), X(n+1, m-1), X(n+1, m), X(n+1, m+1)
$$

The cost of lookahead is an increase in the size of the sample window. The sample window increases by one column for every level of lookahead.

4. OPTIMAL RECURSIVE WOS FILTERS UNDER THE MAE CRITERION

The optimal recursive WOS filter is found in the binary domain by using threshold decomposition [1]. An $M$-valued signal, $X(n)$, can be translated into $M$ binary signals $x(z)^1, \ldots, x(z)^M$ using threshold decomposition defined as in [1, 3]. Threshold decomposition can be applied to the sample window of Eq. 4 to generate a decomposed sample window, $x^m(n)$.

The output of the WOS filter can be described in the binary domain as

$$
\hat{S}(n) = \text{WOS}(\tilde{R}(n)) = \sum_{m=1}^{M} U(\tilde{W}^T x^m(n) - T_{\text{wo}})
$$

where $U$ is the unit step function and $\tilde{R}(n)$ describes the appropriate sample window with lookahead.

As in linear IIR filtering [7], we attempt to optimize the recursive WOS filter, shown in Fig. 1, such that mean absolute error (MAE), between the filter output and the desired signal, is minimized. The mean absolute error for this filter is the difference between the desired output and filter output,

$$
\text{J}(\tilde{W}) = E[|\hat{S}(n) - \hat{S}(n)|]
$$

where $\hat{S}(n)$ represents the desired output and $\hat{S}(n)$ is the filter output defined in Eq. 5. We can rewrite this using binary signals as in [3]. Our optimization problem becomes

$$
\min_{\tilde{W}} \text{J}(\tilde{W}) = \sum_{m=1}^{M} E[|s^m(n) - \hat{s}^m(n)|]^2
$$

where we constrain the weights in the weight vector to be greater than or equal to zero. Note, in the binary domain the MAE is equivalent to the MSE (mean square error) [3].

We use an iterative solution to minimize the error function. Our iterative solution is based on a gradient search of the MSE performance surface. The weights are updated iteratively using

$$
\tilde{W}(n+1) = \tilde{W}(n) - 0.5 \mu \nabla \text{J}(\tilde{W}(n)).
$$

To find $\nabla \text{J}$ we approximate the non-linear unit step function with a linear function of $\tilde{W}$,

$$
\hat{s}^m(n) = \tilde{W}^T x^m(n) - T_{\text{wo}}.
$$

Then the gradient of $\text{J}$ becomes

$$
\nabla \text{J}(\tilde{W}(n)) = -2 \sum_{m=1}^{M} (s^m(n) - \hat{s}^m(n)) \nabla \hat{s}^m(n)
$$
Figure 1: Optimal recursive WOS filter

\[ \nabla \delta^m(n) = \tau^m(n) + \sum_{i=1}^{i=N} W_{-i}(n) \nabla \delta^m(n-i). \]  

The gradient of \( \delta^m(n) \) is unfortunately a complicated recursive formula. Since \( \nabla \delta^m(n) \) is small compared to \( \tau^m(n) \) we can estimate it with \( \nabla \delta^m(n) \approx \tau^m(n) \).

5. SIMULATIONS

We decided to use our iterative solution for the recursive WOS filter to train the weights of the filter on several experiments. First we demonstrate that clustered lookahead does not significantly affect the quality of the one-dimensional recursive WOS filter. In Table 1 we show the results of filtering two sets of test data corrupted with 5% or 15% impulsive noise with recursive WOS filters with varying amounts of lookahead. The test data was taken from two gray-scale images "Lenna" and "Sailboat". These images were used as a readily available source of test data. We measured the mean absolute error (MAE) and the mean square error (MSE) between the filtered image and the original image. From the error measurements it is apparent that clustered lookahead does not cause a significant degradation of the filter performance. We achieved similar results with scattered lookahead.

More interesting results were obtained when filtering in two dimensions. Again we used the gray-scale images "Lenna" and "Sailboat". We corrupted these images with varying amounts of impulsive noise. In Table 2 we show the results of filtering the test two images with both non-recursive WOS filters trained as in [3] and recursive WOS filters trained using our method. For each trained filter there is a corresponding set of weights. Many of these weights are very small and are equivalent to zero. In Table 2 we show the number of nonzero weights for each sample window. The recursive WOS filter requires fewer sample points to provide similar results and thus requires a smaller sample window. Therefore the recursive implementation is less complex than the non-recursive WOS implementation.

In Table 3 we show the results of filtering our test images with the data dependencies required by lookahead. Each level of lookahead allows us to add another level of pipelining to speed up the recursive implementation. The results compare favorably with Table 2. We show the original sample window for each test but the actual sample window increases by one column for each level of lookahead. However it is apparent from

<table>
<thead>
<tr>
<th>Window</th>
<th>K</th>
<th>Noise</th>
<th>Test MAE</th>
<th>Test MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5x3</td>
<td>5%</td>
<td>1</td>
<td>2.28</td>
<td>41.34</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>1</td>
<td>4.08</td>
<td>89.27</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>2</td>
<td>1.96</td>
<td>43.87</td>
</tr>
<tr>
<td>3x3</td>
<td>15%</td>
<td>2</td>
<td>4.46</td>
<td>105.01</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>1</td>
<td>1.78</td>
<td>40.83</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>2</td>
<td>4.39</td>
<td>95.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Window</th>
<th>K</th>
<th>Noise</th>
<th>Test MAE</th>
<th>Test MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5x3</td>
<td>5%</td>
<td>1</td>
<td>2.28</td>
<td>41.34</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>1</td>
<td>4.08</td>
<td>89.27</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>2</td>
<td>1.96</td>
<td>43.87</td>
</tr>
<tr>
<td>3x3</td>
<td>15%</td>
<td>2</td>
<td>4.46</td>
<td>105.01</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>1</td>
<td>1.78</td>
<td>40.83</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>2</td>
<td>4.39</td>
<td>95.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Window</th>
<th>K</th>
<th>Noise</th>
<th>Test MAE</th>
<th>Test MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5x3</td>
<td>5%</td>
<td>1</td>
<td>2.28</td>
<td>41.34</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>1</td>
<td>4.08</td>
<td>89.27</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>2</td>
<td>1.96</td>
<td>43.87</td>
</tr>
<tr>
<td>3x3</td>
<td>15%</td>
<td>2</td>
<td>4.46</td>
<td>105.01</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>1</td>
<td>1.78</td>
<td>40.83</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>2</td>
<td>4.39</td>
<td>95.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Window</th>
<th>K</th>
<th>Noise</th>
<th>Test MAE</th>
<th>Test MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5x3</td>
<td>5%</td>
<td>1</td>
<td>2.28</td>
<td>41.34</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>1</td>
<td>4.08</td>
<td>89.27</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>2</td>
<td>1.96</td>
<td>43.87</td>
</tr>
<tr>
<td>3x3</td>
<td>15%</td>
<td>2</td>
<td>4.46</td>
<td>105.01</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>1</td>
<td>1.78</td>
<td>40.83</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>2</td>
<td>4.39</td>
<td>95.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Window</th>
<th>K</th>
<th>Noise</th>
<th>Test MAE</th>
<th>Test MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5x3</td>
<td>5%</td>
<td>1</td>
<td>2.28</td>
<td>41.34</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>1</td>
<td>4.08</td>
<td>89.27</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>2</td>
<td>1.96</td>
<td>43.87</td>
</tr>
<tr>
<td>3x3</td>
<td>15%</td>
<td>2</td>
<td>4.46</td>
<td>105.01</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>1</td>
<td>1.78</td>
<td>40.83</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>2</td>
<td>4.39</td>
<td>95.75</td>
</tr>
</tbody>
</table>

Table 1: Measured errors for filtering noisy test data with trained recursive 1D WOS filters and K levels of lookahead.

<table>
<thead>
<tr>
<th>Window</th>
<th>Noise</th>
<th>Test MAE</th>
<th>Test MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5x3</td>
<td>5%</td>
<td>1.80</td>
<td>41.74</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>4.06</td>
<td>89.37</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>1.96</td>
<td>43.87</td>
</tr>
<tr>
<td>3x3</td>
<td>15%</td>
<td>4.46</td>
<td>105.01</td>
</tr>
<tr>
<td>5x5</td>
<td>5%</td>
<td>1.78</td>
<td>40.83</td>
</tr>
<tr>
<td>5x5</td>
<td>5%</td>
<td>4.39</td>
<td>95.75</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>2.32</td>
<td>245.57</td>
</tr>
<tr>
<td>5x5</td>
<td>2.64</td>
<td>283.01</td>
<td></td>
</tr>
<tr>
<td>5x5</td>
<td>5.88</td>
<td>183.71</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Measured errors and nonzero weights for filtering noisy images with trained non-recursive and recursive 2D WOS filters.

<table>
<thead>
<tr>
<th>Window</th>
<th>Noise</th>
<th>Test MAE</th>
<th>Test MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5x3</td>
<td>5%</td>
<td>1.80</td>
<td>41.74</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>4.06</td>
<td>89.37</td>
</tr>
<tr>
<td>3x3</td>
<td>5%</td>
<td>1.96</td>
<td>43.87</td>
</tr>
<tr>
<td>3x3</td>
<td>15%</td>
<td>4.46</td>
<td>105.01</td>
</tr>
<tr>
<td>5x5</td>
<td>5%</td>
<td>1.78</td>
<td>40.83</td>
</tr>
<tr>
<td>5x5</td>
<td>5%</td>
<td>4.39</td>
<td>95.75</td>
</tr>
<tr>
<td>5x5</td>
<td>15%</td>
<td>2.32</td>
<td>245.57</td>
</tr>
<tr>
<td>5x5</td>
<td>2.64</td>
<td>283.01</td>
<td></td>
</tr>
<tr>
<td>5x5</td>
<td>5.88</td>
<td>183.71</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Measured errors and nonzero weights for filtering noisy images with trained pipelined recursive 2D WOS filters.
the number of nonzero points in each sample window that in reality the cost of lookahead does not cause the window size to increase significantly above its original size. Therefore lookahead does not significantly increase the complexity of the recursive WOS filters.

Finally in Fig. 5 we show the results of filtering the "Lenna" image with both a 5x5 non-recursive WOS and several recursive WOS filters. The recursive WOS filters were trained using a 5x5 window but the actual filtering window had several zeroed weights, as shown in Tables 2 and 3, while the non-recursive WOS filter had only a single zeroed weight. Also it is apparent that lookahead does not significantly impact the optimal filtering results.

6. REFERENCES


Adaptive Skeletonization using Multistage Boolean and Stack Filtering

Doina Petrescu*  Ioan Tăbuș†  Moncef Gabbouj

Signal Processing Laboratory
Tampere University of Technology
P.O. Box 553, FIN-33101 Tampere, Finland

Abstract This paper develops novel and robust adaptive procedures for skeletonization. The robustification of the image skeletonization process is achieved by multilayer structures of boolean or stack filters. The morphological operators used in the classical skeleton transform are replaced with adaptively designed multilayer stack (boolean) filters, in which different layers are assigned for solving the task of noise rejection and the task of shape (geometry) transformation.

The reconstruction process, which in the noise free case perfectly recovers the original image from its skeleton, now leads to an image very close to the original (uncorrupted) image. This property generates the possibility of using the pair of adaptive skeletonization and reconstruction process, possessing good noise suppression capabilities, as a very efficient tool in image coding and decoding.

In order to illustrate the performance of the new methods, the adaptive skeleton transform is applied to a commutative noise corrupted image, the results obtained showing dramatic improvement over the classic skeletonization method.

\[
S^{sk}(X) = (X \ominus nB) - ((X \ominus nB) \circ B)
\]

\[
S^R(X) = \bigcup_{n=0}^{\infty} S^{[n]}(X)
\]

where \( \ominus, \circ \) and \( \circ \) denote respectively the discrete morphological erosion, dilation and opening, using structuring element \( B \).

One strong limitation in the use of skeleton is its extreme sensitivity to noise, which may alter the results to such an extent that the recognition process is not anymore possible or the coding efficiency is severely degraded.

In this paper, we propose an approach for robustifying the skeletonization process against noise, using filtering stages based on boolean or stack filters dedicated to noise rejection and to geometrical shape transformation.

To solve the design problem of optimal stack filters for an arbitrary window size, we developed in [4] a multilayer filtering architecture. Each layer represents a boolean or a stack filter and the outputs of the intermediate filtering layers provide some partial solutions for the optimization problem while the final solution is provided by the last output layer.

1 Introduction

The skeleton transform was first introduced in connection with continuous images, as a way to find a "medial axis" of the objects in the image.

The skeleton for a binary discrete image object is given by [1],

\[
S^{sk}(X) = (X \ominus nB) - ((X \ominus nB) \circ B)
\]

\*On leave from Polytechnic Institute of Bucharest, Department of Electronics and Telecommunications

†On leave from Polytechnic Institute of Bucharest, Department of Control and Computers

2 Adaptive Skeletonization

In adaptive signal processing, the classical assumption is that part of the "target" signal is known, the goal being to design a processing structure able to transform the "perturbed" signal into an "output" signal as close as possible to the target.

We assume that the image \( X_0 \) for which the skeleton is required is corrupted by noise, but one subset \( I_0 \) of the original image is still available. Then the following three processes are carried on (possibly in
parallel):

a) the skeleton transform is applied to the subset \( I_0 \) of the original image;

b) an adaptive multistage filter design procedure runs, receiving as reference images for the various stages, the images obtained in the intermediate stages of the skeletonization of \( I_0 \), and determines the functions \( f[m] \) which are needed to filter the noisy \( I_0 \) in order to be as close as possible to the way the skeleton transform processes the subset of the original image;

c) the functions \( f[m] \) are used to process the whole noisy image \( X_\nu \), resulting in the adaptive skeleton.

We define the "image distance" \( d_I \) between the binary images \( I_1 \) and \( I_2 \) as the sum of absolute errors between all corresponding pixel values

\[
d_I(I_1, I_2) = \sum_{i=1}^{nrow} \sum_{j=1}^{ncol} |I_1(i, j) - I_2(i, j)|. \tag{2}
\]

The problem of adaptive skeletonization may now be stated:

**Problem (Adaptive skeletonization)** Given

- a subset \( I_0 \) of the original image \( X_\nu \),
- the noisy image \( X_\nu \) (with \( I_0 \) the noisy set resulting after corrupting \( I_0 \) ),

*find* for all stages \( m \) the multistage transforms (i.e. the filter functions and the template functions)

\[
\Psi f[m, l], T_{temp}[n, m] \odot \ldots \odot \Psi f[l, m], T_{temp}[l, m] \tag{3}
\]

which minimize the set-difference distance function

\[
J_d(m) = d_I(S[m](I_0), \Psi f[l, m], T_{temp}[l, m] \odot \ldots \odot \Psi f[l, m], T_{temp}[l, m](I_0)) \tag{4}
\]

3 Stack Filter Based Adaptive Skeleton

To solve the problem stated above, we propose the Adaptive Skeleton Transform of the noisy image \( X_\nu \) using the structuring element \( B \). It is assumed that the upperleft quarter, \( I_0 \) of the original image \( X \) is available and one recursive structure computes its standard Morphological Skeleton Transform. The resulting subsets in the intermediate stages of the computation of the skeleton of \( I_0 \) are used as targets for designing the optimal filters for the implementation of the Adaptive Skeleton Transform. We use robust filters, namely symmetric stack (SymS) filters in the skeletonization process. They are less sensitive with respect to the noise realisation and noise distribution and less signal dependent[3] than the fast (but lacking good generalization property) optimal boolean filters.

These filters are then applied to the whole image, according to the structure of the multistage filter, producing the components of the skeleton for the whole noisy image.

Three types of stack filters are designed:

1. The noise rejecting filter designed to be the optimal one in recovering the image \( I_0 \) from \( I_\nu \). The template \( Temp^{[0]} \) associated with it can be an arbitrary one.

2. The erosion type filters, are filters with template \( Temp^{[0]} = B \), designed to be close to \( B \) structuring element erosion. In the noise free case the filter obtained is identical to erosion.

3. The opening type filters are filters with template \( Temp^{[0]} = B \oplus B \) designed to be close to \( B \) opening. They are obtained as a cascade of two filtering stages, the first being designed to be close to the erosion by \( B \) and the second to estimate the dilation by \( B \). In the noise free case these filters are identical to morphological erosion and dilation.

The structure of the adaptive algorithm is depicted in Fig.5.

4 Experimental Results

In order to compare the performance of the Adaptive Skeleton Transform with those of the standard Morphological Skeleton Transform, both transforms were applied to a noisy test image. The original binary (512 × 512) test image in Fig.1 was perturbed with uniform distributed **commutative** noise (10% of image pixels are complemented).

The standard morphological skeleton of the original image using \( 3 \times 3 \) structuring element is shown in Fig.2. When applied to the noise corrupted image the standard Morphological Skeleton Transform produces a "skeleton" completely buried in noise, as may be seen in Fig.3.
In contrast, Fig. 4 shows the excellent ability of the Adaptive Skeleton Transform to recover, from the noisy image, a "skeleton" carrying useful information about the medial axis of the objects.

The adaptive skeleton procedure utilizes, in the first stage of filtering, an optimal multistage stack filtering (with an equivalent window size of $5 \times 5$).

Analyzing the filters obtained through adaptation in the different stages of the design procedures leads to the conclusion that some of the filters obtained are identical to erosion and dilation, especially for the last stages of the adaptive procedure, but in the first stages they differ from erosion and dilation filters.

Applying the reconstruction procedure based on dilations to the adaptive skeleton sets, the resulting image is much closer to the original image (only 0.3% of the pixels are different) than the perturbed image (where 10% of the pixels are different).

5 Conclusions

We proposed a robust skeletal decomposition of binary images substituting at each stage of the current implementation of the Morphological Skeleton Transform the standard erosion and standard dilation with adaptively designed stack filters.

The Adaptive Skeleton Transform is identical to the standard Morphological Skeleton Transform for noise-free images, but for corrupted images they differ significantly, the former producing a useful "medial axis" of the objects in the image, while the latter delivering an extremely distorted skeleton loosing any significance as medial axis.

Our results emphasized that the adaptive skeletonization procedure has also an important filtering effect over the perturbed images, recommending the application of skeleton in noisy image coding, when the accuracy of reconstructed images is of prime interest.

References


Figure 3: Skeleton of noise corrupted IMTEST (10% of image pixels are complemented)

Figure 4: Adaptive skeleton (based on symmetric stack filtering) of noise corrupted IMTEST

Figure 5: Adaptive Skeleton Transform based on Multistage Stack Filtering
Optimal L-filters for vector magnitude filtering

Nikos NIKOLAIDIS and Ioannis PITAS

Department of Electrical & Computer Engineering, University of Thessaloniki, P. O. Box 463, GR-54006 Thessaloniki, GREECE, Tel/Fax: +30 (31) 996 361 / 274 868, E-Mail: nikolaid@vergina.eng.auth.gr
Tel/Fax: +30 (31) 996 304 / 274 865, E-Mail: pitas@vergina.eng.auth.gr

Abstract. Vector magnitude filtering on vectors represented in polar coordinates can be useful in cases where vector direction is either of no importance or must be preserved intact. This paper focuses on the design of optimal L-filters for the vector magnitude filtering of certain types of two-channel vector fields (rotational, translational and zooming motion vector fields) corrupted by additive noise. Expressions for the unconstrained, unbiased and location invariant L-filter coefficients are derived. Experiments conducted on artificially generated vector fields prove the superior performance of the proposed filters.

1. Introduction

Vector field processing is an important sub-task in many signal processing applications including image sequence processing and stereoscopic image processing. In the case of image sequences, motion vector fields are used in motion compensated coding and compression while disparity vector fields derived from stereo image pairs can be utilized in disparity compensated coding and also provide us with scene depth information. Vector field processing is also closely related to multichannel signal processing (e.g. color image processing) since multichannel observations are essentially vector-valued observations.

So far, vector fields were represented and treated in a cartesian coordinate system. However, vector representation in polar coordinates (i.e. representation in terms of vector magnitude and vector direction) might be preferable in certain situations [1, 2]. Filtering of vector magnitude without affecting its direction (e.g. in estimating wind velocity or car speed from noisy data) is such a case. It is obvious that this task cannot be performed successfully by filtering in cartesian coordinates. Furthermore, transforming certain types of vector fields (e.g. rotational vector fields) in polar coordinates can result in vectors exhibiting interesting properties, like the constant magnitude property. This paper deals with the design of optimal (with respect to the mean square error) L-filters [3] for the filtering of the vector magnitude of two-channel vector fields corrupted by additive bivariate noise. Unconstrained as well as unbiased and location invariant L-filters are studied.

2. Derivation of the optimal filter coefficients

2-D velocity vector fields corresponding to certain types of object motion can be handled more efficiently if transformed to polar coordinates. Such a type of motion is the constant velocity rotation around a fixed point (Figure 1). In this case, the velocity vectors are of constant magnitude on concentric circles around the center of rotation and have the same direction \( \theta \) along each radiat (in fact they are perpendicular to the radiat). A similar case occurs when a camera zooms on a still object, or when an object moves towards a fixed camera.

It is well known [3] that the design of single-channel L-filters for the filtering of additive noise requires the signal to be constant. In the sequel, we shall show that this is also true in the design of L-filters for the filtering of vector magnitude when vectors are corrupted by bivariate additive noise. In the case of the above-mentioned vector field types, vector representation in terms of the polar coordinates \((r, \theta)\) exhibit the constant vector magnitude property when we are moving over certain paths (concentric circles around the center of rotation in the case of rotational fields). Therefore, we can exploit this property to design optimal L-filters. Another type of vector fields that possesses the constant signal property and therefore can be filtered using L-filters is constant velocity translational vector fields. We will present L-filters for this type of vector fields along with filters for rotational fields.

Let us suppose (Figure 1) that the observed two-dimensional vector signal in the case of the constant velocity rotational motion

\[
\mathbf{w}(x, y) = [w_x(x, y), w_y(x, y)]^T
\]

can be expressed as a sum of a signal vector

\[
\mathbf{s}(x, y) = [s_x(x, y), s_y(x, y)]^T
\]

and a zero mean noise vector

\[
\mathbf{n}(x, y) = [n_x(x, y), n_y(x, y)]^T
\]

whose components are distributed according to some bivariate distribution \(f_{n_x, n_y}(n_x, n_y)\). The signal vector \(\mathbf{s}\) can be expressed in polar coordinates as \((r, \theta)\):

\[
r = \sqrt{s_x^2 + s_y^2}, \quad \theta = \arctan \frac{s_y}{s_x}
\]

\(r\) is constant on circles around the center of rotation. Since the vectors in each point are perpendicular to the radiant from the center of rotation to this point, \(\theta\) also defines the position of the vector on the circle. The squared magnitude \(\mathbf{w}\) of \(\mathbf{w}\) can be expressed as follows:

\[
\|\mathbf{w}\|^2 = (n_x + s_x)^2 + (n_y + s_y)^2 = n_x^2 + n_y^2 + r^2 + 2n_x r \cos \theta + 2n_y r \sin \theta = n^2 + r^2 + 2rn \cos \theta
\]

where:

\[
c = [\cos \theta, \sin \theta]^T
\]
where $A$ is the Lagrange multiplier. By differentiating (13) with respect to $a_1, \ldots, a_N, \lambda$, and setting the result to zero, and then solving for $a$, we come to the following expression:

$$a = A^{-1} \left( b - \frac{b^T(A^{-1})^T(r^2 e + \mu N) - r^2}{(r^2 e + \mu N)^T(A^{-1})^T(r^2 e + \mu N)}(r^2 e + \mu N) \right)$$

(14)

An L-filter whose sum of coefficients equals unity i.e. $a^T e = 1$ is said to be location invariant. The coefficients vector that minimizes MSE subject to the previous constraint can also be found by applying the method of Lagrange multipliers and is given by the following expression:

$$a = A^{-1} \left( b - \frac{e^T A^{-1} b - 1}{e^T A^{-1} e} \right) e$$

(15)

In order to proceed with the filter design, the correlation matrix $R_N$ and the mean vector $\mu_N$ for the order statistics vector $N$ have to be evaluated using the expressions for the joint pdf of two order statistics [4]. The evaluation of the moments that appear as elements of $R_N, \mu_N$ was done in a discrete way by quantising the two-dimensional vectors of continuous random variables. The pdf $f_{\nu}(\nu)$ of the random variable $\nu$ must be derived for this calculation. The angle $\theta$, i.e. the position on the circle where the filtering is performed is involved in these calculations. Therefore, different filter coefficients must be evaluated for each position around the circle, resulting in considerably large computational complexity. To overcome this, two different approaches were tested. First, it was assumed that $\theta$ is constant when moving within a small arc on the circle. The coefficients will have to be evaluated for every one of these arcs in order to apply the filter on the entire circle. The computation of $f_{\nu}(\nu)$ under this assumption is done in two stages. First, we use an auxiliary variable $h = n_2$ and evaluate the joint pdf of $\nu, h$, using the formula for the evaluation of the pdf of functions of random variables [5] :

$$f_{\nu|h}(\nu| h) = \frac{f_{\nu_2}(h) - \sin(\theta) + \sqrt{r^2 + (h + \cos\theta)^2}}{2r^2 + (h + \cos\theta)^2}$$

(16)

where:

$$g_1(\nu, \theta) = -r \cos \theta - \sqrt{r^2 + \nu}$$

(17)

$$g_2(\nu, \theta) = -r \cos \theta + \sqrt{r^2 + \nu}$$

(18)

Note that $f_{\nu|h}(\nu| h)$ is zero outside the above-mentioned interval for $\nu, h$. Then, $f_{\nu}(\nu)$ is calculated by integrating (16) with respect to $h$ :

$$f_{\nu}(\nu) = \int_{g_1(\nu, \theta)}^{g_2(\nu, \theta)} \frac{f_{\nu_2}(h) - \sin(\theta) + \sqrt{r^2 + (h + \cos\theta)^2}}{2r^2 + (h + \cos\theta)^2} dh$$

(19)

The integrals in (19) were calculated using numerical integration.

Filtering the rotational field on a small arc over the circle, where $\theta$ is assumed to be constant, can be considered as equivalent to the filtering of a constant velocity translational vector field. Thus, the coefficients derived under this model can be used for the filtering of the vector magnitude of such a translational vector field, provided that the vector direction $\theta$ is known.

Figure 1. Constant velocity rotational motion field.

The output of an L-filter of length $N$ operating on the squared magnitude sequence is:

$$y_i = \sum_{j=1}^{N} a_j u_{i-j} = a^T R = a^T (r^2 e + N)$$

(6)
The second approach was to consider $\theta$ as being uniformly distributed around the circle. Then the filter coefficients will be the same for every filtering position around the circle. In this case, another assumption concerning the joint pdf of the three random variables $n_x, n_y, \theta$ has also to be made. The following model has been adopted:

$$f_{n_x, n_y, \theta}(n_x, n_y, \theta) = f_{n_x, n_y}(n_x, n_y) f_\theta(\theta)$$  \hspace{1cm} (20)

where

$$f_\theta(\theta) = \frac{1}{2\pi}, \quad 0 \leq \theta < 2\pi$$  \hspace{1cm} (21)

This is a realistic model, since the noise components $n_x, n_y$ do not have any dependency from the angle $\theta$ that defines the position on the circle. Again the evaluation of $f_\theta$ is done in two steps. We use an auxiliary random variable $h = n_x$, evaluate the joint pdf of $\nu, h, \theta$ and integrate this expression with respect to $h, \theta$ to obtain $f_\nu(\nu)$:

$$f_\nu(\nu) = \int_{0}^{2\pi} \int_{0}^{\infty} f_{n_x, n_y, \theta}(h, n_y, \theta) d\nu dh d\theta$$

where

$$f_{n_x, n_y, \theta}(h, n_y, \theta) = \frac{1}{\sqrt{2\pi \sigma_x \rho \sigma_y}} \exp\left(-\frac{h^2}{2\sigma_x^2} - \frac{(n_y - \mu_y)^2}{2\sigma_y^2} - \frac{\rho h (n_y - \mu_y)}{\sqrt{\sigma_x^2 \sigma_y^2}} \right)$$

Using these pdfs we can calculate the correlation matrix $R_N$ and the mean vector $\mu_N$ whose elements appear in (11),(14),(15) and obtain the optimum unconstrained, unbiased or location invariant L-filter coefficients.

A practical consideration about the L-filters that have been presented is that the constant vector magnitude $r$ must be known. In fact $r$ is not known, actually this is what we are trying to estimate, and, therefore, we must use an initial estimate. For example, we can filter the vector field with a 2-D mean or marginal median filter, obtain an estimate for $r$ and then use it to design an L-filter that will give a better estimate for the vector magnitude.

### 3. Experimental results

The performance of the L-filters that have been designed in the previous section has been evaluated using simulations and their noise suppression ability has been compared to that of other filters. Three sets of experiments have been conducted to test the performance of the L-filters designed both for constant direction angle $\theta$, (i.e. for filters that can be used for filtering of rotational vector fields in a small arc over a circle around the center of rotation) and for uniformly distributed $\theta$. In the first set of experiments, samples from a constant signal $r = 1$, $\theta = 0.5$ rad or $s_x = 0.877$, $s_y = 0.479$ in cartesian coordinates were corrupted by additive zero mean bivariate Gaussian noise

$$N(0, 0, \sigma_x, \sigma_y, \rho), \quad \sigma_x = \sigma_y = 1, \quad \rho = 0.5$$  \hspace{1cm} (23)

acting on the $x$, $y$ components of the signal. L-filters of length $N = 5, 7$ were designed for this kind of noise, assuming $r, \theta$ to be known, and their performance has been compared to the performance of the following single channel filters performing on the squared vector magnitude: arithmetic mean, median and minimum filter. The performance of two-channel arithmetic mean and marginal median filters performing on the components $u_x, u_y$ of the noisy signal has also been evaluated. Since our final aim was to filter the vector magnitude $r$, the vector output of the two-channel filters was transformed to polar coordinates. The quantitative performance criterion that was used for the comparisons was the Noise Reduction Index

$$NRI = 10 \log \frac{\sum (y_n - r^2)^2}{\sum ((w_n - r^2)^2)}$$  \hspace{1cm} (24)

Comparative results are presented in Table 1. It can be seen that the unconstrained L-filter has the best performance among all filters, the biased L-filter being the second best. Both filters perform far more better than the rest. It is interesting to note that the minimum filter has the best performance among the other filters. The superiority of the L-filters is further illustrated in Figures 2-5. Figure 2 shows the vector magnitude of the noisy vector sequence described in the previous experiment. Figures 3, 4 and 5 show the same sequence filtered with minimum filter, arithmetic mean filter operating on $u_x$, $u_y$ components and unbiased L-filter respectively. All filters were of length $N = 5$. It is obvious that L-filter is much more better than the two other filters.

In the second set of experiments we designed and tested L-filters under the constant direction angle $\theta$ assumption for the suppression of bivariate Gaussian noise contaminated with uniform noise:

$$\epsilon N(0, 0, \sigma_x, \sigma_y, \rho) + (1 - \epsilon) U([0, 1.5], [0, 1.5]), \quad \sigma_x = \sigma_y = 1, \quad \rho = 0.5, \quad \epsilon = 0.8$$  \hspace{1cm} (25)

where $U([0, 1.5], [0, 1.5])$ denotes the pdf of uniformly distributed outliers in the domain $[0, 1.5] \times [0, 1.5]$. This type of noise was used to corrupt the same constant signal that was described in the previous simulation. Results can be seen in Table 2. The superiority of the L-filters is obvious.

The last set of experiments aimed at testing the performance of the L-filters designed under the hypothesis of $\theta$ being uniformly distributed in the interval $[0, 2\pi]$. The coefficients found under this approach can be used to perform filtering of the vector magnitude on the entire circle. In order to test the validity of this hypothesis we used a vector sequence whose vectors were of constant magnitude $r = 1$ while $\theta$ changed from 0 to $2\pi$ as we moved from the first to the last sequence point, i.e. vectors corresponding to a circular path around the center of rotation in a rotational vector field. This signal was corrupted by bivariate Gaussian noise (23). Results are tabulated in Table 3. Once again L-filters are superior than the rest of the filters and their performance is similar to the performance of the filters designed under the constant $\theta$ hypothesis. This can be considered as a proof that the proposed model fits very well the real situation.

<table>
<thead>
<tr>
<th>Filter</th>
<th>$N=3$</th>
<th>$N=5$</th>
<th>$N=7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unconstrained L</td>
<td>-17.61</td>
<td>-18.67</td>
<td>-19.40</td>
</tr>
<tr>
<td>Unbiased L</td>
<td>-15.21</td>
<td>-18.25</td>
<td>-18.45</td>
</tr>
<tr>
<td>Location Invariant L</td>
<td>-12.04</td>
<td>-15.68</td>
<td>-17.78</td>
</tr>
<tr>
<td>Minimum on magnitude</td>
<td>-12.05</td>
<td>-14.55</td>
<td>-14.37</td>
</tr>
<tr>
<td>Mean on magnitude</td>
<td>-2.61</td>
<td>-3.85</td>
<td>-4.33</td>
</tr>
<tr>
<td>Median on magnitude</td>
<td>-4.27</td>
<td>-5.39</td>
<td>-7.84</td>
</tr>
<tr>
<td>Marginal median</td>
<td>-5.62</td>
<td>-8.63</td>
<td>-10.87</td>
</tr>
<tr>
<td>Mean</td>
<td>-7.26</td>
<td>-10.69</td>
<td>-11.93</td>
</tr>
</tbody>
</table>

Table 1. Performance of the various filters in the filtering of bivariate Gaussian noise from a vector sequence of constant magnitude and direction.

### 4. Conclusions

Optimal L-filters for the filtering of vector magnitude in certain types of vector fields have been proposed. Unconstrained as well as unbiased and location invariant L-filters have been designed. Experimental results verify the superiority of L-filters compared to other already known filters.
Table 2. Performance of the various filters in the filtering of bivariate Gaussian noise contaminated with uniformly distributed outliers from a vector sequence of constant magnitude and direction.

<table>
<thead>
<tr>
<th>Filter</th>
<th>N=3</th>
<th>N=5</th>
<th>N=7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unconstrained L</td>
<td>-18.12</td>
<td>-15.93</td>
<td>-20.63</td>
</tr>
<tr>
<td>Unbiased L</td>
<td>-16.76</td>
<td>-15.60</td>
<td>-19.92</td>
</tr>
<tr>
<td>Location Invariant L</td>
<td>-10.51</td>
<td>-14.62</td>
<td>-16.97</td>
</tr>
<tr>
<td>Minimum on magnitude</td>
<td>-10.29</td>
<td>-13.94</td>
<td>-14.51</td>
</tr>
<tr>
<td>Mean on magnitude</td>
<td>-2.43</td>
<td>-3.14</td>
<td>-3.49</td>
</tr>
<tr>
<td>Median on magnitude</td>
<td>-3.28</td>
<td>-4.45</td>
<td>-5.22</td>
</tr>
<tr>
<td>Marginal median</td>
<td>-4.21</td>
<td>-6.12</td>
<td>-7.28</td>
</tr>
<tr>
<td>Mean</td>
<td>-5.68</td>
<td>-8.03</td>
<td>-9.42</td>
</tr>
</tbody>
</table>

Table 3. Performance of the various filters in the filtering of bivariate Gaussian noise from a vector sequence having constant magnitude and direction that varies uniformly from 0 rad to 2π rad.

<table>
<thead>
<tr>
<th>Filter</th>
<th>N=3</th>
<th>N=5</th>
<th>N=7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unconstrained L</td>
<td>-17.24</td>
<td>-19.08</td>
<td>-20.39</td>
</tr>
<tr>
<td>Unbiased L</td>
<td>-15.63</td>
<td>-17.95</td>
<td>-19.48</td>
</tr>
<tr>
<td>Location Invariant L</td>
<td>-11.47</td>
<td>-15.70</td>
<td>-17.87</td>
</tr>
<tr>
<td>Minimum on magnitude</td>
<td>-11.48</td>
<td>-14.54</td>
<td>-14.59</td>
</tr>
<tr>
<td>Mean on magnitude</td>
<td>-2.73</td>
<td>-3.58</td>
<td>-4.04</td>
</tr>
<tr>
<td>Median on magnitude</td>
<td>-4.03</td>
<td>-5.80</td>
<td>-6.83</td>
</tr>
<tr>
<td>Marginal median</td>
<td>-6.02</td>
<td>-9.68</td>
<td>-12.69</td>
</tr>
<tr>
<td>Mean</td>
<td>-7.94</td>
<td>-11.76</td>
<td>14.52</td>
</tr>
</tbody>
</table>

References


Angular Filtering in the Context of Frequency Estimates

T. LOUPAS

Ultrasonics Laboratory, Division of Radiophysics,
Commonwealth Scientific, Industrial and Research Organisation (CSIRO),
126 Greville Street, Chatswood NSW 2067, Sydney, Australia
E-mail: iloupas@at.rp.csiro.au

Abstract. This paper examines the problem of filtering frequency estimates derived from discrete-time signals. The concept of wrapped probability distributions is used to address the fact that such frequency estimates are wrapped around a range whose width is equal to the sampling frequency. Since this is equivalent to wrapping the corresponding angular frequency values around the circumference of a unit-circle, a number of angular filtering schemes based on mean and median operations are introduced, and their output probability distributions are analysed. Also, simulations are presented to quantify the performance of the angular filters as well as their linear (i.e., operating off data distributed along the line of real numbers) counterparts, for inputs belonging to wrapped Gaussian, Laplacian and Cauchy distributions. The results show that conventional linear filters become increasingly inappropriate for processing wrapped data as the input distribution's mode moves away from the centre of the wraparound range. They also suggest that, in analogy with the theory of order-statistics for data along the line, the angular median outperforms the angular mean for long-tailed wrapped distributions.

1. Introduction

Frequency estimation represents one of the most important digital signal processing operations, with a variety of applications in a number of fields such as communications, radar, sonar, geophysics and medicine [1]. The statistical characteristics of a frequency-estimate sequence may vary considerably, depending on the nature of input data and form of the algorithm employed. However, due to the sampling theorem, all estimators applied to discrete signals share the same fundamental property: their output is restricted to the range \([-F_s/2, F_s/2]\), where \(F_s\) denotes the sampling frequency. As a consequence, an individual estimate \(F_x\) which is either forced by noise or genuinely lies outside this range, is wrapped around to produce an aliased value \(\Phi_x\). Using normalised notation \((f_x = F_x / F_s, \ \phi_x = \Phi_x / F_s)\), wrapping around the primary range \([-0.5, 0.5]\) can be described as a modulo-1 transformation

\[
\phi_x = W(f_x) = [(f_x + 0.5) \mod 1] - 0.5 \quad (1)
\]

Multiplying eq.(1) by \(2\pi\) is equivalent to wrapping the angular frequency \(2\pi f_x\) around the circumference of a unit-radius circle. Let us assume that a random variable \(f\) distributed along the line of real numbers \((-\infty < f < +\infty)\) has a probability density function (pdf) \(g_f(f)\) and cumulative density function (cdf) \(G_f(f)\). By making use of the properties of circular data [2], it can be shown that the random variable \(\phi\) obtained from eq.(1) obeys the wrapped pdf and cdf given by

\[
g_{\phi}(\phi) = \sum_{k=-\infty}^{\infty} g_f(\phi + k) \quad (2)
\]

\[
G_{\phi}(\phi) = \sum_{k=-\infty}^{\infty} \left\{ G_f(\phi + k) - G_f(k - 0.5) \right\} \quad (3)
\]

where \(k\) assumes integer values and \(\phi\) belongs to the primary range \([-0.5, 0.5]\). For random variables on the line which are wrapped around an arbitrary range \([\Phi_L, 1 + \Phi_H]\), eq.(2) remains unchanged, while the term 0.5 must be replaced by \(\Phi_H\) in eq.(3).

Wrapped distributions have important implications for the choice of location or smoothing estimators which are applied to sequences of frequency values derived from discrete data. Pdfs of the form defined by eq.(2) exhibit heavy tails and are in general asymmetric, unless their mode coincides with the midpoint of the wraparound range. As result of this asymmetry, estimators which may be optimum for a given probability distribution on the line of real-numbers, perform poorly when applied to sequences obeying its wrapped version. The pdf asymmetry can be removed when the circular nature of wrapped random variables is taken into account, by means of the transformation

\[
\phi \rightarrow \left\{ x, y \right\} = \left\{ \text{Re}{(e^{2\pi x i})}, \text{Im}{(e^{2\pi x i})} \right\} \quad (4)
\]

Eq.(4) maps individual values \(\phi\) onto vectors which have unit-length and form angles of \(2\pi x\) radians with respect to the x-axis. Figures 1 and 2 illustrate the effect of the above transformation on a sequence of wrapped frequency estimates.

![Figure 1. Mean Doppler frequency output of a time-domain estimator and its histogram. The input data had a constant mean frequency of 0.4 and were heavily corrupted by Gaussian noise.](image-url)
Figure 2. Mapping the data points of Figure 1 onto a circle and the corresponding circular histogram.

2. Angular Filters

Angular filters rely on circular operations to handle samples that may be apart in a numerical, but close in a wraparound sense. From this point of view, they are closely related to filters operating on circularly distributed data. A comprehensive treatment of the theory of circular random variables can be found in [2], while an application of circular location estimators in colour image processing and an overview of their properties has been recently reported in [3]. In the context of the wrapped frequency values defined by eq.(1), the most straightforward location estimator (or, equivalently, smoothing filter) is the angular mean. Expressed as a running-window filtering scheme, the angular mean is derived from the mean angle of the unit vectors in eq.(4)

$$\text{ang\_mean}(i) = \frac{1}{2\pi} \arg \left\{ \sum_{n=-N}^{N} e^{j2\pi \phi_{\text{norm}}(n)} \right\}$$

(5)

where \(2N+1\) represents the filter size. For wrapped frequency sequences of varying nature, such as those obtained from frequency-modulated signals, appropriate smoothing coefficients can be introduced in the sum of eq.(5), resulting in a "modulo-convolution" operation [4].

Alternatively, if preserving sharp transitions in the wrapped frequency sequence is of primary interest, or if the original non-wrapped samples obey a long-tailed distribution, median-based operations can be used. In contrast to the median of a sequence distributed on the line, which is unique, a number of schemes can be used for wrapped samples. A marginal median filter can be derived according to

$$\text{marg\_med}(i) = \frac{1}{2\pi} \arg \left\{ \text{median}[\text{Im}[e^{j2\pi \phi_{\text{norm}}(n)}]] \right\}$$

with \(n = -N, \ldots, N\). This filter is essentially identical to the one used in multi-channel image processing [5], although in this case only the angle of the output vector is of interest. The marginal median has the advantage that it is relatively straightforward to compute. However, since the components of the unit-vectors are ordered separately along the x- and y-axis, the marginal median’s output does not necessarily coincide with one of the input samples.

If such an attribute is regarded as essential, ordering must be carried out in an angular manner. For a population of wrapped frequency samples derived from eq.(1), an angular median can be defined as the value \(\phi_k\) satisfying

$$\int_{\phi_0}^{\phi_0+0.5} g_{\phi}(\phi) d\phi = \int_{\phi_k}^{\phi_k+0.5} g_{\phi}(\phi) d\phi$$

(7)

The equivalent operation for the angular frequency population is to find a value \(2\pi\phi_0\) so that the diameter defined by \(2\pi\phi_0\) and \(2\pi\phi_0 \pm \pi\) divides the circle in two segments containing the same number of samples. Note that the integral limits in eq.(7) are allowed to fall outside the primary range \([-0.5, 0.5]\), because of the periodicity of the wrapped pdf in eq.(2). Also, note that, if \(\phi_k\) satisfies eq.(7), then the same applies for \(\phi_k \pm 0.5\). This ambiguity can be removed by selecting \(\phi_k\) only if it is associated with a higher probability than the other two candidates, i.e.

$$g_{\phi}(\phi_k) > g_{\phi}(\phi_k - 0.5) = g_{\phi}(\phi_k + 0.5)$$

(8)

The combined definition of eq.(7)-(8) results in a unique population value, provided that the wrapped distribution is unimodal. The equivalent condition for a finite-size sequence of population samples is that its histogram exhibits a global maximum, a condition that does not hold in general. However, in direct analogy to the circular case [2], the population angular median can be shown to minimise the sum of wrapped distances between itself and the population members. Based on this property, a sample angular median filter can be uniquely determined from

$$\text{ang\_med}(i) = \phi_k \quad \text{where} \quad \phi_k \quad \text{satisfies}$$

$$\sum_{n=N}^{N} \left\{ 0.5 - 0.5 \cdot \left[ (\phi_{i+n} - \phi_k) \mod 1 \right] \right\} \rightarrow \min$$

(9)

with \(k = i - j, \ldots, i + N\). The sample angular median definition is equivalent to the “arc distance median” [3] and “vector directional” [6] filters for colour image processing. It is also closely related, but not identical, to the “vector median” filter introduced in [7].

Because of the various nonlinearities in the definition of the angular mean and marginal median filters, their distributions tend to be very complex and, for most input distributions, can only be obtained in either a numerical manner or in terms of approximations. An approximate expression for the pdf of the angular mean has been given in [4]. The pdf of the marginal median can be found by means of the following procedure. The starting point is the joint pdf and pdf of the x- and y-components associated with the unit-vectors in eq.(4). The joint pdf can be derived in terms of the probability masses of the random variable \(2\pi \phi\), which are all concentrated on the circumference of the unit-circle. For example, given a point \((x_0, y_0)\) which belongs to the first circle quadrant, the joint pdf can be written explicitly as

$$G_{x,y}(x_0, y_0) = G_{\phi}(0.5) - G_{\phi}\left((2\pi)^{-1}(\pi - \arcsin(2\pi y_0))\right) + G_{\phi}\left((2\pi)^{-1}(-\arccos(2\pi y_0))\right) - G_{\phi}(-0.5)$$

(10)
From the theory of random-variable functions [8], the joint pdf can be expressed as

\[ g_{x,y}(x,y) = (2\pi)^{-1} \sqrt{1 - x^2 - y^2} g_{\phi}(\arctan \frac{y}{x}) \]

(11)

The joint pdf \( g_{x,y}(x,y) \) of the marginally-ordered median components, which has been derived in [5], can now be evaluated in terms of eq.(10)-(11). This is subsequently transformed, again by using the theory of random-variable functions [8], to the pdf of the marginal median filter. For example, if \(-0.25 \leq \phi \leq 0.25\),

\[ g_{\text{marg, med}}(\phi) = \frac{\int_{-\infty}^{\infty} [g_{x,y, \text{med}}(y \tan 2\pi \phi, y)dy]}{(2\pi (1 + \tan^2 2\pi \phi))^{-1}} \]

(12)

In contrast to the complexity of the marginal-ordering case, a simple expression can be obtained for the pdf of the angular median by approximating the definition of eq.(7) as an offset linear-ordering operation. More specifically, for a distribution of wrapped values described by eq.(2)-(3) and having a true median equal to \( \phi_0 \), this operation involves wrapping the distribution members around the range \([-\phi_0 - 0.5, \phi_0 + 0.5]\), finding the median value by means of linear ordering in this range, and re-wrapping the median value back to the primary range \([-0.5, 0.5]\). It is worth noting that the procedure outlined above is equivalent to ordering the angular frequencies \(2\pi \phi\) on the circle, with the zero direction corresponding to the anti-diametric point of the median. For a sample size of \(2N+1\), it is straightforward to show that the pdf of the offset linear-ordering median is equal to

\[ g_{\text{off, med}}(\phi) = \frac{(2N+1)!}{N!N!} g_\phi(\phi) \left( p(\phi) \right)^N \left( 1 - p(\phi) \right)^N \]

(13)

where

\[ p(\phi) = \begin{cases} G_\phi(\phi) - G_\phi(\phi_0 - 0.5) & 0.5 < \phi \leq 0.5 \\ G_\phi(\phi_0 - 0.5) - G_\phi(\phi) & 0 \leq \phi \leq \phi_0 - 0.5 \end{cases} \]

(14)

3. Results

Figure 3 plots the approximate pdf curves, calculated from eq.(13)-(14) assuming a wrapped version of a Laplacian distribution with \( \mu = 0.4, \sigma = 0.2 \), as well as simulation-derived output histograms of the angular median, for three filter sizes. The filters were applied to an input sequence which was obtained by generating 128K samples from the same distribution as that used for evaluating the approximate pdf curves, followed by a wrapping operation according to eq.(1). Note that the histogram points have been multiplied by a common scaling factor, chosen so that the histogram and approximate pdf values at 0.4 were equal for a filter size of 1 sample, to facilitate comparisons. Overall, this figure indicates a reasonable agreement between the simulation-derived distributions and their theoretical approximations. Similar comparisons covering a variety of input distributions have shown that the agreement becomes almost ideal as the value of the input pdf at its anti-mode approaches zero.

Figure 4, which plots the theoretically expected pdfs of the conventional median for the same input distribution and filter sizes as those used in Figure 3, highlights the inappropriateness of processing wrapped data with techniques of a linear (input on the line of real numbers) nature. Although the linear median is the maximum-likelihood estimator of location for the Laplacian distribution, the curves of Figure 4 predict a poor performance for the wrapped version of the Laplacian pdf. Note, in particular, the introduction of significant bias which is demonstrated by the shift of the pdf peak towards zero as the filter size increases. The bias is primarily determined by the degree of asymmetry present in the input pdf. Consequently, it is a function of the spread of the input pdf as well as the distance between the mode of the input pdf and the centre of the wraparound range. It is worth noting that at the limiting case where this distance is zero, the linear median coincides with the offset linear median.

![Figure 3. Pdfs of the offset linear median (solid curves) and simulation-derived histogram values of the angular median (points), for a wrapped Laplacian input distribution and filter sizes equal to 1 (i.e., no filtering), 5 and 9 samples.](image)

![Figure 4. Theoretically expected pdfs of the linear median for the same input distribution and filter sizes as those used in Figure 3.](image)
CRMSE = \left( \frac{1}{M} \sum_{m=1}^{M} W^2(\hat{\phi}_m - \phi_{mode}) \right)^{1/2} \tag{15}

where \( \phi_{mode} \) represents the mode of the input distribution, \( \hat{\phi}_m \) denotes the output of a given filter for the \( m \)-th repetition, and \( W(\hat{\phi}) \) is the wraparound transformation defined in eq. (1). Overall, the mode varied from 0 to 0.475 in steps of 0.025, and \( M = 16 \)K repetitions were performed for each mode value.

Figure 6. Circular root-mean-square error of an original input sequence obeying the wrapped Gaussian pdf (mode varying from 0 to 0.475) as well as the outputs of the mean, median, angular mean and angular median 9-sample filters.

Figure 6. As in Figure 5, for a wrapped Cauchy pdf.

The graphs of Figures 5 and 6 demonstrate that the circular error introduced by angular filters is independent of the exact location of the wrapped input distribution. On the other hand, the error introduced by linear filters is affected by the amount of wraparound present in the input distribution and, consequently, increases abruptly near the end of the normalised frequency axis. The CRMSE graphs provide a clear demonstration of the superiority of angular filters over their linear counterparts. Similar conclusions were reached when a linear measure of performance (the commonly used root-mean-square error) was adopted. The CRMSE graphs also show that the angular mean is more suited to short-tailed wrapped distributions whereas the angular median is better for long-tailed distributions, an observation which has been well-documented for data distributed along the line of real numbers [9]. Finally, it should be mentioned that the CRMSE graphs of the marginal median were found to be almost identical to those obtained from the angular median. Therefore, since the angular median tends to be considerably more demanding from a computational point of view and no performance differences between the two filters were documented, it appears that the marginal median is preferable, at least for the type of input signals considered here.

4. Conclusions

This paper has examined issues related to the wraparound nature of frequency estimates which have been derived from sampled signals. In particular, definitions for the cdf and pdf of wrapped random variables were given and a number of mean- and median-based estimators of location, which take explicitly into account the angular nature of wrapped data, were presented. The output distributions of these filters were briefly analysed, and an approximate approach for determining the pdf of the angular median was proposed and its validity tested by means of simulations. The performance of the angular filters, as well as their linear counterparts, was documented as a function of the input distribution's mode, for short- and long-tailed wrapped distributions. In summary, the results suggest that the angular nature of frequency estimates should be taken into account explicitly, to avoid the significant errors introduced by conventional linear filters. They also confirm that median-based angular estimators are more effective than the angular mean as the tails of the wrapped input distribution become more pronounced.

References

Introducing the Fuzzy Median Filter

Fabrizio RUSSO - Giovanni RAMPONI

Dipartimento di Eletrotecnica Elettronica Informatica, Universita' degli Studi di Trieste, Trieste I-34127 Italy, Tel/Fax: +39 (040) 676 3411 / 676 3460, E-mail: rusfab@univ.trieste.it ramponi@univ.trieste.it

Abstract. It is demonstrated in this paper that members of a class of recently introduced fuzzy filters (FIRE operators) can behave exactly as median operators if their parameters are suitably chosen. This result of course does not merely lead to a different realization of a median filter (which would be uneconomical) but rather permits to observe that FIRE filters are a wide and comprehensive framework which includes one of the most widespread nonlinear operators. The above referred equivalence is formally demonstrated in a simple case, but its extension to a more general predicament is in principle straightforward. The effects of some changes in the FIRE operator parameters are also shortly presented.

1. Introduction

Fuzzy filters constituted by single-ELSE rulebases (FIRE filters) are a novel class of nonlinear operators which use approximate reasoning for solving image processing problems. Originally proposed in [1], these operators have demonstrated to be able to perform image smoothing, sharpening, edge detection and to outperform classical operators in some very critical issues, such as the enhancement of noisy and blurred images [2]. Two key elements constitute the basis of the performances of these nonlinear operators: the possibility of easily defining processing strategies in form of rules and the capability of successfully realize them by means of an appropriate inference mechanism.

The operator structure, based on a specifically developed then-else inference mechanism which also allows fuzzy rules to be grouped into many subrulebases, has been described in [3]. The nonlinear behaviour of the operators has been analyzed for a particular case in [4]. In this work the latest member of this family is presented: a fuzzy operator which aims at restoring images degraded by "salt and pepper" noise. As it will be shown, the choice of an appropriate rulebase and of suitable parameters make the operator behaviour identical to that of a median filter. This observation opens up new vistas in the study of the median filter: the new family of fuzzy filters lends itself as a wide and comprehensive framework including one of the best-known and universally diffused operators. This framework, apart from its own already established properties and advantages, could be used for designing and evaluating novel extensions of the median operator.

This paper is organized as follows: Sect.2 describes the operator structure, while more details about the inference mechanism are presented in Sect.3. Sect.4 is devoted to demonstrating the equivalence of a simple FIRE filter and a 1-D, three-point median filter. The effects produced by different parameter choices are briefly indicated in Sect.5.

2. The operator structure

Let $x_0$ be the pixel to be processed and $x_1, x_2, x_3$ and $x_4$ the neighbouring pixels at its top, left, bottom and right side, respectively (see Fig.1).

![Figure 1](image-url)

The corresponding output value will be $y_0=x_0+dx$, where $dx$ is the correcting term yielded by the fuzzy operator. In order to obtain a correcting term $dx$ capable of cancelling the effect of a noisy (black or white) value for $x_0$, we can use a simple set of fuzzy rules: for example, a very simple fuzzy algorithm could be expressed in a plain linguistic form by the following statements:

- if $x_0$ is much darker than $x_1, x_2, x_3$ and $x_4$ then make it much brighter;
- if $x_0$ is much brighter than $x_1, x_2, x_3$ and $x_4$ then make it much darker;
- else leave it as it is.

A slightly more complex set of rules can be adopted in order to make the filter more effective. It should be observed, indeed, that the above mentioned group of rules has no effects on any couple of adjacent noisy pixels: in order to overcome this limit we shall take into account
only groups of three pixels at a time. The set of rules becomes in this case:

- if \( x_0 \) is much darker (brighter) than \( x_1, x_3, \) and \( x_3 \)
  then make it much brighter (darker);
- if \( x_0 \) is much darker (brighter) than \( x_2, x_3, \) and \( x_4 \)
  then make it much brighter (darker);
- if \( x_0 \) is much darker (brighter) than \( x_3, x_4, \) and \( x_1 \)
  then make it much brighter (darker);
- if \( x_0 \) is much darker (brighter) than \( x_4, x_1, \) and \( x_2 \)
  then make it much brighter (darker);
- else leave it as it is.

It is easy to verify that these rules permit the elimination of couples of horizontally- or vertically-adjacent noise pixels.

Once we have defined the basic principle of the processing strategy to be adopted, we can easily translate it into a more formal design. For this purpose we shall define fuzzy variables, sets and rules as follows. The variables which are input to the fuzzy processor will be the luminance differences \( u_i=x_j-x_0 \) (i=1,2,3,4); the output variable will be the quantity \( dx \) mentioned above. Let us also define the three fuzzy sets negative (N), zero (Z) and positive (P) as sketched in Fig.2.

**Figure 2. Fuzzy sets.**

The rulebase we adopt for the operator is described below:

- IF \( u_1 \) is P AND \( u_2 \) is P AND \( u_3 \) is P THEN \( dx \) is P
- IF \( u_2 \) is P AND \( u_3 \) is P AND \( u_4 \) is P THEN \( dx \) is P
- IF \( u_3 \) is P AND \( u_4 \) is P AND \( u_1 \) is P THEN \( dx \) is P
- IF \( u_4 \) is P AND \( u_1 \) is P AND \( u_2 \) is P THEN \( dx \) is P
- IF \( u_1 \) is N AND \( u_2 \) is N AND \( u_3 \) is N THEN \( dx \) is N
- IF \( u_2 \) is N AND \( u_3 \) is N AND \( u_4 \) is N THEN \( dx \) is N
- IF \( u_3 \) is N AND \( u_4 \) is N AND \( u_1 \) is N THEN \( dx \) is N
- IF \( u_4 \) is N AND \( u_1 \) is N AND \( u_2 \) is N THEN \( dx \) is N

**ELSE** \( dx \) is Z

The rulebase includes 8 THEN-rules and one ELSE-rule. The THEN-rules are grouped into two sub-rulebases having \( P \) and \( N \) as consequent sets, respectively; thus, each sub-rulebase includes four THEN-rules. Each THEN-rule is composed of three antecedent clauses representing the input conditions and one consequent clause which defines the corresponding output action. Only the consequent action is specified in the ELSE-rule, because (as we shall see) its degree of activation directly depends on that of the THEN-rules. It should be noticed that the second, third and fourth rule of each sub-rulebase are generated by \( 90^\circ \) degree rotations of the first rule. A graphical representation of the rulebase is shown in Fig.3

**Figure 3. Graphical representation of the fuzzy rulebase.**

3. The fuzzy inference mechanism

The inference mechanism of the proposed operator is briefly described below (for a detailed description of the general inference mechanism which the fuzzy operators are based on, see [3,4]).

The fuzzy sets are described by triangular-shaped membership functions defined in the domain of the signed integers \( U \). As it is well known [5], for a given fuzzy set \( S \) and for each value \( u \in U \) the corresponding membership function \( m_p(u) \) returns the degree of membership of \( u \) to that fuzzy set: this degree is a real number in the interval \([0,1]\), where \( 1 \) denotes full membership and \( 0 \) no membership.

For a given set of input values \( u_1, u_2, u_3, u_4 \) the degree of satisfaction \( \lambda_p(i) \) of the \( i \)-th rule of the first sub-rulebase is given by:

\[
\lambda_p(i) = \text{MIN} ( m_p(u_1), m_p(u_2), m_p(u_3), m_p(u_4) ) \quad i=1,2,3,4 ;
\]

\( \oplus = \text{modulo 4 addition} \)

The overall degree of satisfaction \( \lambda_p \) of the first sub-rulebase is evaluated as:

\[
\lambda_p = \text{MAX} ( \lambda_p(i) : i=1,2,3,4 )
\]

Relationships very similar to (2) and (3) are to be adopted for evaluating the overall degree of satisfaction \( \lambda_N \) of the second sub-rulebase. The degree of satisfaction \( \lambda_Z \) of the ELSE-rule is given by:

\[
\lambda_Z = \text{MIN} ( (1-\lambda_p), (1-\lambda_N) )
\]

The output term \( dx \) is evaluated by combining the different effects of the two sub-rulebases and of the ELSE-rule [3]:

964
\[ dx = \frac{(cp + cN + pN + cz + pZ)}{(p + pN + pZ)} \]  
(5)

where \( pp, pN \) and \( pZ \) are defined here by the following relationships:

\[ pp = \lambda p wp; \quad pN = \lambda N wN; \quad pZ = \lambda Z wZ \]  
(6)

(From a fuzzy point of view it can be observed that this choice corresponds to the adoption here of a product instead of a minimum correlation inference mechanism. See [5] for an outstanding introduction to the fuzzy inference methods).

In order to illustrate the performance of the proposed operator, we have degraded the well known test image Lena by using impulsive noise with probability 10% (Fig.6). The result of the application of the filter is shown in Fig.7: this result is identical to the one obtained by using a 5-point median filter. In this experiment the \( N, Z \) and \( P \) set centers \( c_N, c_z, c_p \) are located at the abscissa values \(-c, 0, c\) respectively; the half-width of the nonzero parts of the same sets is \( w_N = w_Z = w_P = c \). If the input pixel values \( x_i \) are positive integers in the range \([0,255]\), we select \( c = 255 \).

4. Median operators as a subset of fuzzy operators

We shall see in the following how a median operator is a particular case of a rule-based fuzzy filter. For the sake of simplicity, we consider an elementary three-point, 1-D operator. To this purpose, let \( x_0 \) be the sample to be processed at a given time \( n \), and let \( x_1 \) and \( x_2 \) be its left- and right-side neighbours, respectively. The input variables are the differences \( u_1 = x_1 - x_0, u_2 = x_2 - x_0 \). The fuzzy sets are again those of Fig.2. The rulebase (1) becomes in this case:

\[
\begin{align*}
\text{IF} \ (u_1 \text{ is } P) \ & \text{AND} \ (u_2 \text{ is } P) \ \text{THEN} \ (dx \text{ is } P) \\
\text{IF} \ (u_1 \text{ is } N) \ & \text{AND} \ (u_2 \text{ is } N) \ \text{THEN} \ (dx \text{ is } N) \\
\text{ELSE} \ (dx \text{ is } Z)
\end{align*}
\]  
(7)

We want to demonstrate that \( y_0 = \text{MEDIAN}(x_1, x_0, x_2) \).

Let us maintain the same parameter choices as in the previous section: \( c_N = -c, \ c_Z = 0, \ c_P = c, \ w_N = w_Z = w_P = c \), \( c = 255 \); \( -255 \leq u_1, u_2 \leq 255 \). Formally, for two given values of \( u_1 \) and \( u_2 \), the overall levels of satisfaction of the two THEN-rules, respectively are:

\[
\begin{align*}
\lambda_P &= \begin{cases} 
\text{MIN} \ (u_1/c, u_2/c) & \text{if } u_1 > 0, u_2 > 0 \\
0 & \text{otherwise}
\end{cases} \\
\lambda_N &= \begin{cases} 
\text{MIN} \ (-u_1/c, -u_2/c) & \text{if } u_1 < 0, u_2 < 0 \\
0 & \text{otherwise}
\end{cases}
\end{align*}
\]

Relationship (5) yields in this case:

\[ dx = c (\lambda_P - \lambda_N) / (\lambda_P + \lambda_N + \lambda_Z) \]  
(8)

Notice that only three cases are possible:

1) \( u_1 > 0, u_2 > 0 \) \( \Rightarrow \lambda_P \neq 0, \lambda_N = 0 \) \( \Rightarrow dx = c \lambda_P \)

2) \( u_1 < 0, u_2 < 0 \) \( \Rightarrow \lambda_N \neq 0, \lambda_P = 0 \) \( \Rightarrow dx = -c \lambda_N \)

3) otherwise \( \Rightarrow \lambda_P = \lambda_N = 0 \) \( \Rightarrow dx = 0 \)

Let us examine the case 1). Substituting in the previous relations we obtain:

\[ dx = c \text{MIN}(u_1/c, u_2/c) = \text{MIN}(u_1, u_2) = \text{MIN}(x_1, x_2) - x_0 \]

but in this case we have \( u_1 > 0, u_2 > 0 \), which imply \( x_1 > x_0, x_2 > x_0 \); hence

\[ y_0 = x_0 + dx = \text{MIN}(x_1, x_2) = \text{MED}(x_1, x_0, x_2). \]

Similarly, for the case 2):

\[ dx = -c \text{MIN}(-u_1/c, -u_2/c) = -\text{MIN}(-u_1, -u_2) = -\text{MIN}(x_0 - x_1, x_0 - x_2) = \text{MAX}(x_1, x_2) - x_0 \]

we have now \( u_1 < 0, u_2 < 0 \), which imply \( x_1 < x_0, x_2 < x_0 \); hence

\[ y_0 = x_0 + dx = \text{MAX}(x_1, x_2) = \text{MED}(x_1, x_0, x_2). \]

Finally, in the case 3):

\[ y_0 = x_0 = \text{MED}(x_1, x_0, x_2). \]

Indeed, in this case we have

\[ u_1 \geq 0, u_2 \leq 0 \] \( \text{which imply} \ x_1 \geq x_0 \geq x_2 \) \( \text{or} \ u_1 \leq 0, u_2 \geq 0 \] \( \text{which imply} \ x_1 \leq x_0 \leq x_2 \).

q.e.d.

The demonstration above is limited, for reasons of space, to a 1-D case; on the other side, we can show that it is readily extended in order to establish the equivalence of the FIRE filter presented in Sect.2 to a five-point, 2-D median filter. This result is obtained by taking into account each of the various pixel configurations shown in Fig.3, ordering the involved input differences and observing that each subrulebase of the operator yields an activation strength \( \lambda \) which is a function of a single input difference. The global relation (8) can finally be used.

Further extensions to more complicated operators are relatively easy in principle but formally troublesome: they require the choice of two suitable subrulebases, formed in turn by a large number of rules each of which deals with a fraction of the input data.

5. The general framework

As we have seen in the previous section, the FIRE operator is equivalent to a median filter only for particular choices and conditions; more in general, we can obtain different behaviours by suitably changing the involved parameters or the rules.

5.1 Different parameters

Let us examine, for example, the effects of a simple change in the values of the center position \( c \) of the \( N \) and \( P \) sets and of their width \( w \). Notice that we are still considering a trivial operator, which does not take into account any 2-D geometrical relation among the data to be processed and is thus much less powerful than the operators used in the cited literature. As a consequence, the example below gives only a very simplified hint of the possible responses of the fuzzy operator.
can introduce a change in behaviour also by adding further rules, such as those shown in Fig.5.

\[ \text{Figure 5. Additional set of fuzzy rules.} \]

In this way we obtain an operator which is able to take into account a 9-pixel neighbourhood of the processed point, permitting the cancellation of a larger part of the noise pulses (see Fig.8). It must be observed that such an operator is no longer equivalent to a median filter. From a performance point of view, it can be located midway between a 5-point and a 9-point median filter; the latter, indeed, achieves a better noise cancellation but introduces relevant distortions on the image details.

Acknowledgements

This work was supported by the European Research Project ESPRIT BRA NAT 7130.

References


Knowing when we're wrong: The Run-Distribution Test

Andrew W. FITZGIBBON and Robert B. FISHER

Department of Artificial Intelligence, Edinburgh University
5 Forrest Hill, Edinburgh EH1 2QL. Email: andrew@ed.ac.uk

Abstract. In this paper, we are concerned with the problem of deciding whether a fitted model accurately describes the data to which it has been fitted. We have developed an effective method of testing the lack-of-fit of a parametric model to data, with applications to the computer vision problems of robust estimation, model selection, and curve and surface segmentation. The test looks at the distribution of the residual errors, rather than basing statistics directly on their values. It is essentially distribution independent, and requires no knowledge of the sensor noise level. We present results of experiments that compare the test with the standard \( \chi^2 \) statistic. The experiments are designed to represent typical computer vision tasks, namely feature tracking, robust fitting, and segmentation. We show that our test is superior to the chi-square in all cases but those where the noise level is perfectly known.

1 The Problem

It is very common in computer vision to wish to represent some large dataset in a concise way in order to extract geometric properties, attenuate noise, or simply to reduce the volume of data. In almost all cases, this is achieved by fitting an appropriate parametric model to the data set in the least squares sense. It is then vital to have some way of telling when the fit is wrong, and the model is not 'appropriate' to the data. Simple least squares techniques [7] assume the noise in the data to be strictly Gaussian of known variance, and then use the \( \chi^2 \) test to give an estimate of the probability that, under that assumption, the data fits the model. Robust estimators [5] approach the problem more directly, by effectively ignoring data points which do not fit the model. Robust models are, however, even more expensive to fit than unbiased nonlinear models, and do not help when the model is already fitted to the data, and simple verification is all that is needed. Our argument asserts that least squares is adequate for most purposes, until its assumptions are violated. Of course it is precisely these boundaries, at which the assumptions are violated, that are of most importance to the visual process. Hence, a quick and effective test which identifies such errors will allow a cheap estimator to be used on most of the signal, while the more expensive techniques are held in reserve until the cheaper methods fail.

2 Goodness-of-fit Testing

We denote the data points to which the model is to be fitted by \( \{x_i\}_{i=1}^{n} \) and the parameters of the model by \( \{a_i\}_{i=1}^{p} \). We also assume that we have a distance metric \( D(a, x) \) which measures the signed distance between a particular data point and the fitted model. The model fitting process is assumed to have found the value of a for which \( \varepsilon = \sum_{i=1}^{n} \phi(D(a, x_i)) \) is minimized. The function \( \phi(x) \) is an influence function, which for classical least squares is \( \phi(x) = x^2 \). We do not need to know the form of \( \phi \), simply that it must be symmetric or antisymmetric about \( x = 0 \). Having found the value \( a \), we can define the set of residuals \( R = D(a, x_i)_{i=1}^{n} \). The task of goodness-of-fit testing is to determine, based on the values of the residuals, whether it is likely that the model describes the data. Lack-of-fit statistics say whether the model is unlikely to describe the data*.

2.1 Chi-Square Test

Whatte [9] provides an accessible summary of the chi-square testing technique. The basic assumption is that each observed point \( X \) is the exact point corrupted by an isotropic zero-mean Gaussian noise process of variance \( \sigma^2 \). If \( \sigma^2 \) is known, the chi-square statistic \( \chi^2 = \sum_{i=1}^{n} \left( R_i / \sigma \right)^2 \) has a known distribution. The disadvantages of the \( \chi^2 \) test are well known: the Gaussian noise model has repeatedly proved unrealistic in computer vision and the noise variance is often difficult to know in general.

2.2 Median Absolute Deviation

The median absolute deviation (MAD) measure is not strictly a test, in the sense of providing a probability of error. However, because it is essentially the error metric used in robust estimators, it is interesting to see how its response compares with the RD test. The measure is simply the median of the absolute values of the residuals, and may be evaluated in about \( O(n \log \log n) \) time. To use this measure as a test of goodness of fit, we need an estimate of the noise level. For Gaussian distributed residuals with a standard deviation \( \sigma \), the median \( M \) of the absolute values of the residuals satisfies \( \text{erf}(M / \sqrt{2}) = 1/2 \). From this, we can calculate the expected value of \( M \) and threshold the MAD value accordingly.

*The distinction between lack of fit and goodness of fit is subtle and of great interest to statisticians, but we shall not make it here, treating the two terms as equivalent.
2.3 "RANSAC" Maximum Run Length Test:

The "RANSAC" system of Fischler and Bolles [2] is the most similar test reported in the vision literature. Their system considers the maximum run length (see below) observed for a set of residuals. In our experiments, we have found this measure to be noise sensitive.

3 Run-distribution Test

We now introduce our test, which we have called the run-distribution test. We describe the idea behind the test, the noise model which we assume, the actual test, and how it differs from similar tests in the literature.

The tests discussed above essentially extract one number from the set of residuals, and use that as a basis for discrimination. Instead we want to look at the set of residuals $R$, and decide whether that set is what we would expect, given data which is in concordance with both our parametric and noise models.

3.1 Noise model

We allow each point to be corrupted in each dimension by a scalar noise component sampled from a symmetric zero-median process plus an outlier process. Note that this is a very wide range of distributions, trivially including the normal distribution. Moreover, this particular type of distribution is common in computer vision. With such a distribution, the residuals after least-squares fitting will be similarly distributed. We can therefore detect outliers by quantifying the extent to which the distribution of the residuals matches our noise model.

3.2 Algorithm Overview

We begin by creating the set $S = \text{sign}(R - \text{median}(R))$. By detecting the zeroes at the median from $S$, we now have a set whose elements may be represented as either + or -. Following von Mises [8, page 184] we define a run as a sequence of one or more symbols of the same sign. For example the set $S = \{-+++++---\}$ contains runs of lengths 1,1,3,2,1 respectively. Intuitively, we would expect that if the model fits well, there will be a large number of short runs, with long runs of positive or negative residuals indicating that the model has been biased. This idea was used by Besl [1] to decide whether a model was of high enough order to describe the data.

Measuring the likelihood of a particular distribution of runs is a problem that has been approached in the statistical literature [3, 4, 6]. In particular, having decided to measure the runs, the question arises as to how to quantifiy the deviation of a particular example from the general population. Kemphorne et al [4, page 234] calculate the expected value and variance of the total number of runs ($E[M] = n + 1, E[M^2] = \frac{n(n-1)}{2}$), and approximate the distribution by a Gaussian in order to calculate probabilities.

Figure 1: Empirically derived distributions of run frequencies for two values of $n$, the number of data points.

This approach, taken also by Brownlee[3], von Mises [8] and Mood[6], simplifies the analysis, but reduces the sensitivity of the test. In this paper, we instead compare the "actual" distribution to the observed distributions using a modified Kolmogorov-Smirnov test [3].

3.3 Comparing the distributions

If we make a histogram $H(j)$ where bin $j$ contains the number of runs of length $j$ in the residuals, then the sequence

$$C_k = \sum_{j=1}^{k} H(j), \quad 1 \leq k \leq n$$

will approximate the cumulative distribution function. By comparing this function to the predicted cdf $P$ given by a zero-median process (see Figure 1), we can determine the extent to which the outlier process has corrupted the fit. Comparison of cdfs normally entails use of the Kolmogorov-Smirnov test, where the likelihood is calculated from the known distribution of $D = \max | C_k - P(k) |$. However, this has the well-known [3] disadvantage that the sample variance of $D$ varies with $k$. Our alternative, arrived at experimentally, was to calculate the weighted sum of distances

$$D = \sum_{k=1}^{n} \frac{(P(k) - C_k)w_k}{\sum w_k}$$

In the experiments described below, the weighting function used was a simple quadratic $w_k = k^2$ chosen to give more importance to longer runs.

3.4 Determining the Actual Distribution

To enable use of the Kolmogorov-Smirnov test, we must know the expected distribution of our measure. To this end
we performed a Monte-Carlo simulation of the fitting process and recorded the results. For this example, we modelled the sensor noise process as a Gaussian plus quantization, which is an appropriate model for the laser range finder in use in our laboratory.

The distributions (graphed in Figure 1) were calculated as follows: For a given number of points $n$, the line $y = x^2 + 1, x = 1...n$ was corrupted by Gaussian noise of $\sigma = 5$, then quantized to the next lowest integer. The runs histogram was calculated using the residuals of a linear least-squares fit. Repeating this process 5000 times, and measuring the cumulative frequencies for each length of run gave the distributions shown. This technique was chosen because it was felt that the particular choice of this line would not alter the results. To test this conjecture, the line slope and noise were varied widely and the experiment repeated. Results were comparable to within about 0.2 percent. However, changing the model to a quadratic altered the frequencies by up to 10 percent, suggesting that in real applications, it is important to "train" the test on the models expected.

We note that although the histogram should be calculated for all possible values of the number of data point $n$ (up to $10^6$ in a 2D system), there was no significant change in the frequencies after about $n = 100$, lightening the computational load significantly.

4 Experiments

A number of experiments were performed to assess the performance of the new test and compare it to existing tests. The three tests were designed to be representative of ‘everyday’ computer vision tasks.

4.1 Tracking

Here we consider the problem of tracking a point through time or space while maintaining an estimate of its trajectory. The tracking can often be foiled when one point passes in front of another and the program begins to follow the second point. The error may be detected by examining the fit between the trajectory model and the data. In this experiment the track is represented by a line at 45 degrees which proceeds for 100 points. The false trajectory is then represented by a second line of 50 points joining the first at an angle of 90 degrees. The response is observed for two different noise levels.

4.1.1 Procedure

The following experiment was performed 1000 times for each noise level:

1. Gaussian noise was added to the trajectory described above, and the measurements quantized.
2. For each $n$ between 3 and 150, a line was least-squares fitted to the noisy data points and the results of the three goodness-of-fit tests were recorded.

This generates 3 by 1000 traces of 147 response values.

4.1.2 Results

To combine these results, we considered the mean and 98th percentile responses for each $n$. The mean value gives a smoothed impression of the abilities of the tests to reject the incorrect model, but we omit the traces here due to lack of space. The 98th percentile response indicates the potential for

---

1 Although the choice of 90° may seem arbitrary, using smaller angles proved to be equivalent to increasing the noise level on the 90° case.
false negatives with each method. To ensure a false negative rate of less than 2%, it is necessary to threshold the test at a value above the highest 98th percentile response. These traces appear in Figure 2.

4.1.3 Discussion

The graphs of Figure 2 may be interpreted as follows. To the left of the dotted vertical line, false rejections will occur if the response is high. To the right, low values imply false acceptances. A perfect test will be a step function going from 0 on the left to 1 on the right. The sensitivity of a test may be thought of as the slope of the response curve at the breakpoint. The greater the slope, the more likely the test will correctly reject outliers.

The top left graph, for the low noise case, shows all three tests performing well, particularly for large $n$. The $\chi^2$ test, having been applied using the known noise variance shows the greatest sensitivity. Despite the tendency towards false rejections, as seen on the top right, a threshold of 0.95 will give excellent rejection. With the RD test, the low false rejection rate means that a much lower threshold will give similar results.

The real advantage of the RD test becomes apparent as noise is increased. The $\chi^2$ test, with an slightly incorrect a priori noise model ($\sigma = 4$ rather than $\sigma = 5$) fails drastically, rejecting almost every point.

4.2 Segmentation

The test was applied to the problem of conic curve segmentation, with results as shown in Figure 3. This experiment indicates the ability of the test to identify subtle changes in model, at the $C_2$ discontinuity between line and circle for example. Curves were fitted to the 2D boundary of a 3D plane using Taubin's generalized eigenvector fit and the RD test used to identify outliers. This model was chosen to be similar to that used by Whaitie [9], but the results are not comparable without knowing the use to which the segmentation is intended to be put.

Figure 3: Segmentation results. The tracked edge data on the left has been segmented into the lines and circles shown on the right. The RD test is used to identify the breakpoints (shown as dots on the right).

5 Conclusions

We have introduced a new method of testing the hypothesis that some unknown data set is a noisy instance of a parametric model. Our method is superior to existing methods that make unrealistic assumptions about the noise characteristics of the input data. The method is fast, and can in most cases be made to have $O(n)$ time and space complexity. Sensitivity to small deviations in the model is high, while the false rejection rate is extremely low, even when the data are heavily corrupted by noise. The major advantage of our test however is that there is no need to know the input noise level.

The main disadvantage is that the user of the test does need at least to be able to simulate their noise process in order to generate the a priori histograms. In addition, if the distribution is nonsymmetric, the test will not be as sensitive.

A problem with the system is that in situations where quantization error grossly exceeds sensor error, the noise model is violated and the false rejection rate increases sharply. This can be avoided by adding a little Gaussian noise to the data, but this is obviously not an ideal solution.

6 REFERENCES


A TIME AND SLOPE DOMAIN THEORY OF MORPHOLOGICAL SYSTEMS: SLOPE TRANSFORMS AND MAX–MIN DYNAMICS

Petros Maragos

School of Electrical & Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332–0250, USA

ABSTRACT

In this paper we present a theory for a broad class of nonlinear systems obeying a supremum/infimum-of-sums superposition and a collection of related analytic tools, which parallel the functionality of and have many conceptual similarities with ideas and tools used in linear systems. In the time domain, the equivalence of these systems with morphological dilation or erosion by their impulse response is established, and a class of nonlinear (max–min) difference equations is introduced to describe their dynamics. Finding that the affine signals \( \alpha t + b \) are eigenfunctions of such morphological systems leads to developing a slope response for them, as a function of the slope \( \alpha \), and related slope transforms for arbitrary signals. These ideas provide a transform (slope) domain for morphological systems, where dilation and erosion in time corresponds to addition of slope transforms, time lines transform into slope impulses, and time cones transform into bandpass slope-selective filters.

1. MORPHOLOGICAL SYSTEMS

Morphological systems have found many applications in image analysis and nonlinear filtering. All are based on (simple or complex) parallel or serial interconnections of morphological dilations \( \ominus \) or morphological erosions \( \oslash \) [7, 6]

\[
(x \ominus g)(t) = \bigvee_{\tau} x(\tau) - g(\tau - t), \quad (x \oslash g)(t) = \bigwedge_{\tau} x(\tau) - g(\tau - t)
\]

where \( \bigvee \) denotes supremum and \( \bigwedge \) denotes infimum. So far their analysis has been done only in the time domain by using their algebraic properties and lacked a transform domain.

In this paper we first endow morphological systems with various concepts and analytic methods that enable us to determine their output and several properties of these nonlinear systems in the time domain based on their impulse response. Specifically, we call a signal operator \( D : x \mapsto y = D(x) \) a dilation translation-invariant (DTI) system if it is time-invariant and obeys the morphological supremum superposition principle \( c \in \mathbb{R} \)

\[
D \left[ \bigvee_{i} c_{i} + x_{i}(t) \right] = \bigvee_{i} c_{i} + D[x_{i}(t)]
\]

(1)

For DTI systems we assume input and output signals \( x : E \rightarrow \mathbb{R} \) with a continuous (E=R) or discrete domain (E=Z) and whose range is any subset of \( \mathbb{R} = \mathbb{R} \cup \{-\infty, \infty\} \). The useful information in a signal \( x \) analyzed by a DTI system exists only at times \( t \) in its support \( \text{Sp}(x) = \{ t : x(t) > -\infty \} \). The morphological zero impulse \( \mu \) and zero step \( \lambda \)

\[
\mu(t) \triangleq \begin{cases} 0, & t = 0 \\ \infty, & t \neq 0 \end{cases}, \quad \lambda(t) \triangleq \begin{cases} 0, & t > 0 \\ \infty, & t < 0 \end{cases}
\]

are two elementary signals useful for analyzing morphological systems. For example, the impulse response \( g(t) = D[\mu(t)] \) uniquely characterizes a DTI system in the time domain and determines its causality and stability. Specifically [4]

\[
D \text{ is DTI} \iff D(x) = x \oplus g, \quad g \triangleq D(\mu)
\]

Thus any DTI system is equivalent to a morphological dilation of the input with its impulse response. Further, \( D \) is causal iff \( g(t) = -\infty \forall t < 0 \) and stable iff \( \sup \{|g(t)| : t \in \text{Sp}(g)\} < \infty \).

Operators \( E : x \mapsto y = E(x) \) that are time-invariant and obey an infimum superposition, i.e. as in (1) but with \( \bigwedge \) replaced by \( \bigvee \), are called erosion translation-invariant (ETI) systems. These are equivalent to a morphological erosion, because \( E \) is ETI iff \( E[\mu(t)] = \mu(t) \oplus E(-t) \), where \( f = E(-t) \) is defined as its impulse response.

To describe the time dynamics of DTI systems we also develop nonlinear difference equations. Inspired here comes from the linear difference equations which can describe a very large class of discrete linear time-invariant (LTI) systems. Replacing sum with maximum and multiplication with addition gives us the following max difference equation

\[
y[n] = \left( \bigvee_{k=1}^{N} a_k + y[n-k] \right) \vee \left( \bigvee_{m=0}^{M} b_m + x[n-m] \right)
\]

(2)

All coefficients \( a_k, b_m \) are from \( \mathbb{R} \cup \{-\infty\} \). \( N \) is the order of the equation, assuming \( a_N > -\infty \). The vast majority of discrete-time morphological dilations used in applications employs a finite structuring element, and they can be modeled by (2) by ignoring the recursive part (i.e., if all \( a_k = -\infty \)). The only exception is the 1st-order recursive dilation \( y[n] = \max(y[n-1], y[n]) \), which can generate the distance transform of binary images, useful for image analysis. Whenever the max equation has a recursive part, we show that this corresponds to dilating the input signal with an infinite-support structuring function.

To create a transform domain for morphological systems, after finding that the line signals \( \alpha t + b \) are their eigenfunctions, we introduce a 'slope response', a function of the slope variable \( \alpha \), which enables us to understand the systems behavior in a transform domain—the slope domain. The
affine signals \( z(t) = \alpha t + b \) are eigenfunctions of any DTI system \( \mathcal{D} \) or ETI system \( \mathcal{E} \) because

\[
\mathcal{D}[\alpha t + b] = \alpha t + b + G(\alpha), \quad G(\alpha) \overset{\Delta}{=} \int g(t) - \alpha t \\
\mathcal{E}[\alpha t + b] = \alpha t + b + F(\alpha), \quad F(\alpha) \overset{\Delta}{=} \int f(t) - \alpha t
\]

We call the corresponding eigenvalues \( G(\alpha) \) and \( F(\alpha) \) the slope response of the DTI and ETI system. It measures the amount of shift in the intercept of the input lines with slope \( \alpha \). It is also conceptually similar to the frequency response of LTI systems which is their multiplicative eigenvalue for input exponentials, whereas \( G \) (or \( F \)) is the additive eigenvalue of DTI (or ETI) systems for input lines. This nonlinear analysis leads to developing signal transforms called slope transforms whose properties and application to morphological systems has some striking conceptual similarities with Fourier transforms and their application to LTI systems.

This paper is a summary of our results in [3, 4, 5].

2. SLOPE TRANSFORMS

The following two (sup/inf-based) slope transforms, originally introduced by Maragos [3, 4, 5] in the context of morphological systems, were motivated by the algebraic expression of their eigenvalues corresponding to their eigenfunctions \( \alpha t + b \). (Recall that the Fourier transform can be similarly inspired by the form of the eigenvalues (frequency response) of LTI systems corresponding to their exponential eigenfunctions.) Thus, viewing the slope response as a signal transform with variable the slope \( \alpha \), we define for any signal \( x(t) \) its upper slope transform as the function \( X_U : \mathbb{R} \to \mathbb{R} \) and as lower slope transform\(^1\) the function \( X_L : \mathbb{R} \to \mathbb{R} \) defined, for each \( \alpha \in \mathbb{R} \), as

\[
X_U(\alpha) \overset{\Delta}{=} \bigvee_{t \in \mathbb{R}} x(t) - \alpha t, \quad X_L(\alpha) \overset{\Delta}{=} \bigwedge_{t \in \mathbb{R}} x(t) - \alpha t
\]

A geometrical intuition behind the slope transforms can be obtained by realizing that a line that has slope \( \alpha \) and passes from a point \((t, x(t))\) on the graph of a signal \( x(t) \) has an intercept equal to \( X = x(t) - \alpha t \). Thus the upper and lower slope transforms are the max and min intercepts of lines with varying slopes intersecting the signal's graph. These extreme intercepts occur when the line becomes tangent or intersects the graph at only one point. In general, \( x(t) \) is covered from above by all the lines \( X_U(\alpha) + \alpha t \) whose infimum creates an upper envelope \( \hat{z}(t) \) and is covered from below by all the lines \( X_L(\alpha) + \alpha t \) whose supremum creates the lower envelope \( \underline{z}(t) \):

\[
\hat{z}(t) \overset{\Delta}{=} \bigvee_{\alpha \in \mathbb{R}} X_U(\alpha) + \alpha t, \quad \underline{z}(t) \overset{\Delta}{=} \bigwedge_{\alpha \in \mathbb{R}} X_L(\alpha) + \alpha t
\]

We view the signals \( \hat{z}(t) \) and \( \underline{z}(t) \) as the 'inverse' upper and lower slope transform of \( z(t) \), respectively.

**Theorem 1** [5]. For any signal \( z : \mathbb{R} \to \mathbb{R} \),

(a) \( X_U(\alpha) \) and \( X_L(\alpha) \) are convex, whereas \( X_U(\alpha) \) and \( X_L(\alpha) \) are concave. (b) For all \( t \), \( \hat{z}(t) \leq x(t) \leq \underline{z}(t) \).

(c) At any time instant \( t \)

\[
\hat{z}(t) = x(t) \iff x(t) \geq \frac{px(t-q) + qx(t+p)}{p+q}, \quad \forall p, q > 0.
\]

At any \( t, x(t) = \hat{z}(t) \) if the \( \geq \) sign in (3) is replaced by \( \leq \).

(d) \( \hat{z}(t) = x(t) \) for all \( t \) if \( x \) is concave, and \( \hat{z} = \underline{z} \) if \( x \) is convex. (e) \( \hat{z} \) is the smallest concave upper envelope of \( x \), and \( \underline{z} \) is the greatest convex lower envelope of \( x \).

Thus, there is one-to-one correspondence between \( X_U(\alpha) \) and the signal envelope \( \hat{z}(t) \). However, all signals between \( x(t) \) and \( \underline{z}(t) \) will have the same upper slope transform.

Tables I and II list several properties and examples of the upper slope transform. Their proofs are in [6]. The most striking is Property 8, i.e., that dilation in the time domain corresponds to addition in the slope domain. Note the analogy with LTI systems where convolving two signals in time corresponds to multiplying their Fourier transforms.

Consider the rectangular time pulse \( w(t) \), equal to 0 for \( t \leq [T] \) and \(-\infty \) else, added to a signal \( x(t) \). The upper slope transform of the time-limited signal \( x(t) + w(t) \) is the erosion of the original signal's slope transform \( X(\alpha) \) by the negative of the window's slope transform \( W(\alpha) = T[\alpha] \). See Fig. 1. This is a kind of nonlinear blurring. Consider the analogy with the blurring that occurs when we multiply a signal \( x(t) \) by a time window in which case the original Fourier transform of \( x \) is convolved with the window's Fourier transform.

There is a duality between the time and slope domain, similar to the duality between time and frequency domains of Fourier transform pairs. For example, Table II implies that time lines, half-lines, and cones transform respectively into slope intervals, steps, and pulses, and vice-versa.

Whatever we discussed for upper slope transforms also applies to the lower slope transform, the only differences being the interchange of suprema with infima, concave with convex, and dilation with erosion.

For differentiable signals, the maximization or minimization of the intercept \( x(t) - \alpha t \) involved in both slope transforms can also be done, for a fixed \( \alpha \), by finding its value at the stationary point \( t^* \) such that \( x'(t^*) = \alpha \) where \( x' = dx/dt \). At the point \((t^*, x(t^*))\) the line becomes tangent to the graph. This extreme value of the intercept (as a function of the slope \( \alpha \)) is the Legendre transform of the signal \( x \):

\[
X_L(\alpha) \overset{\Delta}{=} \sum_{a(x')^{-1}(\alpha)} -a(x')^{-1}(\beta)
\]

Figure 1. (a) Original parabola signal \( z(t) = t^2/2 \) (in dashed line) and a time-limited version (in solid line) resulting from adding to the signal a rectangular pulse with support \([-5, 3]\). (b) Upper slope transform of the parabola (in dashed line) and of its time-limited version (in solid line).
where $f^{-1}$ denotes the inverse of a function $f$. It is extensively used in mathematical physics [1]. If the signal $x(t)$ is concave or convex and has an invertible derivative, its Legendre transform is single-valued and equal (over the slope intervals it is defined) to the upper or lower transform. Examples 7–12 deal with such signals $x$ with invertible derivatives.

If a differentiable signal is neither convex nor concave or if it does not have an invertible derivative, the Legendre transform is multi-valued; i.e., $(x')^{-1}(a)$ and hence $X_L(a)$ is a set of real numbers for each $a$. For example, consider the cosine $x(t) = \cos(\omega t)$ over all time, which is an infinite sequence of convex and concave cosine pulses. Then

$$X_L(a) = \{Y(\alpha) + \alpha kT, -Y(\alpha) + \alpha kT : k = 0, \pm 1, \pm 2, \ldots\}$$

where $Y$ is the slope transform of a single concave cosine pulse (Example 12, Table II). In general, the number of different functions in the multivalued Legendre transform is equal to the number of consecutive convex and concave pieces making up the signal. This could be finite or infinite. This multivalued Legendre transform is defined in [2] as a 'slope transform' and is expressed via stationary points; i.e., $X_L(a) = \{x(t) - a\alpha : x(t) = a\}$, its properties in [2] seem similar to the properties of the upper/lower slope transform, but there are some important differences (see [4, 5]) stemming from the fact that operations among multivalued Legendre transforms are actually set operations.

An arbitrary signal can be analyzed using slope transforms toward at least two different goals: signal reconstruction, or envelope reconstruction. For exact signal reconstruction, we should segment the signal into concave and convex and concave pieces and find the slope transform of each piece. The set collection of slope transforms of the signal pieces can reconstruct the signal exactly. The disadvantage here is the multivaluedness of the transform. Alternatively, for extracting information about the long-time behavior of the signal, as manifested by its upper and lower envelope, we could compute its upper and lower slope transforms and take their inverses, which give us the two envelopes. Examples of this latter case include the impulse responses of recursive DTI systems (discussed later) and amplitude-modulated signals where we seek to estimate their envelope.

Consider sampling a continuous-time signal $x_c(t)$ at time instants $n = nT$ and obtaining the sampled signal $x_s(t) = \sum_{n=-\infty}^{\infty} x[n] \dot{\mu}(t-nT)$, where $x[n] = x_s(nT)$ is the discrete-time signal. Let $X_\alpha(x_c), X_\alpha(x)$ be the continuous-time upper slope transforms of the signals $x_c(t), x_s(t)$. We define the upper slope transform of the discrete signal $x[n]$ by

$$X\alpha(a) \triangleq \sum_{n=-\infty}^{\infty} x[n] - an, \quad \alpha \in \mathbb{R}$$

Then $X_\alpha(a) = X_\alpha(a/T) \leq X_\alpha(a/T)$.

Another effect of sampling is to replace parts of the slope transform of the continuous-time signal with supporting lines. Further, if $x_c(t)$ is a concave piecewise-linear signal and the sampling time instants $t = nT$ include all the times at which its corner points occur, then $X_\alpha(a) = X_\alpha(a)$ for all $\alpha$ and the original signal $x_c(t)$ can be exactly reconstructed from its samples by applying an upper slope transform on $x_c(t)$ followed by its inverse transform.

The definitions and (almost all) properties of discrete upper and lower slope transforms and their inverses are identical to the continuous-time case, except that the time variable is discrete. Examples 1–6 of slope transform pairs in Table II hold in discrete time.

### 3. MAX DIFFERENCE EQUATIONS

In this section we consider discrete-time signals and view (2) as a nonlinear system $\Psi : x \rightarrow y = \Psi(x)$. To solve (2) in forward time $n \geq n_0$, we need $N$ initial conditions $IC[n_0]$, where $IC[n] = \{y[-1], y[-2], \ldots, y[-N]\}$. If all the values in $IC[n_0]$ are $-\infty$, the initial state of the system does not affect its output. We define the impulse response $g$ of $\Psi$ as its output when the input is the impulse and $IC[0] = -\infty$. The solution of the 1-st order ($N = 1, M = 0$) max equation (2) is found by induction on $n \geq 0$ to be

$$y[n] = (x[n] + y[n]) \vee (a_1(n + 1) + y[-1])$$

where $g[n] = a_1n + b_0 + \lambda[a]$ is the impulse response. Thus the general solution of (2) for $N = 1$ is the maximum of the $(-\infty)$-state response (i.e., the dilation $x \triangleright y$) and the $(-\infty)$-input response due only to the initial condition $y[-1]$. The system is stable only if $a_1 = 0$. Similar results are also true for the general $N$th-order max difference equation.

Theorem 2. The max difference equation (2) corresponds to a causal DTI system if (i) whenever $x[n] = -\infty$ for all $n < n_0$ then $y[n] = -\infty$ for all $n < n_0$, where $n_0$ is an arbitrary but otherwise fixed time instant, and (ii) the required initial conditions $IC[n_0]$ are $-\infty$.

Henceforth we shall make the two assumptions of Theorem 2 for systems described by (2). There are two major subclasses of such DTI systems:

Finite Impulse Response (FIR) DTI systems, when $a_k = -\infty$ for all $k$. Then (2) has no recursive part, and the impulse response has finite support because $g[n] = b_0$, if $n = 0, 1, \ldots, M$ and $g[n] = -\infty$ elsewhere. All these systems are stable. This class is identical with the class of all morphologically dilations with finite-support structuring elements.

Infinite Impulse Response (IR) DTI systems, when $a_k \neq -\infty$ for at least one $k$. The example of the 1-st order system demonstrates that such systems have an impulse response of infinite support. Their stability is controlled by the absolute maximum value of $g$.

Henceforth, we focus only on the recursive part of (2) by setting $b_0 = 0$ and $b_m = -\infty$ for $m > 0$. A 1-st order system $g[n] = \max\{y[n-1] + a_1, x[n]\}$ has impulse response $g[n] = a_1n + \lambda[n]$ and slope response $g(x) = -\lambda(a_1 - a_1)$. It acts as a 'slope highpass' filter since it passes from the input signal only those segments whose upper slopes are $\geq a_1$. For a system order $N > 1$, finding a closed-form expression for the impulse response is generally not possible. However, we can first find the slope response $G$ and then, via inverse slope transform, find the impulse response $g$ or its envelope $\hat{g}$. Thus, applying upper slope transform to (2) and using the fact that $X(x) = G(x) + X_V(x)$ yields

$$G(x) = \max[G(x) - a + a_1, \ldots, G(x) - N\alpha + a_N, 0]$$

A nontrivial (i.e., different than $-\infty$) solution $G$ is

$$G(x) = -\lambda(a_1 - a_0), \quad a_0 = \max_{k} \frac{a_k}{k}$$

The inverse slope transform on $G$ yields the upper envelope $\hat{g}$ of the impulse response

$$\hat{g}[n] = \max[a_0 + \lambda[n], \geq g[n]]$$

Over short time periods $g$ has the shape induced by the sequence $\{a_k\}$ and dominates the output of the recursive DTI system during time periods when the slope of the input signals is smaller than $a_0$. But over time scales much longer
than the length of the coefficient sequence \( \{a_k\} \) it behaves like its upper envelope \( \tilde{g} \). Together \( G \) and \( \tilde{g} \) can describe the long-time dynamics of the system where they predict a behavior approximately equivalent to a 1st-order system whose cutoff slope is \( \alpha_0 \). In addition, if \( g \) is a line, then the above analysis is also exact for the short-time behavior. Note also that by appropriately choosing the coefficients \( \{a_k\} \) we can give the short-time variations of \( g \) many different patterns, even periodic [3, 4].

3.1. SLOPE FILTERS

Consider the causal recursive DTI system

\[ y[n] = \max[0, y[n-1] + a_1, x[n]] \]

with \( a_1 < 0 \), which is a morphological dilation of the input by the causal line \( y[n] = a_1 n + \lambda[n] \). The output \( y[n] \) provides a type of upper envelope of \( x[n] \). As Fig. 2 shows, when computing \( y[n] \) in forward time, during periods where the signals peaks keep decreasing \( y[n] \) falls linearly with slope \( a_1 \) in between these consecutive peaks. When the envelope peaks start increasing, \( y[n] \) continues to fall between peaks, whereas it should rise. The slope response of this system is \( G_1(\omega) = -\lambda(\omega - a_1) \) and hence rejects all negative slopes \( < a_1 \). To be able to also reject some positive slopes we must pass the input through an anti-causal system

\[ y[n] = \max[0, y[n+1] + a_2, x[n]] \]

with \( a_2 > 0 \), run backwards in time (see Fig. 2). It corresponds to a morphological dilation of the input by the anti-causal line \( y[n] = a_2 n + \lambda[-n] \). Its slope response is \( G_2(\omega) = -\lambda(a_2 - \omega) \) and hence it rejects all positive slopes \( > a_2 \). To symmetrize this process we can take the maximum \( y = y_1 \lor y_2 \) of the two envelopes as the final estimated upper envelope of the input. The mapping \( x \rightarrow y \), i.e., the maximum of two DTI systems, is another DTI system with overall impulse response \( g = y_1 \lor y_2 \) and overall slope response \( G = G_1 \lor G_2 \):

\[
g[n] = \begin{cases} 
  a_1 n, & n \geq 0 \\
  a_2 n, & n < 0 
\end{cases}, \quad G(\omega) = \begin{cases} 
  0, & a_2 \leq \omega \leq a_1 \\
  +\infty, & \text{else}
\end{cases}
\]

This is an ideal-cutoff bandpass slope-selective filter. To design a symmetric slope filter we select \( a_2 = -a_1 = \alpha_0 > 0 \) which passes upper slopes with magnitude \( \leq \alpha_0 \) unchanged and rejects all other slopes. This is the case in Fig. 2. Such bandpass slope filters, implemented via 1st-order recursive max/min equations, have been applied to envelope detection from AM signals in [3, 4].

4. REFERENCES


### TABLE I: Properties of Upper Slope Transform

<table>
<thead>
<tr>
<th>No.</th>
<th>Signal: ( x(t) )</th>
<th>Transform: ( X(\omega) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>( \sqrt{c_1 + x(t)} )</td>
<td>( \sqrt{c_1 + X(\omega)} )</td>
</tr>
<tr>
<td>2.</td>
<td>( x(t - t_0) )</td>
<td>( X(\omega - \alpha t_0) )</td>
</tr>
<tr>
<td>3.</td>
<td>( x(t) + \alpha t )</td>
<td>( X(\omega - \alpha) )</td>
</tr>
<tr>
<td>4.</td>
<td>( x(\tau t) )</td>
<td>( X(\omega / \tau) )</td>
</tr>
<tr>
<td>5.</td>
<td>( x(-t) )</td>
<td>( X(-\omega) )</td>
</tr>
<tr>
<td>6.</td>
<td>( x(t) x(t) )</td>
<td>( X(\omega) X(-\omega) )</td>
</tr>
<tr>
<td>7.</td>
<td>( r(t) x(t) )</td>
<td>( r(\omega) X(\omega) )</td>
</tr>
<tr>
<td>8.</td>
<td>( x(t) y(t) )</td>
<td>( X(\omega) Y(\omega) )</td>
</tr>
<tr>
<td>9.</td>
<td>( x(t) + y(t) + t )</td>
<td>( X(\omega) + Y(\omega) )</td>
</tr>
<tr>
<td>10.</td>
<td>( x(t) x(t) )</td>
<td>( X(\omega) X(\omega) )</td>
</tr>
<tr>
<td>11.</td>
<td>( x(t) )</td>
<td>( X(\omega) )</td>
</tr>
<tr>
<td>12.</td>
<td>( y(t) )</td>
<td>( y(t) )</td>
</tr>
</tbody>
</table>

### TABLE II: Examples of Upper Slope Transforms

<table>
<thead>
<tr>
<th>No.</th>
<th>Signal: ( x(t) )</th>
<th>Transform: ( X(\omega) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>( \alpha t )</td>
<td>( -\alpha(\omega - \alpha) )</td>
</tr>
<tr>
<td>2.</td>
<td>( \alpha t + \lambda t )</td>
<td>( -\lambda(\omega - \alpha) )</td>
</tr>
<tr>
<td>3.</td>
<td>( m(t - t_0) )</td>
<td>( -\alpha t_0 )</td>
</tr>
<tr>
<td>4.</td>
<td>( \lambda(t - t_0) )</td>
<td>( -\alpha t_0 - \lambda(\omega) )</td>
</tr>
<tr>
<td>5.</td>
<td>( c_1,</td>
<td>t</td>
</tr>
<tr>
<td>6.</td>
<td>( -\alpha t, \alpha &gt; 0 )</td>
<td>( 0, \alpha \leq \alpha_0 )</td>
</tr>
<tr>
<td>7.</td>
<td>( \sqrt{1 - r^2} ), (</td>
<td>t</td>
</tr>
<tr>
<td>8.</td>
<td>( -\alpha^2 )</td>
<td>( \alpha^2 / 2 )</td>
</tr>
<tr>
<td>9.</td>
<td>( -</td>
<td>p</td>
</tr>
<tr>
<td>10.</td>
<td>( \exp(\lambda t) )</td>
<td>( \alpha(1 - e^{\omega^2}) )</td>
</tr>
<tr>
<td>11.</td>
<td>( \tanh(t), \tanh &gt; 0 )</td>
<td>( \sqrt{1 - \alpha^2} )</td>
</tr>
<tr>
<td>12.</td>
<td>( \cos(\omega t) ), (</td>
<td>t</td>
</tr>
</tbody>
</table>

Figure 2. Dotted line shows input signal \( x[n] = (1 + 0.5\cos(2\pi n/500)) \cos(2\pi n/100) \). The solid (resp. dashed) line is the output of the recursive equation \( y[n] = \max[0, y[n-1], x[n]] \) run in forward (resp. backward) time. Final upper envelope is the max of the solid and dashed curves.
Chamfer Distances in Anisotropic 3D Images

Jean-François MANGIN†, Isabelle BLOCH†, Jaime LÓPEZ-KRAHE† and Vincent FROUIN‡

†: Département Images, Télécom Paris, 46 rue Barrault, 75634 Paris Cedex 13, France,
Tel | Fax: +33 1 45 81 75 85 | 45 81 37 94, E-Mail: mangin@ima.enst.fr
‡: Service Hospitalier Frédéric Joliot, CEA, 4 place du 6e Lecle, 91401 Orsay cedex, France,
Tel | Fax: (+33 1) 69 86 77 70 | 69 86 77 08, E-Mail: frouin@uruens.shfj.cea.fr

Abstract. A number of registration methods relying on the computation of a distance map have been proposed recently in the field of 3D medical imaging. This distance map is usually computed using a fast distance transformation which approaches the Euclidean distance to the object of interest. Until now, 3D distance transformations were designed for a cubic lattice and consequently were not adapted to the voxel anisotropy of the images produced by most of the scanners. In this paper, we propose a process which allows the adaptation of the chamfer transformation, the most popular distance transformation, to any anisotropic lattice. This adaptation is performed in a context specific to registration applications, where the accuracy of the Euclidean distance approximation is only required near the object. The new distance transformation consists in an adaptive chamfer transformation followed by a postprocessing which gives the exact Euclidean distance in a neighbourhood of the object.

1. Introduction

Distance transformations (DT) are used in a number of image processing techniques to estimate the Euclidean distance between points and objects on a lattice. Among a wide variety of DT, the most popular is the chamfer transformation (CT), which approximates the Euclidean distance by propagating local distances [1, 2]. The increasing number of medical imaging modalities producing 3D images with anisotropic voxels calls for an extension of the CT to anisotropic lattices. For instance, the use of chamfer distance maps to design accurate multi-modality image registration algorithms has recently become a topic of interest in many laboratories [3, 4, 5, 6, 7]. Whereas the extension of the CT principle to anisotropic lattices is straightforward, the extension of the works which aim at finding the parameters of the CT which realize the best trade-off between computation cost and quality of the estimation of the Euclidean distance is much more problematic.

In this paper, we first propose a study of the general behaviour of the CT on 2D and 3D anisotropic lattices in order to show to what extent results of previous works for square or cubic lattices may be extended. In particular, we prove that assumptions usually made are not always satisfied in the 3D case. Next, we propose an original method to choose the parameters of the CTs used in registration applications. This method consists in designing adaptive DTs yielding the exact Euclidean distance near the object and elsewhere, an estimation whose quality decreases when the distance to the object increases, unlike the standard CTs which provide good approximations far from the object. Proofs of the results stated in this paper may be found in [8].

2. Anisotropic Chamfer Transformation

Given an anisotropic orthogonal lattice, a way to define a chamfer metric is:

- define a set $P$ of prime\(^1\) vectors on the lattice [9];
- assign a local distance $d_i$ to each $p_i$ of $P$.

Let consider the graph whose nodes are the lattice points and whose arcs are defined by $P$. The length of an arc defined by vector $p_i$ is the local distance $d_i$. The chamfer distance between two points $A$ and $B$ is the minimum path length among all the paths between $A$ and $B$ in this graph. It will be noted $d^P_E(A, B)$.

The computation cost of the CT depends linearly on the cardinal of $P$. Hence, searching for a good CT for a given lattice geometry amounts to searching the set of prime vectors satisfying the requirements on the user. Because of the complexity of the discrete geometry, it would be unrealistic to search for optimal solutions. Therefore, a reasonable approach seems to choose $P$ heuristically and to search for suboptimal local distances given $P$. In order to show that this heuristic choice is highly dependent on the user requirements mentioned above, we propose a general framework allowing to study the behaviour of the CT.

Notations are now introduced in order to state results about the chamfer distance of any point to a given point $O$ of the lattice.

\(^1\)Two vectors of the set $P$ can not be colinear and of same direction.
1. **2D CT**: Let \( \bar{p}_i \) and \( \bar{p}_j \) be two linearly independent vectors of \( \mathcal{P} \); we note:

- \( \mathcal{C}_{ij} \), the discrete half-cone defined as the set of lattice points \( M \) verifying:
  \[
  \exists (a_i, a_j) \in \mathbb{R}^{+2}, \quad O\bar{M} = a_i\bar{p}_i + a_j\bar{p}_j;
  \]
  
- \( \mathcal{Q}_{ij} \) the discrete parallelogram defined as the subset of \( \mathcal{C}_{ij} \) corresponding to \( (a_i, a_j) \in [0, 1]^2 \);

- \( \mathcal{Q}_{ijk} \) the subset of \( \mathcal{Q}_{ij} \) corresponding to \( (a_i, a_j) \in [0, 1]^2 \).

2. **3D CT**: For a family of 3 independent vectors of \( \mathcal{P} \), \( \{ \bar{p}_i, \bar{p}_j, \bar{p}_k \} \), we define a similar way the half-cone \( \mathcal{C}_{ijk} \), the parallelepiped \( \mathcal{Q}_{ijk} \) and its inner \( \mathcal{Q}_{ijk}^\circ \).

3. **2D or 3D**: For \( \bar{p}_i \in \mathcal{P} \), we note \( P_i \) the point of the lattice verifying \( \overrightarrow{OP_i} = \bar{p}_i \).

**Definition 1** A half-cone \( \mathcal{C}_{ij} \) (respectively \( \mathcal{C}_{ijk} \)) is a half-cone of influence, denoted by \( \mathcal{C}_{ij}^{\circ} \) (resp. \( \mathcal{C}_{ijk}^{\circ} \)), if \( \forall l \in \{1, \ldots, n_P\} \):

\[
P_l \in \mathcal{Q}_{ij} \quad \text{resp.} \quad \mathcal{C}_{ijk} \Rightarrow l \in \{i, j\} \quad \text{resp.} \quad \{i, j, k\}.
\]

We extend this notion to parallelogram (resp. parallelepipeds) of influence. This definition is related to [10].

For \( M \in \mathcal{C}_{ij} \), a simple majoration of \( d_p^2(O, M) \) can be inferred from the natural tessellation of the half-cone generated by the corresponding parallelogram \( \mathcal{Q}_{ij}^\circ \):

\[
OM = a_i\bar{p}_i + a_j\bar{p}_j, \quad (a_i, a_j) \in \mathbb{R}^{+2}
\]

\[
= \overrightarrow{OX} + |a_i|\bar{p}_i + |a_j|\bar{p}_j, \quad X \in \mathcal{Q}_{ij}^\circ \cup \{O\};
\]

hence

\[
d_p^2(M, O) \leq d_p^2(X, O) + |a_i|d_i + |a_j|d_j, \quad (3)
\]

where \( |x| \) denotes the integer part of \( x \). The extension of this result to the 3D case is straightforward. Considering now the set of half-convex of influence, we define the particular majoration \( I_p^\circ(M, O) \) for the whole lattice as:

\[
\forall M \exists (i, j, k) \quad \text{resp.} \quad (i, j, k), \quad M \in \mathcal{C}_{ij}^{\circ} \quad \text{resp.} \quad \mathcal{C}_{ijk}^{\circ},
\]

and:

\[
I_p^\circ(M, O) = d_p^2(X, O) + |a_i|d_i + |a_j|d_j + |a_k|d_k, \quad (4)
\]

It should be noted that when \( M \) belongs to more than one half-cone of influence, \( I_p^\circ(M, O) \) is the same whatever the considered half-cone.

A number of methods have been proposed to choose a good set of \( d_i \), given \( \mathcal{P} \) for a square or cubic lattice [2, 9]. For the sake of simplicity, they all systematically assume

\[
d_p^2(O, M) = a_i d_i + a_j d_j + \cdots, \quad (5)
\]

Thus, they assume that \( a_i, a_j, \ldots \) (resp. \( a_k \)) are systematically integers (or \( X = O \)) and

\[
I_p^\circ(O, M) = I_p^\circ(M, O). \quad (6)
\]

We have shown [8] that these assumptions are not always verified, especially in the 3D case. In 2D, we have shown that there exists a neighbourhood of \( (||\bar{p}_i||, \ldots, ||\bar{p}_m||) \) where \( (d_1, \ldots, d_n) \) may be chosen to assure the validity of Eq.5 for all the points of the lattice belonging to the sub-lattice of their half-cone of influence generated by \( \bar{p}_i \) and \( \bar{p}_j \) (\( M \in \mathcal{C}_{ij}^{\circ}, X = O \) in Eq.2). Unfortunately, this is not true for all points of this type in the 3D case because a vector of \( \mathcal{P} \) external to \( \mathcal{C}_{ijk}^{\circ} \) may be close to a face of \( \mathcal{Q}_{ijk}^\circ \) (see Figure 1).

**Figure 1**: The vector \( a_i\bar{p}_i \) is close to \( a_j\bar{p}_j + a_k\bar{p}_k \) and the vector \( \bar{p}_m \) is close to \( \bar{p}_i \). In consequence, whereas \( M \) belongs to the half-cone of influence \( \mathcal{C}_{ij}^{\circ} \), and to the lattice generated by \( \bar{p}_i \), \( \bar{p}_j \), and \( \bar{p}_k \), the shortest path from \( O \) to \( M \) is not constituted by a combination of \( \bar{p}_i \), \( \bar{p}_j \), and \( \bar{p}_k \).

When \( X \neq O \), one can find configurations where Eq.6 is not true in 2D as in 3D. It can be shown that the problems with Eq.6 mainly occur when the angles (resp. solid angles) of adjacent half-cones of influence are not of the same order.

In fact, in 2D, the situation is very favourable with conventional sets \( \mathcal{P} \) which are usually chosen as the maximal set of prime vectors included in a square [2]. Indeed, one may show that in those cases [11]:

\[
\forall \mathcal{C}_{ij}^{\circ \circ}, \quad \mathcal{Q}_{ij}^{\circ \circ} = \emptyset. \quad (7)
\]

Once again, the 3D case is more intricate, but a number of usual sets still verify this property. Nevertheless, the choice of the maximal set of prime vectors included in a cube is not sufficient to assure property 7. For instance, let us consider the set \( \mathcal{P} \) constituted of the prime vectors \( (1, 0, 0), (1, 1, 0), (1, 1, 1), (2, 1, 0), (2, 1, 1), (2, 2, 1) \) (lattice base) and of the symmetrical vectors (maximal set included in a cube of size two), and the point \( M \) defined by \( \overrightarrow{OM} = 2\bar{i} + 2\bar{j} + \bar{k} \) [9]. The three following vectors clearly define an half-cone of influence \( \mathcal{C}_{ijk}^{\circ \circ} \):

\[
\bar{p}_i = 2\bar{i} + \bar{j},
\]

\[
\bar{p}_j = 2\bar{i} + \bar{j} + \bar{k},
\]

\[
\bar{p}_k = 2\bar{i} + 2\bar{j} + \bar{k}.
\]

Nevertheless \( \overrightarrow{OM} = \frac{1}{2}(\bar{p}_i + \bar{p}_j + \bar{p}_k) \) implies \( M \in \mathcal{Q}_{ijk}^{\circ \circ} \).

Unfortunately, when the anisotropy of the lattice becomes important, more complex \( \mathcal{P} \) have to be chosen. Indeed, since the quality of the CT in a half-cone of influence decreases with its angle (resp. solid angle), a guiding principle in the choice of \( \mathcal{P} \) is the minimiza-
tion of the half-cone angles. When this principle is applied in the case of a large anisotropy, keeping a low \( n_P \) implies the choice of a loose prime vector set, unlike the standard approaches choosing the maximal set included in a square (resp. a cube). When user requirements do not focus on the neighbourhood of the object, assuming Eq.6 and \( X = O \) is a good choice because \( d^2_e(O, M) - d^2_e(O, X) \) and \( d^2_e(O, X) \) are negligible far from the object. A good heuristic to choose \( \mathcal{P} \) is then to minimize the dispersion of the angles (resp. solid angles) of the half-cones of influence. The choice of the local distances may then be done using an estimation of the maximum error in the largest half-cone of influence using the method of Lagrange multipliers [9]. On the contrary, when accuracy is required near the object as in registration applications, the previous heuristic is not available because it tends to increase the number of half-cone of influence for which Eq.7 is not verified. Therefore we propose for such applications a dedicated adaptive approach combining several CTs.

3. Adaptive Chamfer Transformation

The idea consists in combining two (or more) CTs defined from nested sets \( \mathcal{P}_i \). The larger set \( \mathcal{P}_1 \) is simply chosen as the maximal set of prime vectors included in the Euclidean ball whose size (user defined) depends on the constraints we want to impose on \( n_P \). \( \mathcal{P}_1 \) will yield an accurate chamfer distance in a neighbourhood \( \mathcal{N}_1 \) of the object. A looser set \( \mathcal{P}_2 \) is extracted from \( \mathcal{P}_1 \) following the heuristic approach described in the previous section (see Figure 2). \( \mathcal{P}_2 \) will yield a less accurate chamfer distance outside \( \mathcal{N}_1 \). The choice of the local distances is made in order to optimize a final refinement of the DT described in the next paragraph. The sequential implementation is a direct extension of the CT sequential algorithm [2]. First, the conventional propagation algorithm performed with \( \mathcal{P}_2 \) yields \( d^2_e(O, M) \) for the whole image (where \( O \) denotes the object). Next the propagation is proceeded with the complementary of \( \mathcal{P}_2 \) in \( \mathcal{P}_1 \) conditionally to the subset of the image defined by \( d^2_e(O, M) \leq S(\mathcal{N}_1) \) (where \( S(\mathcal{N}_1) \) is a user chosen constant which defines \( \mathcal{N}_1 \) size. At the end of the propagation, \( \mathcal{N}_1 \) contains \( d^2_e(O, M) \) [8]. The general principle described here for two \( \mathcal{P}_i \) is easily extendable to a larger number of nested \( \mathcal{P}_i \) with nested neighbourhoods \( \mathcal{N}_1 \).

4. Nearest Point Correction

We propose a new improvement of CT particularly interesting in an adaptive context. All the works aiming at optimizing CT only consider chamfer distance to a single point. When considering chamfer distance to an object, we highly improve the approximation of the Euclidean distance by computing after the CT a new distance map where a pixel value is the Euclidean distance to the point of the object which is the nearest for the chamfer metric. An image of labels (one label for each object point), propagated during the CT algorithm, yields a Voronoi diagram for the chamfer metric allowing the immediate computation of this distance map after the CT.

Figure 2: 2D illustration of the adaptive chamfer transformation with two nested prime vector sets.

In an adaptive context, we only apply the nearest point correction in a neighbourhood of the object, which can be achieved in a more efficient way. Indeed, the small number of possible chamfer distances lower than \( S(\mathcal{N}_1) \) allows the precomputation of an equivalence table between values given by the CT and the corresponding Euclidean distance to the nearest point of the object. Therefore the label propagation during the CT is not required. Theoretical results are proposed in [8] in order to choose sub-optimal local distances maximizing the distance to the object at which the first difference with the Euclidean distance after the nearest point correction occurs. We have shown that with this aim in view, the prime vector norms turn out to be a better choice than the usual ones. Moreover, better discrete local distances can be found near from these values using fast simulations with the equivalence table approach. Figure 3 proposes a comparison between the result of the standard CT [5,7,11] [2] and the result of our nearest point correction with the same set of prime vectors but with the local distances (14,20,31) optimizing the correction according to the criterion mentioned above.

5. Conclusion

We have proposed in this paper a simple way to overcome the difficulties induced by the voxel anisotropy of 3D medical images in the computation of distance maps dedicated to registration applications. This method is intensively used in our institution for two years in order to register functional and anatomical images of the brain [6, 7] (see Figure 4). This method could be used in any other 3D registration application relying on a 3D distance map since the principle does not depend on the voxel anisotropy.

References

Figure 3: Relative difference between the Euclidean distance to a circle and CT results with or without the nearest point correction: prime vector set $i, i+j, 2i+j$ and symmetrics, 2D square lattice, the points located at more than 90 pixels from the object are set to zero.

Figure 4: Application of the anisotropic CT: registration of PET and MR images of the head using chamfer matching applied to the brain surface (PET images with the superimposed corresponding MRI edges after registration, three orthogonal slices).

Skeleton Extraction from Noisy Binary Images Using Statistical Morphological Filters and Simulated Annealing

Carlo S. REGAZZONI* and Anastasios N. VENETSANOPoulos#

*Department of Biophysical and Electronic Engineering (DIBE)
University of Genova, Via all'Opera Pia 11A, I-16145 Genova, Italy
Ph. +39 10 3532792, Fax: +39 10 3532134
E-mail: carlo@bibemuit.unige.it
# Department of Electrical and Computer Engineering,
University of Toronto,
10 King's College Road, Toronto, Canada

Abstract: A new shape descriptor obtained by skeletonization of noisy binary images is presented. Skeleton extraction is performed by using an algorithm based on a new class of parametrized binary morphological operators, taking into account statistical aspects. Parameters are adaptively selected during the successive iterations of the skeletonization operation in order to regulate the characteristics of the shape descriptor. A probabilistic interpretation of the scheduling strategy used for parameters is proposed by analogy to stochastic optimization techniques. Skeletonization results on patterns extracted by a change-detection method in a surveillance application are reported. Results show the greater robustness of the proposed method, as compared with other morphological approaches.

1. Introduction

Shape descriptors are feature vectors extracted from image patterns. Shape descriptors can be used by classification and coding systems to synthesize the morphological properties of the objects considered, either to identify them or to transmit their appearances. A Morphological Skeleton (MS) is an information-preserving shape-descriptor [1], i.e., a descriptor that allows one to reconstruct exactly an original shape from the values of the feature vector. The principle for obtaining the MS of a binary pattern is to keep track of the information loss through successive transformations performed by applying morphological operators. Measurements of information losses are made locally within the domain of a structuring element, which corresponds to a mobile observation area in which the image is inspected and processed. An MS does not allow one to distinguish the content of an image, for example, by separating noise from the information content related to an object. Therefore, filtering steps are required to filter out noise, before extracting the MS.

In this paper, a new iterative method for skeleton extraction is presented, namely the Statistical Morphological Skeleton (SMS), which is based on Statistical Morphology [2,3]. The SMS is a shape descriptor that allows information to be almost completely preserved without requiring pre-processing steps, even in the case of noisy images. This is made possible by the adaptive nature of the proposed shape descriptor, which operates through different successive iterations, and removes noise during the first steps.

2. Literature review and paper organization

The discrete morphological skeleton (DMS), also known as the uniform-step MS, was first defined by Serra [4] for hexagonal lattices and then extended by Maragos and Schafer [5] to rectangular grids. In both methods, the same structuring element was used during the processing steps. Maragos [6] proposed a generalized MS, which uses a different structuring element at each skeletonization step. These methods completely preserve the information content of an image, but require a considerable memory occupation. Maragos [7] proposed a new method which allows a reduction in skeleton storage requirements by removing some points from the DMS. This method aims to use the MS for coding purposes. Schonfeld and Goutis [8,9] studied the problem of finding the sequence of morphological filters to restore, in an optimal way, binary patterns from noisy images. The optimality criterion was defined as the minimization of the expected restoration error, corresponding to the distance between the original image without noise and the output of the sequence of morphological filters. Inner points of skeletons were given higher weights within the expectation functional to favour preservation of shape information versus noise.

The work by Schonfeld [9] suggests a statistical optimality criterion to determine the selection of morphological operators to be included within a sequence. A suboptimal computational solution was proposed starting from set-based considerations. In our work, too, set-based aspects are mixed with statistical considerations. However, we also explore an alternative approach: the sequence of operators is fixed in a morphological sense, i.e., an opening is considered. However, the definition of opening is extended in a statistical sense. The starting point is [1], which introduces probabilistic definitions of the basic operators of Statistical Morphology by pointing out their relationships with other statistical filters, e.g., Markov Random Fields (MRF) [10]. Such definitions imply the use of a continuous parameter which regulates the behaviour of a filter from a linear behaviour to a morphological one. In [3] we
applied Statistical Morphology to binary images and described efficient computational methods to evaluate the basic operators of Binary Statistical Morphology (BSM). We also suggested analogies between BSM operators and rank-order filters. Moreover, we observed that it could be feasible to use the analogy to MRFs by scheduling the regulation parameter in applications that require successive iterations of basic BSM filters.

This consideration is exploited here for information-extraction purposes, and an algorithm is proposed which takes advantages of the adaptive properties of BSM filters to obtain a robust skeletonization of binary patterns. The algorithm substantially corresponds to a generalized MS where the regulation parameter is changed at each step according to a logarithmic scheduling law. Logarithmic scheduling was chosen by analogy to the logarithmic scheduling that is used when Simulated Annealing [11] is applied to MRFs.

An application of this method to binary images acquired by change-detection methods for surveillance purposes (i.e., teleoperated vehicle surveillance) is described. It is shown that the method exhibits significant representation properties for noisy images, with a good computational efficiency.

In Section 3, some results by Binary Statistical Morphology, as reported in [12], are provided. In Section 4, the proposed skeleton extraction method, i.e., Statistical Skeleton (SSK), is described. In Section 5, results are presented, and conclusions are drawn.

3. Binary Statistical Morphology

Erosion and dilation are the elementary set-based operators of Mathematical Morphology. If the set of elements in a binary image and the structuring element (i.e., a subset of image points), are denoted by X and B, respectively, erosion and dilation can be defined as:

\[
X \ominus B = \{z : B + z \subseteq X\} \tag{1}
\]

\[
X \oplus B = \{z : B + z \cap X \neq \emptyset\} \tag{2}
\]

where + indicates a translation of B to an image location z. Statistical Morphology operators recently introduced by Yuille et al. [4] provide a method for including a-priori knowledge about the noise level as an additional parameter of an MF. Elementary operators of Statistical Morphology, i.e., Statistical Dilation and Statistical Erosion, are based on probabilistic definitions of set-based operators.

In [3], we described a computationally efficient implementation of an SM filter for the case of binary images by defining Binary Statistical Morphological Filters. For example, Binary Statistical Dilation (BSD) is given by:

\[
X \ominus_B B = \begin{cases} 
1 & \text{if } \beta \geq \ln \left( \frac{N - M_B^L(m)}{M_B^C(m)} \right) \\
0 & \text{if } \beta < \ln \left( \frac{N - M_B^L(m)}{M_B^C(m)} \right) 
\end{cases} \tag{3}\]

where \( E^* \) is the output of the filter at an image location m, \( B_m \) is the structuring element at m, with a cardinality equal to N, and \( \beta \) is a parameter taking into account the noise level. The dependence on an image X is provided by the term \( M_B^C(m) = \text{card} \{ x \in (B + x) \cap X \} \), which is the number of elements in the binary image X in the area of the structuring element B translated to position m. The relationship with Mathematical Morphology is shown by the fact that, in the limit \( \beta \rightarrow \infty \), BSD becomes Mathematical Dilation. For \( \beta = 0 \), BSD corresponds to a thresholded linear average over the structuring element. The threshold is fixed at 0.5 if equally probable inputs are assumed. For intermediate \( \beta \) values, BSD behaves like a rank-order filter [13], i.e., the rank increases for increasing \( \beta \) values. Continuous \( \beta \) ranges can be established over which BSD has the same behavior, i.e., the same rank number. A similar reasoning applies to Binary Statistical Erosion (BSE), denoted as \( X \ominus_B B \) can be defined, if we consider \( \beta = 0 \).

The definition (3) of BSD is obtained as an approximation of the maximum a-posteriori (MAP) probability configuration of the output set of variables H, for a given input image. \( \beta \) plays a well-defined role within the probabilities involved in such a computation.

4. Skeleton extraction

Morphological transformations result from transformations generated by set-based operations and by combinations of dilation and erosion. Morphological Filters (MFs) [7,9] are idempotent and increasing morphological transformations obtained by combining basic operators in a sequential way. MFs form a class of image-processing operators dependent on a single parameter, i.e., the structuring element (B). A particular subclass of MFs, used by Schonfeld [9] for restoration of noisy binary images are Alternating Sequential Filters, (ASFs), which involve sequential applications of erosion-dilation pairs. This author showed that ASFs characterized by longer sequences yield good noise-elimination results, despite their higher computational complexity. However, it is not easy to establish optimal ASFs for a class of noisy images.

Skeleton extraction can be realized through a morphological transformation. ASFs have been proposed for optimal skeleton extraction from noisy images [9]. Here we apply operators of Binary Statistical Morphology to propose a skeleton-extraction method which is based on a Statistical ASF (SASF), i.e., an ASF using statistical operators. The basic element of the SASF is of the simplest type, as it lies in statistical erosion followed by statistical dilation, i.e., a statistical opening.

The reference algorithm for skeleton extraction, say \( SK(X)(B) \), is the one proposed in [9]; it considers as input an image X, and provides as output a representation R(X), i.e., a morphological skeleton.

The algorithm requires the following steps:

a) Initialization: \( n=0 \), \( S_n(X) = X \).

b) Iterations of the following steps:

1) \( S_n(X) = (S_{n-1}(X) \ominus_B B(n)) \)

2) if \( S_n(X) \neq \emptyset \) then

\[ R_n(X) = S_n(X) \ominus_B B(n) \; \text{Stop}; \]

3) \( \gamma_n(X) = S_n(X) \ominus_B B(n) \)

4) \( R_n(X) = S_{n-1}(X) \ominus_B B(n) \)

5) \( n = n + 1 \)

6) if \( n > N_{\max} \) then \( \text{Stop}, \text{else goto1.} \)

The Skeleton is provided as the combination of the representations at the intermediate steps, i.e., \( SK(X)(B) = \bigcup_{n} R_n(X) \).

The algorithm is based on the assumption that the opening of a set, i.e., \( \gamma_n(X) \), is completely contained in X.
4.1 Statistical Skeleton

If we define a binary statistical opening at step \( n \) as
\[ s_{\text{OP}}(X) = (S_{\text{OP}}(X) \circ B(n)) \circ B(n) \]
we can replace steps 1) to 5) of SK with the following expressions:

1) \( S_0(X) = (S_{\text{OP}}(X) \circ B(n)) \)
2) if \( S_0(X) = \emptyset \) then
   \[ s_0(X) = S_{\text{OP}}(X); \text{Stop}; \]
3) \( s_n(X) = S_{\text{OP}}(X) \circ B(n) \)
4) \( S_n(X) = S_{\text{OP}}(X) \circ s_n(X) \)

The statistical skeleton is defined as:
\[ \text{SSK}_X(B) = \bigcup_{n=1}^N s_n(X) \]

It cannot be ensured that the result of statistical erosion is contained in \( S_{n-1}(X) \). The set \( S_n(X) \) which is propagated at step 1) can be larger than \( S_{n-1}(X) \). At the same time, statistical dilation performed at step 3) is not necessary larger than \( S_n(X) \). The statistical opening operation resulting from the composition of the two operators with the same value of \( \beta(n) \) is included "almost everywhere" in \( S_{n-1}(X) \). The only exceptions are bordered subsets of points spatially contained in the convex set obtained from \( X \), but not belonging to \( X \). Such subsets becomes smaller as \( \beta(n) \) increases, due to progressive approximation of statistical opening towards morphological opening.

If noise points form small spatial subsets, it is possible to eliminate them in the first steps of SK. To this end, lower values of \( \beta \) have to be preferred at this stage for their filtering properties. Such filters introduce some not-preserving information (smoothing) effects which depend mainly on the shape of \( B(n) \). During successive iterations noise is supposed to have been eliminated. Consequently, approximation of SK to SK allows one to keep information preserving properties.

4.2 Noise-model parameter scheduling

In general, we let both the structuring element and the \( \beta \) parameter vary through successive iterations. However, in the following, the scheduling of \( \beta \) is discussed for \( B(n)=B \).

The first solution proposed in this paper is an algorithm that uses a fixed \( b \) value, i.e., \( b(n)=b \). This solution can be efficient when large enough \( b \) values are used, i.e., when opening is approximated. The use of a constant, low \( b \) value may cause the algorithm either not to converge to the empty set or to converge slowly. Choosing a higher fixed \( \beta \) value is not optimal in terms of robustness as the stability with respect to increasing noise conditions is damaged.

This led us to develop another algorithm, which is presented here. The algorithm consists in applying a scheduling law to regulate \( \beta(n) \) through various iterations. In particular, the scheduling law is chosen such that \( \beta(0)=0 \) and \( \beta(n) \rightarrow \infty \) for \( n \rightarrow \infty \). The theoretical reason for this choice lies in the analogy of the algorithm to the application of Simulated Annealing [10] to optimization problems in Markov Random Fields (MRFs) [11] approaches. For this reason the algorithm has been called Stochastic Statistical Skeleton, \( S^S_{\text{SS}}(B) \).

In the case of MRFs, it has been shown that the configuration obtained as the estimation of a field of variables is not changed by introducing a further degree of freedom, i.e., the field temperature. Furthermore, an appropriate scheduling of the temperature, joined with successive optimization processes, allow one to obtain the configuration corresponding to the Maximum A-Posteriori estimate of the field. The cooling process consists in obtaining a solution at a certain temperature and then in using this solution as a starting point for reiterating the optimization process at a lower temperature. A decreasing logarithmic scheduling of temperature is usually adopted to ensure a sufficiently slow change (through iterations) in the shape of the probability density function to be maximized. Also in this case the parameter \( \beta \) is used, which can be interpreted as the inverse of the temperature.

When applied to our problem, this strategy requires that different filters be used to extract representations at different steps of the algorithm. The behavior of a filter changes from a linear filter to an opening, after passing through rank-order filters. Linear filtering, corresponding to \( \beta=0 \), obtains noise elimination, and the filters used at the following steps progressively increase the convergence speed towards an empty set. Overall, the algorithm is characterized by greater stability under noisy conditions, mainly thanks to the smoothing effects of the first steps. Moreover, the convergence speed is increased with respect to the fixed \( \beta \) values, thanks to increasing \( \beta \) value through successive iterations.

5. Results and conclusion

Figure 1 shows a reference image and a image from a sequence, which are used by a change detection method [13] to focus attention of a surveillance system (see Figure 2). Figure 2 also presents the result of SK with logarithmic scheduling. A square 3x3 structuring element has been used.

Figure 3 shows results of SK with fixed \( \beta \) values. SK is a special case, i.e., the one at the right. Only points with \( N>5 \) are shown. As one can see for \( \beta=0 \) no skeleton is obtained, due to the existence of fixed points for this value. This fact is pointed out also in Figure 4 where \( S_n(X) \) is represented at successive iterations. As one can see, no change occurs for \( \beta=0 \) after first iteration, due the presence of the fixed point. From Figure 4 one can also notice the increasing convergence speed to the empty set for increasing \( \beta \) values. In Figure 5 \( S_n(X) \) is shown for SK with logarithmic scheduling. An average convergence speed can be noticed with respect to SK for different fixed \( \beta \) values. To estimate robustness of SK under increasing noise conditions, we show in Figure 5 results of SK with percentage of noise added up to 50%. As one can see from Figure 6 the skeleton remains quite stable up to 25% of noise added.

Therefore, results show that the proposed SK algorithm provides a good stability without loosing convergence speed.

References


**Figure 1:** Background image (left) and current image (right).

**Figure 2:** Results of the change-detection algorithm (left) and skeleton extracted by SSK with logarithmic scheduling (right).

<table>
<thead>
<tr>
<th>β value</th>
<th>0</th>
<th>0.4</th>
<th>0.7</th>
<th>1.3</th>
<th>4</th>
</tr>
</thead>
</table>

**Figure 3:** Results of SSK with fixed β value, thresholded at n=5.

**Figure 4:** Convergence of SSK with fixed β value. From top to bottom representations of $S_n(X)$ are plotted for β values of Fig.3. Representations are provided, from left to right, for n=1, 8, 16, 32, 47.

**Figure 5:** Convergence for SSK with logarithmic β scheduling.

**Figure 6:** Negative impulsive noise added to original image with percentages equal to 0%, 10%, 15%, 25%, 50%, from left to right.

**Figure 7:** Results of SSK with logarithmic scheduling for images of Fig.6. The skeleton remains quite stable up to 25% of noise added.
The Euclidean Distance Transform On Curved Space (EDTOCS) With Application To Image Compression

Pekka J. TOIVANEN
Lappeenranta University of Technology
Department of Information Technology
PO Box 20, SF-53851 Lappeenranta, Finland
E-mail: ptoivane@lut.fi

Abstract. In this paper, a new distance transform for gray-level images, called the Euclidean Distance Transform on Curved Space (EDTOCS), is presented. It gives the minimum of all the possible exact Euclidean distances for every point measured from the background. The calculation area does not have to be homogenous. A fast sequential two-pass algorithm to calculate it is given. The EDTOCS is applied to image compression by selecting control points from the gray-level image. The results are good when control points are only in even positions.

1 Introduction

Danielsson [2] invented the Euclidean Distance Transform (EDT), which allows the generation of distance maps with no significant errors. It has been later refined by Yamada [9] and Ragnemalm [5] to generate completely error-free Euclidean distance maps.

Algorithms for 3D with some generalization have been described by Mohr and Bajcsy [3] and Borgefor [1]. Recently, Ragnemalm [6] proposed a number of new algorithms for EDT in 3 and higher dimensions.

This paper presents a new EDT, called the Euclidean Distance Transform on Curved Space (EDTOCS). It gives the exact Euclidean distance transform over a gray-valued image. It is a geodesic transform in the sense that the value of every pixel in the distance image is the length of the shortest path to the nearest pixel in the background.

2 Definition and properties of the EDTOCS

The Euclidean Distance Transform on Curved Space (EDTOCS) between two points is defined as the minimum of all possible paths linking those points. Along this path, each subdistance (and so the whole distance) is Euclidean.

Definition 1. Let \( x \in \mathbb{Z}^2 \) and \( y \in \mathbb{Z}^2 \) be two points in 2-dimensional discrete space and \( X \) be a non-empty subset of \( \mathbb{Z}^2 \). Let \( d(x, y) \) be a function from \( \mathbb{Z}^2 \) to \( R \). \( d_X(x, y) \) denotes the distance value between points \( x \) and \( y \) on set \( X \). Let \( \Psi(x, y) \) be the set of paths in \( X \) linking \( x \) and \( y \). \( d_X(x, y) \) is the minimum of the lengths of the paths \( \gamma \) of \( \Psi(x, y) \) if such paths exist and \(+\infty\) if not.

\[
\forall(x, y) \in X, d_X(x, y) = \min(\Lambda(\gamma), \gamma \in \Psi(x, y))
\]

iff \( \Psi(x, y) \neq \emptyset \). \hspace{1cm} (1)

where \( \Lambda(\gamma) \) is the length of the path \( \gamma \). Here \( d_X(x, y) \) is called the geodesic distance from \( x \) to \( y \) with respect to \( X \).

Definition 2. \( N_C(p) \) denotes the set of the neighbors of a pixel \( P \in \mathbb{Z}^2 \) according to grid \( G \). Let \( X \) be a non-empty subset of \( \mathbb{Z}^2 \). Let \( y \in N_B(x) \) (the 8 neighbors on a square grid). \( \forall(x, y) \in X \), The distance \( d_X(x, y) \) is Euclidean. \hspace{1cm} (2)
Proposition 1. Let $x$ and $y$ be points in $\mathbb{Z}^2$. Let $d_X(x, y)$ be the distance value defined by The Euclidean Distance Transform On Curved Space (EDTOCS) between points $x$ and $y$ on set $X$. This distance satisfies both symmetricity, positive definiteness and triangular inequality:

Symmetricity: $d_X(x, y) = d_X(y, x)$ \hspace{1cm} (3)

Positive definiteness: $d_X(x, y) = 0 \iff x = y$ \hspace{1cm} (4)

Triangular inequality: $d_X(x, z) \leq d_X(x, y) + d_X(y, z)$ \hspace{1cm} (5)

3 The EDTOCS algorithm

The following algorithm is very simple. Only two passes over the image with a chosen kernel are needed.

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>e</td>
<td>f</td>
</tr>
<tr>
<td>g</td>
<td>h</td>
<td>k</td>
</tr>
</tbody>
</table>

Table 1. The $3 \times 3$ kernel used in this paper.

In order to implement the EDTOCS algorithm, two surface models are needed: The original gray-level image, and another, which determines the region or regions in which the transform is calculated. The transform is performed on this image. The part of the surface where the distance function is calculated, $X$, is initialized to maximal representative number of the memory and its complement $X^C$ to 0. It should be noted that the region $X$ in which the following transform is performed does not have to be homogeneous. The algorithm, which applies the EDTOCS, proceeds as follows. Let $G(x)$ denote the original gray-level image and let $F(x)$ denote the binary image which determines the region(s) in which the transform is calculated. $F^*(x)$ means an already calculated point. $F^*(e)$ denotes the new distance value of the point $e$ in the kernel.

1st iteration: The first iteration round proceeds in the "direct video order" (from top to bottom, and from left to right) calculating the new point $F^*(e)$. With 8-neighbor kernel $N_8$ used the iteration proceeds as follows. ($a, b, c, d$) are the neighboring points of $e$ in directions:

$45, 90, 135$ and $180$ degrees.

$$F^*(e) = \min[F(e), \min(da + F^*(a), db + F^*(b), dc + F^*(c), dd + F^*(d))]$$ \hspace{1cm} (6)

where $da = \alpha \sqrt{(G(e) - G(a))^2 + 2}$,
$db = \alpha \sqrt{(G(e) - G(b))^2 + 1}$,
$dc = \alpha \sqrt{(G(e) - G(c))^2 + 2}$ and
$dd = \alpha \sqrt{(G(e) - G(d))^2 + 1}$.

2nd iteration: The second iteration round proceeds in the "inverse video order" (from bottom to up, and from right to left) calculating the new point $F^*(e)$. With the kernel used the second iteration proceeds as follows. ($f, g, h, k$) are points neighboring $e$ in directions:

$0, 225, 270$ and $315$ degrees.

$$F^*(e) = \min[F(e), \min(df + F^*(f), dg + F^*(g), dh + F^*(h), dk + F^*(k))]$$ \hspace{1cm} (7)

where $df = \alpha \sqrt{(G(e) - G(f))^2 + 1}$,
$dg = \alpha \sqrt{(G(e) - G(g))^2 + 2}$,
$dh = \alpha \sqrt{(G(e) - G(h))^2 + 1}$ and
$dk = \alpha \sqrt{(G(e) - G(k))^2 + 2}$.

The above algorithm is in close relation with the Rosenfeld-Pfaltz-Lay algorithm. Features of this algorithm are summarized as follows:

1) It gives the exact Euclidean distance transform over the gray-value image.
2) It only requires two passes over the image.
3) It is easily adaptable to other grids. (For example, the hexagonal grid is quite straightforward).
4) Only two image buffers are needed: The original gray-value image and the binary image which defines the region(s) of calculation.

4 Results

In Eqs. (6) and (7), the differences $da$, $db$, $dc$, $dd$, $dg$, $dh$ and $dk$ are calculated from the original gray-level image. The parameter $\alpha$ governs the amount in which the curvature (difference) is taken into account.
Proposition 2. Let $\delta_X(x, y)$ be the distance value between points $x \in X$ and $y \in X^C$ given by the EDTOCS algorithm if $\alpha = 1$ (or omitted from the equations). If $0.0 < \alpha < 1.0$, the distance $d_X(x, y)$ given by EDTOCS between all points $x$ and $y$ is proportional to $\alpha$

$\forall x \in X, \forall y \in X^C: d_X(x, y) = \alpha \delta_X(x, y)$ (11)

Proof. Consider an arbitrary gray-value image $G(x)$. Let the 3x3 kernel (Table 1.) be at an arbitrary place in the image $F(x)$, which is the distance image, and in the $G(x)$ image, which is the original gray-value image. See 2. Let's assume that the 3x3 kernel lies at $v_8 - v_{22}$.

<table>
<thead>
<tr>
<th>$v_1$</th>
<th>$v_2$</th>
<th>$v_3$</th>
<th>$v_4$</th>
<th>$v_5$</th>
<th>$v_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_7$</td>
<td>$v_8$</td>
<td>$v_9$</td>
<td>$v_{10}$</td>
<td>$v_{11}$</td>
<td>$v_{12}$</td>
</tr>
<tr>
<td>$v_{13}$</td>
<td>$v_{14}$</td>
<td>$v_{15}$</td>
<td>$v_{16}$</td>
<td>$v_{17}$</td>
<td>$v_{18}$</td>
</tr>
<tr>
<td>$v_{19}$</td>
<td>$v_{20}$</td>
<td>$v_{21}$</td>
<td>$v_{22}$</td>
<td>$v_{23}$</td>
<td>$v_{24}$</td>
</tr>
<tr>
<td>$v_{25}$</td>
<td>$v_{30}$</td>
<td>$v_{31}$</td>
<td>$v_{32}$</td>
<td>$v_{33}$</td>
<td>$v_{34}$</td>
</tr>
</tbody>
</table>

Table 2. Pixel values of an arbitrary image $G(x)$.

$F^*(e) = \min[F(e), \min(df + F^*(f), dg + F^*(g), dh + F^*(h), dk + F^*(k))]$

The first term on the right hand side, $F(e)$, is obtained at the previous round:

$F^*(e) = \min[\min[F(e), \min(\alpha \sqrt{(G(e) - G(a))^2} + 2 + F(a), ...)], \min(\alpha \sqrt{(G(e) - G(f))^2} + 1 + F(f), ...)]$

$F^*(e) = \min[\min[F(e), \min(\alpha \sqrt{(G(e) - G(a))^2} + 2 + F(a), ...)], \min(\alpha \sqrt{(G(e) - G(f))^2} + 1 + \min[F(v_{23}), \min(\alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2 + F(v_9), ...)])]$

$F^*(e) = \min[\min[F(v_{15}), \min(\alpha \sqrt{(G(v_{15}) - G(v_8))^2} + 2 + F(v_8), ...)], \min(\alpha \sqrt{(G(v_{15}) - G(v_9))^2} + 2 + \min[F(v_{23}), \min(\alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2 + F(v_9), ...)])]$

Since $F(v_{15}) = F(v_8) = F(v_{23}) = F(v_9) = 255$, it follows that

$F^*(e) = \min[\min(\alpha \sqrt{(G(v_{15}) - G(v_8))^2} + 2 + 255, ...), \min(\alpha \sqrt{(G(v_{15}) - G(v_9))^2} + 2 + 255, ...)]$

Since 255 is the maxint, and maxint + 1 = 0, it follows that

$F^*(e) = \min(\alpha \sqrt{(G(v_{15}) - G(v_9))^2} + 2, ...), \min(\alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \min(\alpha \sqrt{(G(v_{15}) - G(v_9))^2} + 2, ...), \min(\alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \min(\alpha \sqrt{(G(v_{15}) - G(v_9))^2} + 2, ...), \alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

$F^*(e) = \alpha \sqrt{(G(v_{16}) - G(v_9))^2} + 2, ...)]$

Q.E.D.

Corollary.

$F^*(e) = \min[F(e), \min(da + F^*(a), ...)]$, where $da = \alpha \sqrt{(G(e) - G(a))^2} + 2$, is equal to

$F^*(e) = \min[F(e), \min(da + \sqrt{(G(e) - G(a))^2} + 2, ...)]$

Proof. The proof is the same as the proof of Proposition 2.

5 Compression

The EDTOCS was applied to image compression. The algorithm, which applies the transform, proceeds as follows: 1. Create boundary image (binary), mark points with threshold. 2.
Dilate the boundary image. 3. Select a equally distributed points on the image boundary. 4. Initialize the selected control points to maximal representative number of the memory and other points to 0. 5. Do EDTOCS and Assign maximas to new control points while(no more new points).

The maxima of the distance function are selected as new control points. Maxima are points, whose distance value is 1) greater than its neighbors: Height constant defines the difference, 2) greater than Distance constant to avoid unnecessary peaks. Also no control points are allowed to be in the $N_r(p)$ neighborhood of pixel $p$. The algorithm creates an image with control points, each having the right gray value. The control points are further coded by Huffman coding for both their relative position form each other and for their gray value. The decompression is done by the morphological 8-kernels method introduced in citeVep91. The best compression results are achieved if the control points are allowed to take only even positions. Table 2 and Table 3 show the data of the obtained images when only even positions are allowed. Figure 1 depicts the original Leena image of size 512x512x8 bits.

<table>
<thead>
<tr>
<th>Name</th>
<th>Ratio</th>
<th>Points</th>
<th>Pos</th>
<th>Value</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>kB</td>
<td>kB</td>
<td>dB</td>
<td></td>
</tr>
<tr>
<td>Fig.2</td>
<td>1:13.5</td>
<td>22190</td>
<td>2.61</td>
<td>16.30</td>
<td>23.48</td>
</tr>
<tr>
<td>Fig.3</td>
<td>1:18.5</td>
<td>18987</td>
<td>2.38</td>
<td>11.48</td>
<td>22.84</td>
</tr>
<tr>
<td>Fig.4</td>
<td>1:21.4</td>
<td>15541</td>
<td>2.14</td>
<td>9.80</td>
<td>21.73</td>
</tr>
</tbody>
</table>

Table 2. The obtained images.

<table>
<thead>
<tr>
<th>Name</th>
<th>$\alpha$</th>
<th>Dist.</th>
<th>Height</th>
<th>Const.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig.2</td>
<td>1.0</td>
<td>50</td>
<td>170.0</td>
<td>Const.</td>
</tr>
<tr>
<td>Fig.3</td>
<td>1.0</td>
<td>50</td>
<td>175.0</td>
<td>Const.</td>
</tr>
<tr>
<td>Fig.4</td>
<td>1.0</td>
<td>50</td>
<td>180.0</td>
<td>Const.</td>
</tr>
</tbody>
</table>

Table 3. The obtained images.

References


A new approach for scrambling speech signals prior to Analysis-by-Synthesis LPC coding

Dimitris J. HIOTAKAKOS, Costas S. XYDEAS and Colin A. BOYD

Speech Processing Research Laboratory, Electrical Engineering, Dept. University of Manchester, Manchester M13 9PL, U.K., Tel/Fax: +44 (0) 275 4569 / 275 4512, E-Mail: hiotakakos@uk.ac.man.ee.v10
Tel/Fax: +44 (0) 275 4511 / 275 4512, E-Mail: c.xydeas@uk.ac.man
*Communications Research Laboratory, Electrical Engineering, Dept. University of Manchester, Manchester M13 9PL, U.K., Tel/Fax: +44 (0) 275 4526 / 275 4512, E-Mail: boyd@uk.ac.man.ee.shannon

Abstract. In this paper we examine the effect of combined speech coding and encryption as applied to Analysis-by-Synthesis (AbS) LPC based coding techniques. We present a novel approach for scrambling the LPC spectral envelope of the speech signal which is modified prior to the encoding process. In contrast to existing scrambling techniques, this approach takes into account the intrinsic characteristic properties of speech such that the modified signal retains a speech like spectral envelope shape and can thus be encoded more efficiently by the AbS-LPC process. The modified (scrambled) LPC envelope is completely unintelligible.

1. Introduction

Source coding and encryption are linked theoretically by the aim of removing redundancy. However, no attempt has been made, so far, to combine source coding and encryption together, except in lossless compression as applied to text [1-2]. Furthermore, this approach, when applied to speech, can lead into new methods for implementing secure voice systems. In particular, we consider in this paper the issue of secure speech in digital voice systems which involve Analysis-by-Synthesis (AbS) LPC based coders. These coders form a class of time domain speech compression algorithms which are currently widely used in voice communication and in store and forward voice systems [3-5]. A novel Pre-Processing Speech Scrambling Algorithm (PSSA) is proposed, which produces a scrambled signal with speech like characteristics. The resulting signal can then be compressed by a low bit rate speech codec.

The essential difference between PSSA and conventional Analogue Scrambling methods, (ASM) is that the speech signal is modified by taking into consideration its intrinsic properties, within the context of AbS-LPC coding, in a way which effectively retains the compression efficiency of the following coder.

In contrast, common scrambling techniques [6-8] are based on the principle of keyed band splitting and permutation which is applied with no regard to the intrinsic speech properties. In many of these techniques the residual intelligibility left in the encrypted signal is not as low as desired because there is often considerable spectral envelope information still available since the envelope band shapes are not altered [9]. In order to enhance the degree of encryption further, two-dimensional scrambling is normally applied which is based on both time and frequency permutations. The aim is obviously to "randomise" the speech signal and thus render it completely unintelligible to eavesdroppers. However, speech compression algorithms are not effective when applied to random input signals, since they operate on the principle of redundancy removal. In general, there is always a trade-off between scrambling and encoding efficiency.

2. Preliminary Experiments

In order to investigate the effect of combined speech coding and encryption, we distinguish between two different types of experiments. In the first, the 'scrambling importance' of different sets of AbS-LPC codec parameters was assessed in terms of their capacity to reveal the input speech message. In the second type of experiments, each codec parameter is scrambled within the AbS loop, in a keyed manner, and the effect on the recovered speech signal is examined. Notice that although our tests employ a Multi-Pulse Excitation MPE-LPC codec, the investigation strategy adopted here and the resulting conditions can be generalised for any AbS-LPC based codec.

The parameters transmitted in a typical AbS system are:

- The LPC coefficients a(k), where k is the order of the filter, and typically k=10.
- A set of z(n) parameters which effectively define the excitation signal at the input of the LPC synthesis filter. In MPE-LPC, z(n) consists of amplitudes b(i) and positions p(i) defining the excitation pulses. In a CELP type system, z(n) represents the indexes of fixed and adaptive codebooks and the corresponding gains [4, 5].

2.1 Assessing the 'scrambling' importance of the MPE-LPC parameters.

It is a well known fact that the intelligibility of speech signals is directly related to its formant frequencies and thus to its short-time spectral envelope. If the LPC synthesiser is excited with random noise (Fig 1.i), the resulting speech S_x(n) is of poor quality but is highly intelligible. Therefore, even if all the other parameters are scrambled (or encrypted) the LPC coefficients a(k) alone can give away the content of the transmitted message. On the other hand, if the optimised excitation z(n) excitation is used as the input to the LPC filter (Fig 1.ii) whose coefficients are chosen randomly (but within restricted limits) the synthesised speech signal S_x(n) contains a certain degree of intelligibility. The latter should be expected since the short-term residual r(n) as well as the excitation z(n) signals contain
sufficient information to reveal the corresponding speech message; this was confirmed by informal subjective tests.

\[ z(n) = a(n) \]
\[ z(n) = \text{Random} \]

**Figure 1**

2.2 Scrambling within the ABS loop.

In Fig 2 the schematic diagram of an ABS-LPC codec is shown, where the LPC coefficients \( a(k) \) and the excitation signal \( z(n) \) are encrypted, by \( K_L \) and \( K_E \) respectively, in a keyed manner. In our experiments we allowed simultaneous as well as independent changes for \( a(k) \) and \( z(n) \). In many cases, the quality of the synthesised speech is very poor but still intelligible enough to a hypothetical eavesdropper.

**Figure 2** Combined speech coding and encryption

For example, if the LPC parameters are modified within the ABS loop (but prior to the MPE search), then during the MPE search the optimisation process will try to compensate for the poor (modified) definition of the LPC model, and vice-versa. When both \( a(k) \) and \( z(n) \) are modified and the degree of modification is sufficiently high, then the recovered speech becomes gradually unintelligible, but the compression efficiency of the codec is completely lost.

**Figure 3** Desirable Speech Coding - Residual Intelligibility - Undesirable Speech Coding

The effect of combined source coding and encryption, and, in general, the trade-off between residual intelligibility and coding efficiency is demonstrated pictorially by the curves of Fig 3. The behaviour of combined speech coding and encryption is best represented by curves 'A', and 'B'. In fact, conventional scrambling techniques, which are applied prior to coding follow this undesirable behaviour. Thus it has been decided to develop a method, for modifying the input signal prior to encoding, which would retain some of the signal's essential characteristics so that the compression efficiency of the following coder is maintained and the behaviour indicated by curve 'C' is achieved.

3. Algorithm description

Here we introduce a novel approach for scrambling speech signals. The method takes into consideration the characteristic properties of speech. Since knowledge of only the LPC coefficients \( a(n) \) is sufficient to reveal the transmitted message (Section 2), we decided to explore the possibility of scrambling the original speech signal in a way which aims to preserve the compression efficiency of the following ABS-LPC coder.

The block diagram of the proposed scheme is shown at Fig 4. The original speech signal \( s(n) \) is modified in two stages. In the first stage, the LPC coefficients \( a(k) \) are scrambled in a particular way which ensures that the resulting spectral envelope is speech-like and that intra-frame correlation is preserved. The modified envelope is then imposed on the scrambled residual signal \( r(n) \) (second stage), which is produced by scrambling \( r(n) \) with a conventional technique [8]-[9]. In our experiments, a simple approach of keyed permutation of the DCT coefficients of \( r(n) \) has been used.

**Figure 4**

Modifying the spectral envelope of the speech signal prior to compression is an important issue since this process has the potential to affect considerably the speech coder's compression performance. Furthermore, as we discussed earlier, knowledge of only the LPC coefficients is sufficient to reveal the transmitted message. The algorithm presented here is based on shifting the formant frequencies by a keyed offset. For this task the Line Spectrum Pair (LSP) \( l(k) \) representation of the LSP coefficients [11] is conveniently employed. Our choice for LSPs, not only provides a direct association of coefficients-to-formants but also takes advantage of their special properties which can ensure stability of the modified (scrambled) filter. In particular, LSPs are known for their ordering property which guarantees stability of the synthesis filter when the following condition is satisfied:

\[ l(1) < l(2) < l(3) < \ldots < l(9) < l(10) \]  \hspace{1cm} (1)

Another useful property is that each formant is represented by a pair e.g. \( l(1) \) & \( l(2) \) are associated to the first formant, \( l(3) \) & \( l(4) \) to the second and so on.

For each 20ms input speech frame the LPC parameters are derived from the Burg algorithm. These are converted to LSPs and
quantised (using optimum scalar quantisers for each $k$ coefficient $k=1..10$)\(^1\). The key is updated every $n$-LPC analysis frames. When the key is updated, the first coefficient in each pair, i.e. $l_1(i)$ ($i=1,3,5,7,9$), is shifted randomly by a keyed $K_j$ where $j=12...5$, to yield $l_1(i)$. The second coefficient in each pair assumes then an appropriate quantised value as the algorithm attempts to preserve the "distance" $E_1 = l_1(i) - l_1(i+1)$ between the original LSP values of the corresponding pair, that is $l_2(i) = l_2(i+1) - E_1$.

Furthermore, for a frame $m$ between key updates the values of the LSP coefficients are calculated from those of the previously scrambled frame, i.e. $l^{m-1}_{p,k}(k)$, shifted appropriately according to the difference $E_m = l^{m-1}_{p,k}(k) - l^m_{p,k}(k)$ formed between original coefficients.

All modifications to the LSP coefficients are designed to yield a stable filter. A detailed description of this technique is provided by the flow-chart diagram of Fig 6.

4. Results
The performance of the proposed technique can be addressed in terms of: i) the efficiency of the speech compression (MPE-LPC) mechanism, ii) the residual intelligibility of the speech production model based on the scrambled envelope, iii) "speech-like" and "evolution" (intra-frame correlation) characteristics of the magnitude spectra of the resulting scrambled signal. Experiments have been performed using Key-Updates of 1, (i.e. every frame), 5, 10 & 20 frames.

4.1 Properties of Scrambled Spectral Envelope information.

Figure 7 gives an example of scrambled envelope shapes for a number of 20ms frames of a specimen male speech segment. Notice that: i) the produced envelopes have a speech-like shape, and ii) intra-frame correlation is achieved for KeyUpdates $> 1$ frame; both are desired features in terms of maintaining the efficiency of the following coding process.

4.2 Compression efficiency of the spectral envelope modification method.

When coding is performed at 8 and 9.6 kbps informal subjective tests indicated that the quality of the recovered speech at the output of the system in Fig 4 is reduced. However, this signal quality is significantly better than that obtained from a conventional scrambling process followed by MPE-LPC coding. This is true for all different Key-Update rates considered here, i.e. 1, 5, 10 and 20, although best quality is achieved when Key-Update=20. When only envelope modification is applied (no short-term residual scrambling) the compression efficiency of the coder, and hence the quality of $s(n)$, is increased at the expense of reduced voice security.

4.3 Residual intelligibility of the scrambled speech envelope (Fig 5).

In Section 2 we have tested the LPC filter with random input and found that the synthesized signal is quite intelligible. This experiment was repeated for the modified LPC filter with coefficients $a_k(k)$, (Fig 5) in order to assess the residual intelligibility of the scrambled speech envelope. Subjective tests indicated that $s_n(n)$ is completely unintelligible for all Key-Update rates under consideration.

Figure 6 Schematic description of the FSSA algorithm
4.4 Residual Intelligibility of the overall system

Since the transmitted bit stream is not encrypted, an eavesdropper with prior knowledge of the speech coding scheme used in the system, but without any knowledge of the keys employed, could reconstruct the signal $r(t)$ (Fig 4). Although the latter possesses very low intelligibility, a “trained” eavesdropper can still peak up part of the message, if the scrambling action taken on $r(t)$ at the transmitter is not severe. Increasing the degree of scrambling on $r(n)$ leads gradually, to virtually null intelligibility, but at a significantly reduced coding efficiency.

5. Conclusions

We have introduced a novel Pre-processing Speech Scrambling Algorithm, (PSSA) which explores the natural characteristics of speech signals and can therefore be applied prior to speech compression AbS-LPC based algorithms. This technique ensures high compression efficiency and a secure LPC model and thus provides:

- a smooth speech-like spectral envelope shape
- preserves intra-frame correlation between successive LPC analysis frames
- complete “unintelligibility” of the resulting (scrambled) LPC model.

In order to enhance the security of the proposed method, we have introduced a second stage to the algorithm, which modifies the short-term residual signal $r(t)$. However, in terms of speech compression efficiency, we have shown that known scrambling techniques such as band permutation and/or splitting, are not suitable. In a feature paper we will report on advances for “modifying” the fine spectral structure of $r(t)$ without inflicting significant damage to the encoding process of the system.

Acknowledgement

The authors gratefully acknowledge BTL for their financial support to this work.

References


Figure 7 Specimen speech signal (in 20ms frames) and corresponding modified short term spectral envelope.
(Key Update Points are indicated with a circle surrounding the frame number)
Speech Transmission Over a Channel with a Very Low SNR

Raz SHANI, Dov WULICH and Felix FLOMEN *

Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, 85104, Israel.
Email: DOV@BGUEE.BITNET
* AudioCodes Ltd., Or Yehuda, 60265, Israel

Abstract. Speech transmission over a voice-band channel with a very low SNR, where a standard analog as well as vocoder-based transmission are practically impossible, is considered. A new technique, based on simultaneous transmission through the noisy channel of the (unprocessed) speech signal and 10 LPCs (digitally transmitted) which are used for speech enhancement, is proposed. A Data Over Voice, based on decision feedback, is employed for such a transmission. The performance of the proposed system was evaluated and compared to the standard method, for SNR<4dB, using an intelligibility test. The test was based on a known dictionary of 1/4 different Hebrew words. It was found that the proposed system is significantly better than the standard analog transmission and a vocoder based systems.

1. Introduction.
There are many practical situations in which it is required to transmit a clean speech signal over a noisy channel. Important examples are long-distance communication over a noisy radio channel or cellular radio telephone systems which suffer from channel noise in addition to local background noise. To improve the performance of such a voice communication systems in the sense of minimizing the effect of channel noise (in order to improve perceptual aspects such as quality and intelligibility) usually speech enhancement techniques are used.

There are two places where the speech can be processed in order to be enhanced: (a) in the transmitter to "immunize" the (still clean) signal prior to being affected by the channel noise, or/and (b) in the receiver by using methods which will enhance the signal contaminated by noise.

Speech enhancement which can operate on the clean signal prior to its degradation by noise achieves a significant improvement in intelligibility [ref. 6-13 in 1]. The processing method consists of high-pass filtering followed by infinite amplitude clipping. High-pass filtering is used to enhance the second formant frequency relative to the first formant frequency. Infinite amplitude clipping is used as a technique for increasing the power of the weak speech events (consonants) relative to the weak speech events (vowels) [2].

Enhancement methods at the receiver side are based on knowledge of certain parameters of the clean speech signal. Among a variety of methods, as representative examples we can consider the suppression of non-harmonic frequencies or re-synthesis using a vocoder. The needed a priori knowledge about the signal is the fundamental (pitch) frequency in the first method and the LPC coefficients in the second method [3]. However, when the speech signal is contaminated by noise the parameters are extracted from the noisy signal which leads to poor enhancement performance. A good example to illustrate the problem can be found in the Prolog to [1], where it is clearly stated that "The H/M-based enhancement approach in its current state of development was not found effective for the very noisy input SNR of 5dB".

In this paper, we will examine the simultaneous transmission with the speech signal of a certain amount of vital information on the clean speech signal which is available at the transmitter side and to use it for enhancement at the receiver side. To be more specific we propose the following: (a) to extract some parameters from clean (immunized or not) speech signal at the transmitter; (b) to send this information in a digital form simultaneously with the (analog) speech signal by using a Data Over Voice (DOV) technique similar to that described in [4]; (c) to use the additional information for effective speech enhancement by using one of the known methods.

It should be clearly stated at this point that the method proposed employs the fact that even in a noisy channel which already carries a speech signal there is still a "place" for additional information. Of course, the amount of this additional information depends on the noise level and power constraints of the channel-or, concisely, Signal- to-Noise Ratio (SNR).

2. The statement of the problem and principles of the solution.
We are given a voiceband channel with a very low SNR and we wish to transmit a speech signal through the channel. We suppose that the SNR in the channel is so low that direct transmission of the speech signal will result in very poor intelligibility at the receiver end. It is also assumed that it is
3. Description of the proposed system.

3.1 Transmitter

A block diagram of the transmitter which sends both the voice signal \( v(t) \) and digital data represented by \( s(t) \) is shown in Figure 2. The unprocessed speech signal is summed with a digitally modulated signal which carries 10 LPC coefficients, \( 3 \) pitch gain coefficients and pitch frequency for each 22.5 msec frame (180 samples at \( f_s = 8 \) kHz sampling frequency).

3.1.1 Vector quantization of LPC

The \( 10 \) LPC are firstly transformed into \( 10 \) Line Spectral Frequencies (LSF) \([5]\). The transformation from LPCs to LSFs is reversible, i.e., it is possible to compute exactly the LPCs from the LSFs. As shown in \([5]\) the spectral sensitivities of LSFs are localized, i.e., a change in a given LSF produces a change in the LPC power spectrum only in its neighborhood. The localized spectral sensitivity property of LSFs makes them ideal for split Vector Quantization (VQ). For the needs of this study the LSF vector is split into \( 3 \) parts: the first part contains the first three LSFs, the second part contains the next three LSFs and finally the third part contains the remaining four LSFs. As a result of a trade off between complexity and performance of the split VQ, the total bits available for LSF quantization are divided as follows: the first \( 3 \) LSFs are quantized with \( 8 \) bits, the next \( 3 \) LSFs are quantized with \( 7 \) bits and the last \( 4 \) LSFs are quantized with \( 5 \) bits yielding a total of \( 20 \) bits per frame.

The speech data base used here consists of 30 minutes of speech recorded from a few different FM radio stations\(^1\). The first 24 minutes of the speech (from about 150 speakers) is used for training, and the last 6 minutes of speech (from 33 speakers, different from those used for training) is used for testing. Additional testing was performed on four short records of 12.5 sec: two English speakers (female and male) and two Russian speakers (female and male) to prove the independence of the obtained code-book from the language. The LSF distance measure is the weighted distance \( d(f, f') \) between the test LSF vector \( f \) and the reference LSF \( f' \) and is given by

\[
d(f, f') = \sqrt{ \sum_{i=1}^{10} w_i (f_i - f'_i)^2 } \tag{1}
\]

where \( w_i \) is the weight assigned to the \( i \)-th LSF. \( w_i = [P(f_i)]^r \)

where \( P(f) \) is the LPC power spectrum associated with the test vector as a function of \( f \) and \( r \) equal to \( 0.15 \). The weights \( c_i \) are determined as follows

\[
c_i = \begin{cases} 
1.0, & \text{for } 1 \leq i \leq 8, \\
0.8, & \text{for } i = 9, \\
0.4, & \text{for } i = 10. 
\end{cases} \tag{2}
\]

The optimal size of each of the three obtained codebooks is found according to listening tests and Segmental VNR (SVNR) criteria.

Finally the \( 20 \) bits/frame split VQ is obtained and used for LPC quantization through this paper.

3.1.2 Pitch period

The pitch period varies from 20 to 146 samples in steps of two (the pitch period can be an even number only) which yields \( 6 \) bits/frame.

3.1.3 Vector quantizer of pitch gains

The \( 3 \) gain coefficients are encoded by using a vector quantizer yielding \( 7 \) bits/frame.

\(^1\) All the records we made are in Hebrew.
3.1.4 Digital modulator

A BPSK modulation format carrying 1465bits/sec with a carrier frequency of 2000Hz was used in the experiment which will be described in the next section. However any other modulation technique can be used which is able to carry the necessary amount of information in the given voiceband channel.

3.2 Receiver

The resulting signal \( x(t) \) is sent through the channel and received by the system shown in Figure 3. The receiver consists of three main parts, the digital demodulator, the adaptive canceller and the speech enhancer. The digital demodulator detects the data considering the speech signal as an interference. The adaptive canceller firstly synthesizes the signal \( s_{d}(t) \) by modulating the received data \( \hat{d}_{a} \) yielding \( \tilde{s}(t) \), and then \( \tilde{s}(t) \) is passed through an adaptive filter which is a duplicate of the channel. The output of the filter \( s_{d}(t) \) is then subtracted from the signal \( y(t) \) resulting in \( v_{r}(t)+n_{d}(t) \). This signal and the decoded data \( \hat{d}_{a} \) are fed to the voice enhancer.

4. Performance of the proposed method.

A simulation of the complete system was made in order to evaluate its performance. Firstly we decided which of the speech parameters, LPC or pitch, are more important for our needs. As a result of a preliminary listening test using several listeners we found that the LPC coefficients are more important than the pitch parameters. Consequently, using the LPC alone decreases the bit rate to 888bit/sec and at the same time decreases the BER.

![Figure 4. Block diagram of the enhancer.](image)

The block scheme of the voice enhancer is presented in Figure 4. The received digital data \( \hat{d}_{a} \) is split into 10 LPC coefficients, \( \beta \) gain coefficients and a pitch period which control two serially connected filters. The first has the transfer function:

\[
H(z) = \frac{10}{1 + \sum_{i=1}^{10} a_{i} z^{-i}} \frac{3^{i} z^{-i}}{1 + \sum_{i=1}^{10} g_{i} z^{-i}}.
\]  \( i \) (3)

where \( a_{i} \) \((i=1,2,...,10)\) are the LPC coefficients and the second has the transfer function:

\[
G(z) = 1 + \sum_{i=-L}^{1} g_{z} z^{-i-L}.
\]

where \( g_{i} \) are the gain coefficients and \( L \) is the pitch period.

The filters \( H(z) \) and \( G(z) \) act in fact as a time varying filters which change their parameters every frame. The filter \( H(z) \) fits its shape according to the magnitude value of the spectrum of the speech signal. Such pole-zero filters are used for adaptive postfiltering in low-speed speech coders [6].

993
Table 1

<table>
<thead>
<tr>
<th></th>
<th>-2dB</th>
<th>0dB</th>
<th>2dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>analog system</td>
<td>1.58</td>
<td>1.55</td>
<td>1.75</td>
</tr>
<tr>
<td>our system</td>
<td>1.74</td>
<td>2.46</td>
<td>2.61</td>
</tr>
<tr>
<td>% of listeners that preferred our system</td>
<td>36%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

It can be seen that for all 3 SNRs the performance of our system is significantly better.

For SNR=4dB a sentence of 12sec (recorded from FM radio) transmitted in a standard way and processed by our system was also presented to the listeners. All listeners decided that it was much more comfortable to listen to the output of our system.

In addition, the system was qualitatively checked on four short records of two English speakers and two Russian speakers (female and male for both languages). The results were practically the same as for Hebrew speakers including speaker identification for all three above mentioned SNRs.

5. Comparison with vocoder based transmission.

As was mentioned in section 3 the speech signal can be sent digitally by using a vocoder. Here we will discuss this matter on the basis of existing literature focusing our efforts on the behavior of the vocoder when there are channel errors.

Using a 4.8kbit/sec CELP vocoder guarantees good intelligibility and quality but under the assumption that there are no channel errors. When error protection is employed the total bit rate increases to 6.4kbit/sec (1.6kbit/sec is used for channel coding). Such a combined source (vocoder) and channel coder provides good quality speech for channel BER as high as 10^-2 [7]. For SNR<4dB the use of such a vocoder is theoretically impossible. The capacity of a voice-band channel for SNR>2dB equals 4.26kbit/sec.

There has been an attempt to transmit a speech signal over a noisy channel using an 800bps vocoder with error protection to 1200bps [8]. This vocoder however suffers, according to [8], from low intelligibility, language dependency and lack of speaker identification.

Consequently, it can be concluded, that the proposed method is better than the vocoder-based approach.

6. Conclusion.

Speech transmission over a voice-band channel with a very low SNR, using both standard analog and vocoder-based transmission is practically impossible. A new method has been proposed where the unprocessed speech signal is transmitted simultaneously with 10 LPCs which are digitally transmitted. A Data Over Voice technique based on decision feedback is used for the simultaneous transmission. The LPCs are used at the receiver end as the parameters of a time-varying filter which enhances the speech signal relative to the noise.

The method was compared to two potential competitors: (a) standard analog transmission over a noisy channel, and (b) a vocoder based transmission where the speech signal is synthesized from some coefficients which are extracted at the transmitter side and transmitted digitally over the noisy channel.

Comparison to the standard transmission based on the comparative intelligibility test resulted in the clear conclusion that for SNR<4dB the proposed method is better. Comprehensive comparison to the vocoder is very difficult to perform due to the large number of existing methods and techniques. Moreover, as a rule vocoders are designed for channels with a high SNR and consequently it is difficult to find, in the literature, a deep study on how the transmission errors influence the performance of the vocoder. Because of these reasons the comparison is based on two works where the problem of channel errors is addressed. We can conclude that using a vocoder which guarantees good quality and intelligibility is theoretically impossible as the channel capacity for the considered SNRs is lower than the vocoder’s bit rate. On the other hand the proposed method is better than the vocoder which is designed to work with a noisy channel because of two relative advantages: language independence and the ability to convey the identity of the speaker.

References

THE DANISH CANDIDATE FOR THE GSM HALF-RATE SPEECH CHANNEL

Keld B. Mikkelsen, Knud J. Larsen*,
Henrik B. Hansen, Henrik Nielsen, Henrik Jakobsen

Tele Danmark Research, Hørsholm, Denmark, E-mail: keld@tdr.dk
*Institute of Circuit Theory and Telecommunication, Technical University of Denmark.

Abstract: This paper describes the results of the optimisation of the Danish GSM half-rate candidate since the second selection phase of the half-rate codec standardization. The optimisation has focused on improving the basic speech quality as well as robustness to transmission errors. The basic speech quality was improved mainly by means of increased adaptive codebook resolution and output speech postfiltering. The channel coder performance was improved significantly by changing the bit allocation and by exploiting speech coder data statistics during the Viterbi decoding.

1 Introduction

During the fall of 1992, four candidate coders from CSELT/AT&T/PRI, ANT, Motorola, and Telecommunications Research Laboratory participated in the second selection test for the GSM half-rate standard. All coders fulfilled the preselection requirements, and the ANT and the Motorola candidates were selected to continue in the standardisation process. From the selection test results we concluded that further improvement of our candidate should be possible both for speech and channel coders.

This paper deals with joint optimisation of the speech and channel coders. The speech and channel coders are slightly modified versions of the candidate submitted for the selection test [1].

2 Speech Encoder

The selection test version of the speech coder is based on the 5.4 kb/s speech coder presented at ICASSP [2]. The most significant modification is that the stochastic codebook excitation contribution is constrained according to a strategy similar to the one described by Shoham [3]. Constrained excitation reduces the roughness which is typical of voiced output speech in CELP coding.

In the latest version, the adaptive codebook covers 128 integer pitch delays, and in the first half of the codebook the pitch delay resolution is increased to 1/3 of a sampling period. This modification improves the quality of voiced speech, especially for high-pitched voices, while maintaining the bit allocation of the previous version. Quantization of the spectral parameters is based on LSP differences rather than absolute values, and a perceptual weighting technique is incorporated in the quantiser design [4]. The constrained excitation strategy is changed such that the excitation gain is adjusted outside the codebook search loop. This is a simpler strategy offering similar performance, and the additional complexity is negligible compared to the codebook search.

The information bits delivered by the speech encoder are rearranged according to their individual robustness characteristics and perceptual importance and divided into two protection classes.

3 Channel Coder

The characteristics of the GSM channel and the TDMA system used in GSM are described elsewhere, e.g. in [1]. For the present purpose, it is sufficient to know that various kinds of interleaving may be used to reduce the correlation between adjacent bits. Unfortunately, the bursty character of the errors cannot be completely removed by interleaving. After several experiments [1], we have found it advantageous to use convolutional codes for the channel coding, even though these codes are actually best suited for correction of random errors.

In the next section, we shall describe the codes chosen. Then we proceed by describing various enhancements to the conventional Viterbi decoding algorithm.

3.1 Convolutional Encoder

As with all error correcting codes the performance of a convolutional code is dependent on the rate and the minimum distance of the code. For these codes the minimum distance is not easily defined, but the so-called free distance is an important parameter for the perfor-
formance. The free distance is the minimum weight of any non-zero path leading from the zero-state and back again. As mentioned in Section 2, the data from the speech encoder exhibit different degrees of sensitivity to transmission errors, and they are grouped into two protection classes. We have used the different requirements of the two classes to reduce the overall rate of the convolutional encoder by varying the code rate along the frame. In order to overcome difficulties that may arise if the code were completely changed between the two classes, we have used rate compatible puncturing of a basic code [5]. The basic code has rate 1/3 and is punctured to rate 2/5 in the most protected class and to 1/2 in the other class.

Even though the free distance is not directly related to the performance of the total system we have chosen to use this parameter as an optimisation criterion. It turned out that the best rate 2/5 and 1/2 codes were not always from the best 1/3 code, so we have chosen the basic code to be one of the codes with maximum free distance for rate 1/2 and among them we have found the best for rate 2/5. Table 1 shows generators and free distances for convolutional codes with different encoder memory. As it may be seen, increasing memory increases the free distance, but unfortunately complexity increases almost exponentially with memory.

<table>
<thead>
<tr>
<th>memory</th>
<th>generators (hex)</th>
<th>free distances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>05</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>1D</td>
<td>2B</td>
</tr>
<tr>
<td>4</td>
<td>2B</td>
<td>3D</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>79</td>
</tr>
<tr>
<td>6</td>
<td>2B</td>
<td>79</td>
</tr>
</tbody>
</table>

Table 1: Generators and free distances for different convolutional codes.

In Table 2 we show the performance of the codes from Table 1 with respect to the ratio of error free frames and the bit error rate after decoding. The performance shown is for one of the GSM test patterns. These patterns were described in [1, 8]. The decoder will be described in the next section. It turns out that the decoded bit error rate actually increases with larger free distance, but on the other hand the number of error free frames increases as well despite the higher error rate. This may appear strange but as mentioned above, interleaving only partly removes the bursty character of the errors. Therefore, the code is temporarily overloaded in bursts with many errors, and in such cases a "good" code performs worse than a normally inferior code. From these results, it may be concluded that a memory of four is a good choice.

The coding is performed in frames, which leaves a decision on how to end the frame. One way is to append a tail with as many 0-bits as the memory but this leaves less room for speech data, or it requires further punctuation of the code. Another way is to let the decoding select the path with the best metric and use no special tail. We have investigated the two possibilities, and

<table>
<thead>
<tr>
<th>memory</th>
<th>decoded bit errors</th>
<th>error free frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.53%</td>
<td>89.2%</td>
</tr>
<tr>
<td>3</td>
<td>1.80%</td>
<td>90.1%</td>
</tr>
<tr>
<td>4</td>
<td>1.95%</td>
<td>90.9%</td>
</tr>
<tr>
<td>5</td>
<td>2.06%</td>
<td>91.1%</td>
</tr>
<tr>
<td>6</td>
<td>2.11%</td>
<td>91.3%</td>
</tr>
</tbody>
</table>

Table 2: Channel decoder performance for GSM test pattern EP8.

found no difference in performance when averaged over each frame. However, leaving out the tail allows us to increase the most protected class of bits, and in total this gives a better result when used in connection with the speech encoder.

### 3.2 Enhanced Viterbi Decoder

The optimal method for decoding convolutional codes is Viterbi decoding, but the basic algorithm may be improved in several ways to suit the application better. Normally, a Viterbi decoder assumes that the coded data have equally probable bit values. Here we shall present a modification that allows the decoder to use information about unequal probabilities to assist the decoding and thus obtain the data sequence with maximum a posteriori probability. We assume that the de-modulator provides the log-likelihood ratio as an indication of the reliability of each received symbol \( s_i \). Then the optimum branch metric for the branch corresponding to the data symbol \( s_i \) is:

\[
\sum_{i=0}^{n-1} \left[ \log P(z_{i+1} \neq y_{i+1}) - \log P(z_{i+1} = y_{i+1}) \right]
\]

where \( y_{i+1} \) are the coded symbols, and \( n \) the number of coded symbols on the branch. Although the indices may indicate that each branch has the same number, \( n \), of symbols, this is not the case when puncturing is used for the branch. In such cases, the sum is over the reduced number of symbols on the branch, but this fact is left out of the formula in order not to complicate the notation further. The probabilities \( P(s_i) \) are found from analysis of the correlation between frames. The improvement of decoder performance using this technique is shown in Table 3.

In order to assist the speech decoding, the Viterbi algorithm is enhanced further to deliver not only the most likely path sequence, but also reliability information together with the output bits. The reliability information is a log-likelihood function, and the method used to produce the log-likelihood output is described by Hagenauer et al. [7]. The calculation of the reliability measure increases the complexity of the algorithm by approximately a factor of two. The algorithm by Hagenauer et al. does not produce the optimum a posteriori probabilities for each bit to be decided if the purpose
<table>
<thead>
<tr>
<th>error pattern</th>
<th>EP1</th>
<th>EP2</th>
<th>EP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C/I</td>
<td>10 dB</td>
<td>7 dB</td>
<td>4 dB</td>
</tr>
<tr>
<td>channel bit error rate</td>
<td>4.5%</td>
<td>8.3%</td>
<td>13.4%</td>
</tr>
<tr>
<td>decoded bit error rate</td>
<td>0.45%</td>
<td>1.95%</td>
<td>6.6%</td>
</tr>
<tr>
<td>improvement</td>
<td>18%</td>
<td>14%</td>
<td>12%</td>
</tr>
<tr>
<td>frame error rate</td>
<td>2.5%</td>
<td>9.1%</td>
<td>28%</td>
</tr>
<tr>
<td>improvement</td>
<td>13%</td>
<td>11%</td>
<td>7%</td>
</tr>
</tbody>
</table>

Table 3: Channel decoder performance for memory four and improvement due to the use of unequal probabilities in decoding. A frame containing one or more errors is referred to as a frame error.

is to minimize the bit-error probability. However, the optimum algorithm for this is very complicated [8] so we have chosen the less complicated modification of the Viterbi algorithm.

The method for calculation of the reliability measure is constructed to give the opportunity to only calculate the measure for a reduced number of bits. The speech decoder actually uses soft decision information for less than half the received bits. This means that complexity is reduced by a similar amount. The next section will describe the benefits of the technique.

4 Speech Decoder

Previously, quantization of the absolute LSPs was considered more robust to residual channel errors, because a few errors usually distort only a limited portion of the spectral envelope, whereas with differential quantization, errors may propagate into other parts of the spectrum. However, with proper rearrangement of the error protection and error concealment, this problem can be alleviated such that differential quantization appears to be subjectively at least as robust as absolute quantization.

At the decoder output an adaptive postfilter \( H(z) = (1 - \mu z^{-1})A(z/\gamma)/A(z/\alpha) \), where \( A(z) \) is the LPC filter, is introduced. The 1st-order section compensates for the additional spectral tilt caused by \( A(z/\gamma)/A(z/\alpha) \), and the output signal level is equalized to compensate for the gain of \( H(z) \).

4.1 Bad Frame Handling and Error Concealment

The availability of soft decision output from the channel decoder facilitates bad frame handling and error concealment. Using the soft decision information, error concealment techniques are applied to the individual speech coder parameters, forming a more refined strategy compared to simple frame substitution. This means that the speech decoder stands a better chance of extracting useful data from frames with relatively few errors, which might otherwise have been discarded.

Now, the problem is how to transform bit level soft decision information into a meaningful reliability measure for the speech coder parameters. For data transmission it might be useful to estimate the probability of the received parameter being identical to the transmitted one, but for speech coder parameters it is often more important to know whether the value of the received parameter is likely to be far away from or close to that transmitted. To try to capture this and since the most significant bits of a parameter are usually better protected and thus more reliable, it was decided to simply ignore the soft decision of the \( k \) least significant bits. The reliability of parameter \( p \), which is allotted \( n \) bits, is then defined as

\[
r(p) = \sum_{i=k}^{n-1} s_d(i)p_2^{i-k},
\]

where \( s_d(i) \) is the soft decision of the \( i \)th bit within \( p \). The reliability measures for the spectral parameters are combined into a common figure which controls the parameter interpolation and substitution. For the adaptive codebook parameters and the stochastic codebook gain factor smoothing is applied to the individual parameters whereas the stochastic codebook index is left unmodified.

Due to the recursive structure introduced by the adaptive codebook it is important to control the gain of the adaptive codebook loop. The adaptive codebook gain factor may temporarily exceed unity under normal conditions but when the received excitation parameters are unreliable the gain should be kept below one in order to avoid loud unpleasant sounds. Furthermore, the content of the entire adaptive codebook is scaled by a factor of 0.5 if the parameters are unreliable. This means that the adaptive codebook contribution will vanish if bad excitation parameters are received for several consecutive frames, and when the channel fade has ended the codebook will be restored fairly slowly.

5 Results

The described candidate codec was tested in a formal subjective test in Danish language similar to the preselection test. The preselection test is designed according to the basic qualification requirements for the GSM Half-rate Traffic channel [6]. The test conditions which will be used to check if the requirements are met by the candidate are also given in [6].

<table>
<thead>
<tr>
<th>error pattern</th>
<th>no errors</th>
<th>EP1</th>
<th>EP2</th>
<th>EP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR (MOS)</td>
<td>3.61</td>
<td>3.18</td>
<td>2.79</td>
<td>2.77</td>
</tr>
<tr>
<td>HR.92 (ΔMOS)</td>
<td>-0.82</td>
<td>-0.51</td>
<td>-0.53</td>
<td>-0.72</td>
</tr>
<tr>
<td>HR.93 (ΔMOS)</td>
<td>-0.46</td>
<td>-0.39</td>
<td>-0.34</td>
<td>-0.23</td>
</tr>
</tbody>
</table>

Table 4: Subjective test results. ΔMOS is the weighted average MOS difference between the half-rate (HR) candidate and the full-rate (FR) codec for A-law quantized, IRS filtered speech at three input levels.

The test showed a statistically significant speech quality improvement compared to the selection test version.
In Table 4, ΔMOS results for A-law [9] quantized and IRS [10] filtered speech input are shown. ΔMOS is the difference between the Mean Opinion Scores [11] of the half-rate candidate and the full-rate reference, and the ΔMOS is a weighted average over three different input levels. The scores for the error free condition, EP1, and EP2 were collected using the quality rating scale and those for EP3 with the listening effort scale. In the table, HR.92 refers to the selection test version and HR.93 is the improved candidate. It can be noticed that for HR.93 the quality difference becomes smaller as more transmission errors are introduced on the channel, which is not the case for HR.92.

6 Discussion

The recent modifications of the speech and channel coders improve the basic output speech quality and especially the robustness against transmission errors compared to the selection test version of the candidate. The channel coding scheme has been optimised with respect to bit allocation, and the decoding algorithm is enhanced using interframe correlation. These modifications increase the error correction capability, improving both the output bit error rate and the frame error rate. In addition to the output bit sequence, the channel decoder calculates soft decision information as well. The soft decision output is utilized in the speech decoder for error concealment purposes by calculating a reliability factor for each parameter. This eliminates the need to calculate a bad frame decision and to transmit CRC check bits. The combined effect of lower error rates and improved error concealment is that the output speech can be more effectively smoothed or extrapolated in error situations. Moreover, error concealment based on soft decision has the advantage that smoothing may be disabled when not needed, ensuring that speech quality is always preserved under good channel conditions.

The subjective test showed that the half-rate coder is now relatively more robust to channel errors than the full-rate coder. Concerning the speech quality under good channel conditions the test also shows that the basic quality is still slightly below the required performance. Therefore further optimisation or more bits allocated for the speech coder is needed. Taking the previous optimisation phase into account, it seems that this particular speech coder structure may not have the potential at 5.4 kb/s to fully match the quality of the full-rate reference, even though a coder of similar structure has proven capable of providing sufficient basic quality near 6.5 kb/s [12, 1]. The bit allocation for the speech and channel coders is a major topic. With the recent advances concerning robustness, it is possible that a better overall solution could be achieved if more bits were allocated to the speech coder, but it could also be necessary to change the structure of the algorithm.

References


DIAMANT

All Digital Frequency Division Multiplexing for 10 Gbit/s Fibre-Optic CATV Distribution System

Heinz G. GÖCKLER and Karlheinz GROTZ
ANT Nachrichtentechnik GmbH, D-71520 BACKNANG, Germany
Tel/Fax: +49 7191 133836 / 134321, E-Mail: Heinz.Goecker@Bk.Bosch.DE

Abstract. The DIAMANT system is being developed for the distribution of up to 64 TV channels (or with some of those channels replaced by an equivalent number of FM, DSR or DAB signals) via a single optical fibre. The channels are efficiently combined to a digitised Frequency Division Multiplex (FDM-PCM) signal, allocated in the standard frequency range below 446 MHz, by sophisticated digital processing of complex-valued sequences in two-stage digital frequency converters. Any group of FDM channels can be replaced by a multitude of future digitally compressed source signals introduced in a Time Division manner (TDM/FDM-PCM). Thus, a subscriber interface essentially consists of a D/A converter and, if applicable, a TDM demultiplexer.

Following advantages of DIAMANT over existing techniques are claimed: i) Higher fibre lengths and splitting factors ii) Higher and reproducible performance iii) High flexibility: reconfigurable channel allocation, sliding transition from FDM to TDM format, apt to interactive services iv) Modularity with impact on extension, design, (ASIC-) realisation and production v) Low power consumption and vi) Low expenditure.

1 Introduction

Present-day optical distribution of TV, FM and DSR channels applies analogue intensity or frequency modulation of the optical carrier by the continuous frequency division multiplex (FDM) signal, and appropriate detection by a photodiode at the receiver front end. However, nonlinearities of the optical modulator or laser diode, frequency shift (chirp) of the optical carrier during modulation, dispersion of the optical fibre and shot noise put severe limitations on the maximum fibre length and splitting factor for a prescribed signal-to-noise ratio.

Hence, a digital (PCM) format of the overall FDM signal lends itself for transmission via optical fibre. Thus, regeneration of the optical signal is possible, allowing to increase the maximum fibre length and splitting factor to almost any desired value. Furthermore, besides the improvement and reproducibility of signal quality due to digital signal processing, a gradual transition from today's analogue channels to future digitally encoded (compressed) source signals is feasible by mere data (de-)multiplexing at the fibre interfaces; Fig.1. In

Figure 1: DIAMANT (TV-)distribution system: FDM-channels time-multiplexed with (future) TDM-channels; DFC (AFC)/DFM: Digital (Analogue) Frequency Converter/Multiplexer, MOD: Modulator, E/O/O/E: Electrical/Optical Interfaces, (DE)MUX: Time Division (De-)multiplexer

either case the bulk of signal processing is per-
formed in "hub" stations, i.e. removed from the subscriber.

2 DIAMANT CATV System

The various digital frequency converters (DFC) of a digital frequency multiplexer (DFM), located at the hub stations of an optical fibre ring or tree structure, are fed with individual channels to be distributed in an FDM manner. These channels are provided as PCM signals (Fig.1)

- either by digitisation (A/D: analogue-to-digital conversion) of baseband signals (video and sound signals, Stereo-MPX signals or DSR data stream) followed by modulation to the desired representation (VSB with FM sound carriers, FM or QPSK, resp.) at an appropriate intermediate frequency (IF)
- or by direct digitisation of the modulated signals at a suitable (standard) IF.

The A/D conversion may be performed either at the hub station calling for an analogue air or cable interface, or removed from the hub station calling for a digital transmission means such as the Synchronous Digital Hierarchy (SDH).

The above features of a DFM are depicted in Fig.1 for TV channels, including also future digital encoding (not to be considered in the following). For mere FDM transmission the only interface at the subscriber premises (or curb etc.) is a fast and highly linear D/A converter (cf. [1]): 896 MHz sampling rate and 11 bit resolution resulting in an overall transmission rate 9.856 Gbit/s of the DIAMANT system.

Contrary to the far end, in the hub station A/D conversion is performed for each channel individually, since

- high speed and high resolution A/D converters are not expected to be available with the required linearity in the near future
- various features (performance, reproducibility, manufacturing etc.) can be improved considerably by shifting the transition from analogue to digital signal processing as close as possible to the generation of the source signals
- flexibility in channel allocation can be gained by (remote) programming of the DFCs.

The challenge, how to design and realise a versatile (flexibly programmable) and relatively low-cost digital frequency multiplexer (DFM) suitable for

- sampling modulated TV, FM or DSR signals at their standard IF, with adequate oversampling capability to diminish the constraints of the analogue antialiasing filter in front of the A/D converters [2]
- shifting the channels to their ultimate frequency position
- interpolating the signals to their final output sampling rate close to 1 GHz
- and combining the individual signals to the digitised FDM signal

is dealt with in the next section.

3 Digital Frequency Multiplexer

A complete (TV-)DFM is depicted in Fig.2. It consists of a cascade of two stages:
i) After A/D conversion, the first stage represents an individual digital frequency converter (IDFC) for one channel. It performs complete rejection of the contiguous channels (BF), and the transition from real to complex (analytical) signal representation [2, 3, 4] (CBF), thus allowing to frequency-shift the channel to be processed to the ultimately desired frequency slot by a complex mixer (CMI) without producing any additional unwanted spectral replica of the signal. Finally, each IDFC increases the sampling rate by a factor of four (C4BF) followed by the superposition of eight adjacent complex-valued channel signals (CA).

In the first stage all operations are carried out at the A/D converter sampling rate of 28 MHz, since the interpolation filter C4BF with complex-valued coefficients is implemented as a polyphase filter [5, 6]: The fourfold increase of the sampling
of the linear-phase FIR (nonrecursive) type being easily realisable due to its modular structure [2, 3]. The importance of linear phase in image processing filters is stressed in [9]. Furthermore, these filters are most suitable for sample rate alteration, are unconditionally stable, and can be designed for low noise contribution with only moderate increase of filter inherent signal wordlength.

For better understanding of DFM operation a closer look is taken at the TV-IDFC of Fig.2a,b:

A: Each IDFC is fed with a modulated TV-VSB signal in spectrally reversed position (centred at the standard analogue TV IF of 36.4 MHz) and sampled at a rate of 28 MHz. Thus, the standard TV signals (channel spacing: 7 MHz) are oversampled by two, while the TV spectra are frequency-shifted to the digital IF \( f_o = 19.6 \) MHz and spectrally inverted to normal position.

B: Aliased spectral remains in the gaps due to insufficient analogue antialiasing filtering are rejected by the bandpass filter BF with real coefficients.

C: The transition from real to complex signal representation is most efficiently performed by a complex halfband filter CBF [2, 4]. Since its passband is centred at \( 3f_s/4 \approx f_o \), the computational burden of CBF is identical to that of the associated halfband prototype with real-valued coefficients. Thus, the desired spectrum centred about \( f_o \) is retained.

D: As a consequence of CBF filtering, the resulting complex-valued signal can freely be frequency-shifted to any desired centre frequency \( f_c = \) (mod \( f_s \)) by mere multiplication (i.e. without additional filtering) of the complex CBF output sequence by the samples of the complex carrier exponent \( j2\pi k(f_c - f_0)/f_s \).

E: The interpolating polyphase filter C4BF selects the desired portion of the signal and rejects all unwanted replica in the frequency range \([0, 0.4f_s = 112 \) MHz\). For minimum computational burden the complex C4BF is derived from a fourth-band prototype lowpass filter [10] by frequency-shifting its passband centre frequency to \( f_c \). Thus, one of the four branch filters of the C4BF polyphase implementation contributing to the real part of the output signal degenerates to a mere delay with one multiplication only, whereas its imaginary counterpart completely vanishes.

F: The FDM group signal is formed by adding

This two-stage approach is imposed by technology. However, it also represents a balanced compromise between maximum flexibility of channel allocation (calling only for IDFCs), and minimum computational complexity (calling only for filter bank methods, see e.g. [5, 7, 8]). All filters are
the various IDFC output signals at their final frequency positions.

4 DFM Design

By applying the approach of [11] for distortion analysis, spectral distortion caused by adjacent channels, and noise contribution due to signal (re-)quantisation were determined for a 64-channel TV-DFM (Fig.2). As a preliminary result, an overall unweighted signal to distortion ratio of 59.4 dB is obtained for a modulated TV channel, being based on following assumptions (cf. [1]):

i) All 64 channels are loaded by TV signals of identical power, uncorrelated with each other
ii) ADC (DFM input) signal wordlength: 10 bit
iii) DAC (DFM output) signal wordlength: 11 bit
iv) Input/output signal wordlength of signal processing (filter/CMI) blocks: 14 bit
v) Filter stopband rejection: 75 dB
vi) Peak factor: \( p_f = 1.52 \).

The above result is mainly governed by the ADC and DAC signal wordlengths and the peak factor. The filter stopband rejection of 75 dB is guaranteed with linear-Phase FIR filters of length N (Fig.2):

i) \( N_{BF} = 127 \), where the stopband attenuation is complemented to 75 dB by the analogue antialiasing filter in front of the ADC
ii) \( N_{CBF} = 27 \): 15 multipliers
iii) CMI (complex mixer): 4 multipliers
iv) \( N_{CABF} = 19, f_s = 28(12) \) MHz: 58 multipliers
v) \( N_{COBF} = 71, f_s = 112(896) \) MHz: 127 multipliers.

The maximum required coefficient wordlength does not exceed 16 bit.

5 Conclusion

The above results have been verified by simulation and measurement. Meanwhile the design of VLSI filter implementation is in full progress, aiming at a minimum number of different chips for the realisation of the various digital filters [6].

The 5 Gbit/s version of DIAMANT (for 32 TV-channels) will be available mid 1995.

References


Transmultiplexing by Recursive (IIR) Polyphase Structures for Onboard Processing (OBP) Satellites

Gholam R. DANESFAHANI, Tony G. JEANS and Barry G. EVANS

Centre for Satellite Engineering Research, University of Surrey, Guildford GU2 5XH, England, U.K., Tel/Fax: +44 (083) 300800/ 259504, E-Mail: eep1gd@ee.surrey.ac.uk

Abstract. In this paper we make a comparison between recursive (IIR) and nonrecursive (FIR) transmultiplexing techniques. The computational complexity, transmultiplexing delay and bit error rate (BER) performance in the presence of the channel non-linearities and finite word length are reported. A binary tree structure has been used for the channelisation process. At every stage of the structure, the input spectrum is halved. The design of the IIR and FIR prototype filters are based on the Ansari – Liu and Parks – McClellan algorithms, respectively. Simulation results show that IIR transmultiplexers require less multipliers and adders, less delay and have a better noise performance. Furthermore, they are highly modular and are, hence, more attractive for ASIC implementations. These results suggest that IIR transmultiplexers are suitable for OBP applications.

1. Introduction

An attractive multiple access scheme for use with OBP satellites is single channel per carrier/ frequency division multiple access (SCPC/FDMA) for the up-link and time division multiplexing (TDM) for the down-link [1]. The translation from SCPC/FDMA to TDM is performed onboard the satellite in a multicarrier digital demodulator (MCDD) whose other main function is the demodulation of each channel. The subject of transmultiplexing by FIR filters has been extensively covered in the literature; see [2] – [4] and their references. The choice of FIR filters has been dominated by their linear phase response and a constant group delay which can not be achieved by the conventional IIR design techniques. However, the IIR filter design algorithm of [5] can be used to design filters which have an approximately linear phase response and show a reasonably small variation of the group delay in the passband. The filters designed by this algorithm have been used as halfband filters (HBF) in a tree structure to demultiplex the received channels. To compare the performance of the developed transmultiplexer (TMUX) with an FIR type TMUX, we have considered the work carried out in [4] in which the ASIC implementation of an 8 channel TMUX using FIR filters has been reported.

2. System Architecture

The simulated system consists of 8 frequency division multiplexed QPSK channels, a travelling wave tube amplifier, an up-link subject to AWGN, and a receiver with an MCDD which is followed by BER counting models. The QPSK channels are statistically independent and uniformly spaced in the frequency range 0 ≤ f ≤ 8W, where W is the channel spacing and is taken to be equal to twice the symbol rate R(= 1/T). Carrier phase and symbol timing errors introduced by the TWTA are corrected by feedback synchronization loops in the MCDD.

Referring to Figures 1 and 2, the IIR TMUX demultiplexes the received SCPC/FDMA signal into the individual channels in 3 stages. At every stage the spectrum is halved; the side channels are in one half and the middle channels will be in the other half. For filtering the side channels no spectral shift is required. However, for filtering the middle channels, the spectrum is shifted by −0.5f_s^n, where f_s^n is the sampling frequency at the n-th stage. After filtering, the FDM signal is decimated by a factor of 2 in HBFs. For the correct operation of the MCDD, it is necessary that the channels are arranged in odd stacking. The odd stacking arrangement is maintained throughout the TMUX up to the last stage. Before filtering is performed in the last stage, the stacking is made even.

The above discussion is also applicable for the FIR TMUX, except that the spectrum is divided into the upper half and the lower half at every stage. This approach leads to some disadvantages in using the FIR TMUX which will be discussed in section 6.

3. Design of the IIR Prototype Filters

The HBFs in Figure 2 are identical with the transfer function [5]:

\[ H(z) = 1 + \sum_{n=2}^{N} z^{n-1} (z^L)^{n_0} \prod_{i=1}^{N_L} \frac{-C_{ni} z^L + 1}{z^L - C_{ni}} \times \prod_{j=1}^{N_0} \frac{(P_{nj} z^L - 1)(\overline{P}_{nj} z^L - 1)}{(z^L - P_{nj})(z^L - \overline{P}_{nj})} \]

(1)

1003
where

- \( C_{nj} \) and \( P_{nj} = R_{nj} e^{j\theta_{nj}} \) are the real and complex poles of the \( s^j \)th first-order and \( j^j \)th second-order sections of the \( n^j \)th subfilter.
- \( N \) is the number of subfilters.
- \( N_{P}^n \) and \( N_{Q}^n \) are the number of first- and second-order sections of the \( n^j \)th subfilter.
- \( L \) is the decimation factor.
- \( T' \) is the input sampling period.
- \( (z^L)^m = \) reduces the inherent delay in the \( n^j \)th subfilter so that it is not in excess of the required fractional delay. A suitable initial value of \( m \) is \( N_{P}^n + 2N_{Q}^n - 1 \). Possibly a search in the neighbourhood of this value must be made to choose the best \( m \).
- \( \omega_k \) is the \( k^j \)th angular frequency in the passband or stop-band.

The transfer function \( H(z) \) consists of a unity transfer function subfilter, and fractional delay subfilters which are arranged in parallel. Each of the latter subfilters is further arranged as a cascade of first- and second-order partial fractions. The numerator and denominator of these partial fractions are mirror-image polynomials (MIP) with respect to the unit circle \( z = e^{j\omega T'} \). As a result each section has a magnitude of unity at all frequencies, i.e. the circuit is a pure phase shifter.

![Figure 1. Halfband (HBF) IIR TMUX.](image)

![Figure 2. Theoretical spectrum at different points of the halfband IIR TMUX.](image)

### 3.1 Phase Response and Group Delay

Two of the main concerns in using IIR filters are the variable phase response and the non-zero differential group delay. Compared to the conventional design techniques, the filters designed by optimization show less variation in group delay in the passband and are reasonably linear. To verify this, the analytical expressions for phase response and group delay of the TMUX are required. From eq.(1) the expression for the phase response, \( \phi(\omega) \), of the polyphase filter at \( \omega_k \) is:

\[
\phi(\omega_k) = \tan^{-1} \left( \sum_{n=2}^{N} \sin(\phi_n(\omega_k)) \right) + \tan^{-1} \left( \sum_{n=2}^{N} \cos(\phi_n(\omega_k)) \right)
\]

where \( \phi_n(\omega_k) \) is the phase response of the \( n^j \)th subfilter, and is given by:

\[
\phi_n(\omega_k) = \omega_k T'L_n + \omega_k T' (n - 1) + \sum_{i=1}^{N} \left[ \tan^{-1} \left( \frac{-C_{ni} \sin(L_w k T')}{1 - C_{ni} \cos(L_w k T')} \right) \right]
- \sum_{j=1}^{N} \left[ \frac{R_{nj} \sin(L_w k T' + \theta_{nj})}{R_{nj} \cos(L_w k T' + \theta_{nj}) - 1} \right] + \tan^{-1} \left( \frac{R_{nj} \sin(L_w k T' - \theta_{nj})}{R_{nj} \cos(L_w k T' - \theta_{nj}) - 1} \right) - \tan^{-1} \left( \frac{\sin(L_w k T') - R_{nj} \sin(\theta_{nj})}{\cos(L_w k T') - R_{nj} \cos(\theta_{nj})} \right) - \tan^{-1} \left( \frac{R_{nj} \sin(\theta_{nj}) + \sin(L_w k T')}{\cos(L_w k T') - R_{nj} \cos(\theta_{nj})} \right)
\]

By taking the negative of the first derivative of eq.(2) and eq.(3) and performing some algebraic manipula-
tions, the following expression for the group delay, \( \tau_g \), can be derived:

\[
\tau_g = -\left\{ \left[ \sum_{n=2}^{N} \cos(\phi_n) \frac{\partial \phi_n}{\partial \omega_k} \right] \left[ 1 + \sum_{n=2}^{N} \cos(\phi_n) \right] + \right. \\
\left. \left[ \sum_{n=2}^{N} \sin(\phi_n) \frac{\partial \phi_n}{\partial \omega_k} \right] \left[ \sum_{n=2}^{N} \sin(\phi_n) \right] \right\} / \\
\left\{ \sum_{n=2}^{N} \sin(\phi_n) \right\}^2 + \left[ 1 + \sum_{n=2}^{N} \cos(\phi_n) \right]^2 
\]

(4)

where

\[
\frac{\partial \phi_n}{\partial \omega_k} = T' L m_n + T'(n - 1) \\
+ \sum_{i=1}^{N_i} \frac{LT'(C_{n_i}^2 - 1)}{1 + C_{n_i}^2 - 2C_{n_i} \cos(L\omega_k T')} \\
+ \sum_{j=1}^{N_j} \frac{LT'(R_{n_j}^2 - 1)}{1 + R_{n_j}^2 - 2R_{n_j} \cos(L\omega_k T' + \theta_{n_j})} \\
+ \frac{LT'(R_{n_j}^2 - 1)}{1 + R_{n_j}^2 - 2R_{n_j} \cos(L\omega_k T' - \theta_{n_j})} 
\]

(5)

A computer program has been developed to plot eqs. (2 – 5) for the frequency range of \( 0 \leq f \leq 0.5f_s \). The peak-to-peak variation of the phase response of the designed prototype IIR filter in the passband is less than 0.0053 radian. The group delay variation in this band is 0.104T. It is near the passband edge that the filter shows a large variation in the phase response and the group delay. Therefore, with the IIR design algorithm of [5] the area near the passband edge must be avoided.

3.2 Quantization Effects

We have also investigated the effect of multiplication roundoff noise when fixed point arithmetic with b-bit representation (excluding the sign bit) is used to realise the subfilters in eq. (1). For this purpose the canonic forms have been used, in which the number of delay elements is equal to the order of the filter. Furthermore, as the numerator and denominator are mirror image polynomials, the coefficients are shared in the realisation. With this arrangement the total number of real multipliers in the realisation of the polyphase filter is:

\[
\sum_{n=2}^{N} N_i^n + 2 \sum_{n=2}^{N} N_p^n 
\]

(6)

From the definition of variance, the following expressions for the variance of uncorrelated output noise as a function of pole position can be found:

\[
\sigma_{n_i}^2 = \begin{cases} 
\frac{q^2}{6(1-C_{n_i})} & \text{first order section} \\
\frac{q^2}{3(1-R_{n_i})} & \text{second order section}
\end{cases} 
\]

(7)

where \( q \) is the quantization step. There has been good agreement between the theory and the simulations performed to measure the variance of noise at the output of the multipliers.

4. Optimization Algorithm

The unconstrained optimization technique in [6] was used to minimize the weighted error between the desired response in eq. (1) and the ideal response \( H(e^{\omega T'}) \), where

\[
H(e^{\omega T'}) = \begin{cases} 
1, & |\omega| < \frac{\pi}{T'} \\
0, & |\omega| \geq \frac{\pi}{T'}
\end{cases} 
\]

(8)

For the optimization, the first derivative of the desired response is required which can be found in [5]. The minimization step performed iteratively. At every iteration a new vector \( \bar{A} \) containing the real poles, \( C_{n_i} \), and the complex poles, \( P_{n_i} \), is found where

\[
\bar{A} = [C_{n_i}, P_{n_i}] \\
\begin{cases} 
2 \leq n \leq N \\\n1 \leq i \leq N_p^n \\\n1 \leq j \leq N_p^n
\end{cases}
\]

(9)

If the error function found in the \( q \)-th iteration is less than or equal to that of the previous iteration for all values of \( C_s \) and \( P_s \), the global minimum is found. Otherwise, the search continues.

5. Software Implementation

The implemented filter design package, which includes the design and optimization algorithms and the plotting routines, can be run on any computer with the SunOS Release 4.1.1 or higher. The package was written in the C++ programming language. C++ is a powerful object-oriented programming (OOP) language with some advanced features, such as operator overloading [7]. Since the filter design algorithm makes extensive use of complex filters, with operator overloading one can redefine the existing algebraic operators to be used with complex numbers. Furthermore, the optimization algorithm is mainly based on using vectors and matrices. C++ offers a neater approach to performing vector and matrix arithmetic.

6. Results

The following results are based on comparisons between our scheme and the FIR TMUX scheme of [4]. The IIR prototype filters have only 2 real poles (in \( z^{-1} \)). The FIR HBFs and the post processing filters have 7 and 15 taps, respectively. The FIR prototype filters were designed by the Parks – McClellan algorithm. In the quoted results the symmetrical property of the FIR filters has been taken into account. The FDM signal and filter coefficients have been quantised to 8 bits (excluding the sign bit). The input back-off of the TWTA\(^1\) is 2dB.

\(^1\)The model of Hughes 261-H-tube, used in the INTELSAT IV satellite, has been simulated.
Figure 3. Noise performance.

- To simulate the noise performance, channel 4 was selected. This channel, similar to channel 5, is subject to the highest adjacent channel interference (ACI) from both sides. As shown in Figure 3, the IIR TMUX has a better performance at all values of signal-to-noise ratios. This improved performance is due to the lower quantization error of the designed IIR prototype filter. For real poles located at $-0.579166$ and $0.121871$, the simulated normalised roundoff noise variance is $3.68 \times 10^{-6}$, which is close to the theoretical normalised value of $3.54$ (see eq. (7)). The simulated normalised roundoff noise variance in the FIR prototype filter is $12.59$, which is about 3.4 times higher than that of the IIR TMUX.
- The IIR TMUX has 76 real multipliers, while the FIR TMUX has 356 real multipliers. This is a saving of 78.7% in the number of multipliers.
- The IIR TMUX has 108 real adders, while the FIR TMUX has 500 real adders. This is also a saving of 78.4% in the number of adders.
- The transmultiplexing delay in the IIR TMUX is $9T$, while the delay in the FIR TMUX is $16T$. This is equivalent to 43.5% faster transmultiplexing.
- To demultiplex 8 channels, the channels must be sampled at $32R$ in the FIR TMUX, while the sampling frequency in the IIR TMUX is exactly half this value, i.e. $16R$. The direct impact of this on the transmitter is that the anti-aliasing anti-imaging filter which is used in the transmitter is half the length of that used in the scheme of [4]. The other impact is that since no channels can be multiplexed in the frequency range of $16R < f < 32R$, half the spectrum remains unusable.
- The filtering modules in the FIR TMUX are highpass filters, lowpass filters and post processing filters, while in the IIR TMUX only lowpass filtering modules are required. This makes our scheme more attractive for ASIC implementations.

7. Discussion and Conclusions

In this paper we have introduced an efficient IIR TMUX for use in OBP satellites. Comparisons with the existing FIR TMUX show that our scheme performs demultiplexing with a lower number of arithmetic operations. Therefore, there will be less complexity in the ASIC design. Under the same channel conditions, an improvement in the noise performance has been achieved. With the FIR TMUX, half the spectrum remains unusable, while with the IIR TMUX the spectrum is fully utilised. The design specifications of the prototype IIR filters must be such that the channels are not too close to the passband edge. Only 5 bits are adequate for quantizing the FDM signal and filter coefficients. And finally, our transmultiplexer operates faster than the FIR TMUX.

References


---

2Oversampling by a factor of 2.
3Critical sampling.
Real Time implementation of the GMDF Alpha algorithm on a multi DSP TMS320C40 board

Grégoire LE TOURNEUR, Jean-Philippe THOMAS, André GILLOIRE

France Telecom CNET Lannion A Route de Tregastel - BP 40
22301 LANNION CEDEX FRANCE
Tel / fax +33 96 05 33 69 / +33 96 05 13 16 - E-Mail: gilloire@lannion.cnet.fr

Abstract. This paper addresses the problem of implementing in real time the recently introduced GMDF Alpha, an efficient frequency domain block adaptive filtering scheme well suited to acoustic echo cancellation for teleconference applications. A parallel implementation of this algorithm on two TMS320C40 floating point digital signal processors supported by a PC board is described, which provides real time operation (at 16 kHz sampling frequency) of an acoustic echo canceller with a filter impulse response of 64 ms. Experimental results illustrate the real time performance of the implemented algorithm.

1. Introduction
The problem of real time implementation of acoustic echo cancellation algorithms is still a challenging one, especially in wide band (7 kHz) teleconference applications, because in this context the acoustic impulse response is generally very long and the sampling frequency is high (16 kHz). Although a large number of algorithms have been proposed which can be used for wide band acoustic echo cancellation [1], most of the implementations performed until now have been based on the popular NLMS algorithm, because it requires moderate computational power and it can be straightforwardly implemented on standard digital signal processors (DSP). However, it is well known that the performance of this algorithm on speech signals suffers limitations; furthermore the computational complexity of the NLMS is still too high for the identification of very long impulse responses found in poorly acoustically damped or large teleconference rooms. Another critical issue is the control of the adaptation, which must be performed in real time with speech signals [2].

New efficient algorithms have appeared during the last five years, which have both very good performance in terms of convergence with speech signals, and reduced complexity (in comparison with the standard time domain NLMS) thanks to block processing of speech samples. One of these new algorithms: the so-called GMDF Alpha (standing for Generalized Multi-Delay adaptive Filter with oversampling factor $\alpha$) [3], was found particularly useful for implementation studies because, in addition with good convergence and tracking performances and low computational complexity, it can be easily split into specific computational tasks which can be carried out simultaneously in a parallel architecture.

We describe in this paper a parallel implementation of the GMDF Alpha algorithm on two Texas Instruments TMS320C40 floating point DSPs supported by a commercial PC board. This "hand made" implementation allowed us to perform real time experiments in a wide band teleconference context. It is hand-made in the sense that we did not use CAD tools which can perform automatically parts of the parallel implementation process [4]. A brief description of the GMDF Alpha algorithm is recalled in section 2. Section 3 presents the main characteristics of the TMS320C40 DSP and of the PC board used. The parallel implementation is described in section 4; some experimental results are presented in section 5.

2. Overview of the GMDF Alpha algorithm
The GMDF Alpha algorithm can be viewed as a generalization of the Frequency-domain LMS (FLMS) algorithm, which is the transposition in the frequency domain of the Block LMS algorithm. The FLMS algorithm, which provides reduced complexity in comparison with the standard LMS, suffers from two main drawbacks. The first one comes from the size of the blocks which introduces an unacceptable delay in the system, since it is equal to the size $L$ of the filter modelling the acoustic impulse response. The second one is a limitation of convergence speed and tracking performance in case of fast variations of the acoustic path, because the adaptation of the filter is done only once every L samples. The GMDF Alpha algorithm offers two significant improvements over the FLMS: first, the segmentation of the impulse response of size L in K blocks of smaller size N leads to short delay; moreover, increased updating rate of the coefficients in the ratio $\alpha$ improves both the convergence speed and the tracking capability of the algorithm.

The basic equations of the GMDF Alpha algorithm as well as a performance analysis can be found in ref. [3]. A block diagram of the algorithm is shown in figure 1. The shaded areas correspond to the different "tasks" implementing the computational parts of the algorithm.

Exact equivalence with time domain adaptation requires the application of a constraint on the adapted impulse response, which costs K direct and K inverse FFTs per iteration of R=\[N/\alpha\] samples. To avoid this large extra
3. A floating point parallel processing oriented DSP: the TMS320C40

The TMS320C4x floating point family has been designed for intensive mathematical computations. Devices in this generation incorporate on-chip hardware to facilitate high-speed interprocessor communication without degrading CPU performance. This product is well suited for quick real time implementation and in-situ evaluation of new complex algorithms without constraints of optimization in a first step.

The architecture of the TMS320C40 includes a high performance CPU capable of 40 MFLOPS, on-chip program cache and dual access single-cycle RAM, two identical external data and address busses to support shared memory systems and high data rates, 6 high speed ports for glueless interprocessor communications, and a 6-channel DMA coprocessor without CPU intervention [5].

This device is fully supported by a set of software development tools to ease and speed the design of systems, such as an advanced C compiler.

We used a DPC/C40B PC board (manufactured by Loughborough Sound Images Ltd.) as the hardware support of our experimental teleconference system. This board supports two TIM modules and an analog I/O interface daughter module. Each TIM module is equipped with 1 DSP TMS320C40, 96Kx32 words of data or program memory, and 16Kx32 words are available for global memory. The layout of this PC board is shown in figure 2.

4. Implementation of the GMDF Alpha algorithm on two TMS320C40 DSPs

4.1 Constraints specific to the application and settings of the algorithm

High quality teleconference systems use "wideband speech" including frequencies up to 7 kHz, which is handled by wideband speech codecs complying to the CCITT G722 Recommendation. The corresponding sampling frequency is 16 kHz; therefore, 1250 cycles are available within each sampling period when using a TMS320C40 DSP running at 50ns cycle time. Since our goal is to estimate impulse responses of large duration (with a minimum of 64 ms), we must optimize the distribution of computations in the algorithm over the time. The processing in the GMDF Alpha algorithm is performed on a block-by-block basis. From lots of simulations on real speech signals, we found an
optimum block length of R=64 samples, which corresponds to a total of 80000 available cycles to implement the GMDF alpha algorithm in real time.

For a filter impulse response of 64 ms, i.e. L=1024 samples, the other settings of the algorithm (also resulting from simulations) were the following:
- impulse response segments size N=256 (i.e. K=4)
- oversampling factor \( \alpha = 4 \)

### 4.2 Steps to the real time implementation

In a first step, a model of the GMDF Alpha algorithm written in C language has been used to obtain a first executable code for the TMS320C40 DSP, thanks to the Texas Instruments' advanced C compiler. Although this first code did not run in real time, it provided us with a verification tool, since (with little modification of the C software: essentially I/O procedures) we got the same results as those obtained from the software version running on UNIX workstations (within slight differences due to non identical floating point formats).

The second step to the real time implementation was the evaluation of the execution time for each procedure ("task") directly obtained from the C code (with Texas Instruments' optimised real FFT). This evaluation led to about 160000 cycles to process each input buffer of R samples. Note that this result holds for the core of the GMDF Alpha algorithm, i.e. without input/output mechanism and control algorithm.

The third step consisted in extracting and optimizing critical loops in terms of computational load; these loops were optimised by writing them directly in assembly language. As shown in Table 1, this process led to reduce the total number of cycles from about 160000 to 88500.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>direct from &quot;C&quot; to assembler</th>
<th>assembler optimised</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT (2) (real, 512 points)</td>
<td>10718 x 2</td>
<td>10550 x 2</td>
</tr>
<tr>
<td>Convolution</td>
<td>60882</td>
<td>12820</td>
</tr>
<tr>
<td>Inverse FFT</td>
<td>14206</td>
<td>13980</td>
</tr>
<tr>
<td>Error computation</td>
<td>1598</td>
<td>850</td>
</tr>
<tr>
<td>Overlap-add</td>
<td>1444</td>
<td>668</td>
</tr>
<tr>
<td>Adaption gains</td>
<td>22792</td>
<td>18924</td>
</tr>
<tr>
<td>Coefficients adaptation</td>
<td>33340</td>
<td>18996</td>
</tr>
<tr>
<td>others</td>
<td>3836</td>
<td>1222</td>
</tr>
<tr>
<td>TOTAL</td>
<td>159534</td>
<td>88560</td>
</tr>
</tbody>
</table>

Table 1. Number of cycles obtained for each main procedure of the algorithm

After this step, the effort required to reduce the total number of cycles under the limit of 80000 cycles was found so high that we stopped the optimisation. It appeared clearly that a parallel implementation on two DSPs would provide sufficient overall computational power to achieve the real time execution of the algorithm. This choice allowed us to focus on the development of control techniques for the adaptation and to develop the I/O interfaces necessary for the real time implementation of the algorithm on two TMS320C40 DSPs. Then we could perform real time measurements and tests in teleconference-like situations.

The last, challenging step was to split the algorithm on the two TMS320C40 DSPs, taking into account the hardware organization of the PC board for data transfers between the DSPs. Many solutions were tested, with large differences in efficiency of data transfers. The chosen solution optimizes the amount of data to be transferred and also the transfer waiting time for each TMS320C40. Only one DMA link between the two DSPs is used. The first DSP is idle approximately 10% of the available time, whereas the second one is idle 75% of that time. Note that, although DMA transfers are effectively executed in parallel with the processing in the CPU, it is difficult to split the algorithm in an optimal way because only a few number of tasks of the algorithm can be executed in parallel. An alternative approach would consist in splitting the code of each task to implement it on the two DSP at the same time. This way needs a lot of work to produce appropriate assembly code, so we rejected it.

In the final version of our implementation, the only "parallelized" tasks are the computation of the adaptation gain parameters (executed at the same time as the convolution and the inverse FFT on the other DSP) and the coefficients adaptation (which is split between the two DSPs). Note that this parallel execution needs important data transfers between the two DSPs, but the single external port limits the transport possibilities. This final implementation is shown in figure 3.

It appears clearly that this implementation leaves lots of available computational time in the second DSP to include other tasks for control, etc., provided that these tasks do not require too much data transfers.

The memory resources used by each of the two DSPs are given in Table 2 below.

### Table 2. Memory resources used by each of the two DSPs

<table>
<thead>
<tr>
<th>DSP</th>
<th>Internal DATA RAM</th>
<th>external DATA RAM</th>
<th>external PROG. RAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRI</td>
<td>2x32</td>
<td>7.7k x 32</td>
<td>2.1k x 32</td>
</tr>
<tr>
<td>SEC</td>
<td>2x32</td>
<td>7k x 32</td>
<td>1.8k x 32</td>
</tr>
</tbody>
</table>

5. Real time measurement of performance

The real time implementation described in section 4 was used to measure the performances of the GMDF Alpha algorithm and to perform tests in real situation. As acoustic front-end, a loudspeaker and a microphone were installed on a table in a small room with low reverberation time (about 150 ms). The distance between the loudspeaker and the microphone was about 50 cm, the back of the microphone facing the loudspeaker. The acoustic coupling path between the loudspeaker and the microphone could be modified by using a cardioid microphone cell or an omnidirectional one. Measurements were performed with the CCITT P.50 artificial voice as the excitation source (loudspeaker signal).
Figure 3. Implementation of the tasks on the two DSPs. Task sizes correspond approximately to their durations.

The results given in figure 4 have been obtained with an omnidirectional microphone, which corresponds to the worst case for acoustic coupling. The figure shows the spectrum of the residual echo after convergence, along with the spectra of the microphone signal and of the microphone background noise. The obtained ERLE is larger than 30 dB at the low frequencies, and the residual echo is generally not far from the background noise. This result (in good agreement with simulation ones) demonstrates the good performance of the real time implementation.

6. Conclusion
The "hand-made" parallel implementation of the GMDF Alpha algorithm on two TMS320C40 DSPs was successful since real time operation of the algorithm was achieved. Nevertheless, it revealed that the data transfers between the two DSPs critically limit the possibility to parallelize the computational tasks. This implementation allowed us to perform real time tests in various situations. Further studies are planned to increase the length of the filter impulse response, and to optimize the present implementation to run the algorithm on a single DSP of the TMS320C3x family, which would provide a cheaper implementation for development of teleconference products.

Figure 4. Spectrum of the residual echo signal after convergence (3). The spectra of the microphone signal (1) and of the background noise (2) are also shown.

Acknowledgements
The authors wish to thank Eric Moulines and Jacques Prado, from Telecom Paris Dept. Signal, for their useful advices and for preliminary versions of the GMDF Alpha software. They also thank Hervé Treguer who performed lots of simulations to optimize the settings of the algorithm.

References

1010
A maximum likelihood solution to blind identification of multichannel FIR filters

K. Abed Meraim and E. Moulines

Telecom Paris / SIG 46, rue Barrault 75634 Paris CEDEX 13 FRANCE
E-mail: abedmerai@sig.enst.fr, Tel.: +33 1 40817515, Fax: +33 1 40817935.

Abstract. In this contribution, a maximum likelihood solution to blind identification of single-input multiple-outputs FIR filter for discrete input is presented. The likelihood of the observations being very modulated, direct solutions or approximations of the likelihood equations are likely to be rather intricate. To alleviate this problem, we resort to the standard complete/incomplete data model, where the observations play the role of the incomplete data while the input data (the unobserved symbol sequence) are the missing data. We then maximize the incomplete likelihood (i.e., the likelihood of the observations) by iteratively maximizing the complete likelihood function using (i) the standard EM algorithm and (ii) simple approximations of it. Simple and robust initialization procedures are also discussed. Some applications to blind deconvolution/ equalization in the digital communication framework are presented.

1. Introduction

Narrowband TDMA digital cellular radio systems require identification and equalization algorithms that are robust to (i) low to very low SNR (say between 5dB to -5dB) (ii) severe multipath fadings. A way to cope with these difficult problems is first to use multiple sensors (say, 2 to 4 sensors) in combination with appropriate non-linear spatial filtering devices.

The interest for spatial diversity equalizer traces back to the early work by Baraban and Saltz [1]. In this contribution, a linear spatial filter is adapted so as to minimize the mean-square error (MSE) at the array output, subject to appropriate constraints [1, 2, 3] (this scheme shares many similarities with the linearly-constrained minimum-variance beamformer). The linear spatial equalizer is adapted on a block-by-block basis (more specifically, for each time slot), making use of a (short) training sequence (say, 16 to 20 symbols). More recently, [4, 5] have proposed schemes for fast blind identification diversity channel parameters resorting only on the second-order statistics of the array output (see also [6]). The channel is then equalized either by using a linear minimum mean-square error equalizer or the standard vectorial Viterbi decoder. Some improvements over these methods have been presented in [7].

In this contribution, we develop a maximum likelihood solution for joint diversity channel parameters (blind) estimation and symbol detection. We resort to the standard complete/incomplete data framework, where the observed signals (array output) play the role of the incomplete data whereas the (unobserved) input symbol sequence are the missing data. The incomplete likelihood is then maximized by iteratively maximizing the complete likelihood. Two algorithms are presented:

- a standard EM algorithm, where the required a posteriori probability of the symbols given the observations and the current values of the parameters are evaluated by resorting to the computationally intensive forward-backward procedure,

- an approximate EM implementation, where the a posteriori probabilities of the symbols are approximated instead of being exactly evaluated, leading to much simpler algorithms, especially when the number of filter coefficients is large.

In both cases, the calculated likelihood functions are then used to provide optimum decisions on information symbols which minimize the symbol error probability. The proposed procedure does not necessitate the transmission of a training sequence, and is thus essentially blind (a training sequence can be used to improve the robustness, but is not needed for the algorithm to work). Simulations in realistic mobile radio communication context (GSM channel, low SNR) demonstrate the usefulness of this approach [5].

2. Problem Formulation

2.1. Diversity Signal and Channel Model

In this contribution, blind equalization of a digital source impinging on an array of $D$ sensors is considered. For convenience, we assume perfect timing synchronization (a realistic assumption in current TDMA systems). Under the standard discrete-time model for linear digital modulation over a linear ISI channel [8], the measurements at each receiver can be represented as

$$y_k^{(i)} = \sum_{l=0}^{M_i} h_l^{(i)} u_{k-l} + n_k^{(i)} \quad 1 \leq i \leq D$$

(1)

where $u_k$ is the symbol sequence and $h_i = [h_0^{(i)}, \ldots, h_{M_i}^{(i)}]^T$. $M = \max_{1 \leq i \leq D} (M_i)$ are the discrete-time composite channel coefficients (that includes the transmit filter, the channel and the receive filter). The following assumptions are made (see [9] for comments): (i) The symbols $u_k$ are a sequence of independent complex random variables, taking their values in a finite alphabet $A = \{a_1, \ldots, a_{2^B}\}$ of size $2^B$ ($B$ is the number of bits per symbol) with equal probabilities. (ii) The noise processes
in the diversity channels are independent from the symbol sequence and are mutually uncorrelated. (iii) The complex noise samples \(n_t^i\) are white and circular Gaussian with variance \(\sigma_r^2, 1 \leq i \leq D\). These hypotheses make however sense in TDMA system, since the typical correlation length of the noise field is very short (note that hypothesis (ii) and (iii) can be weakened; in fact, arbitrary noise covariance matrix can be identified).

The parameter to be identified are the channel coefficients and the noise variances collected in the matrices \(H = [h_1, \ldots, h_D]^T\) and \(\Sigma = \text{diag}\{\sigma_1^2, \ldots, \sigma_D^2\}\). These parameters are collectively referred to as \(\theta = (H, \Sigma)\).

Blind diversity channel parameters estimation

Based on T successive observations of the array output \(Y_t = [y_t^1, \ldots, y_t^D]^T\), the objectives are (i) to estimate the parameters underlying the channels and the noise model and (ii) to determine the posterior probability of the emitted symbol sequence.

The estimation method we consider is an extension to the multichannel case of the solution proposed in [10, 11, 12] in the scalar case. It relies upon the EM algorithm, an iterative method for finding maximum likelihood estimates in incomplete data problems [10, 11, 12, 13, 14, 15].

Define \(S_n\), the vector state process of dimension \(M\) by: \(S_n = (s_{n-1}, \ldots, s_{n-M})\). \(S_n\) is a discrete-time finite-state Markov process, with \(N = 2^M\) states. The states are denoted \(q_n\) and are defined as particular combinations of the alphabet of \(M\) symbols taken in the alphabet \(\{a_0, a_1, \ldots, a_M\}\). The set of states is denoted \(Q = \{q_0, \ldots, q_N\}\). The vector \(x_t = (x_t^1, \ldots, x_t^M)\) of \((M+1)\) symbols associated to the transition between two connected states \(q_n\) and \(q_m\) is denoted \(x_{nm}\). The set of the \(2^M(M+1)\) possible transitions are denoted \(Y\). For convenience, sequences of observations \(Y_t, \ldots, Y_{t-M}\) and states \(S_{t-M}, \ldots, S_t\) are denoted \(Y_{t-Mt}\) and \(S_{t-Mt}\) respectively. In the EM framework, the sequence of states \(S_{t-Mt}\) (not observed) are the missing data. \(Y_{t-Mt}t\) are the (incomplete) observations. \(Y_{t-Mt}, S_{t-Mt}\) is the complete-data vector random variables. Each iteration of the EM algorithm has two steps: an E-step and a M-step. The \((m+1)\)st E-step finds the conditional expectation of the complete-data loglikelihood with respect to the conditional distribution of the 'missing data' \(S_{t-Mt}\) given the 'observations' \(Y_{t-Mt}\) and the current estimated parameter \(\theta^{(m)}\).

\[
\theta = (H, \Sigma) - Q(\theta; \theta^{(m)})
\]

where

\[
Q(\theta; \theta^{(m)}) \triangleq E(\log f_{\theta}(Y_{t-Mt}, S_{t-Mt})|Y_{t-Mt}, H^{(m)}, \Sigma^{(m)})
\]

is referred to as the reestimation function. The \((m+1)\)st M-step then finds \(\theta^{(m+1)}\) to maximize the reestimation function \(\theta \rightarrow Q(\theta; \theta^{(m)})\) with respect to \(\theta\):

\[
Q(\theta^{(m+1)}; \theta^{(m)}) \geq Q(\theta; \theta^{(m)}) \quad \text{for all} \quad \theta \in \Theta
\]

Neglecting terms independent of \(\theta\), the reestimation function \(Q(\theta; \theta^{(m)})\) may be shown to be proportional to:

\[
Q(\theta; \theta^{(m)}) \propto \sum_{t=M}^{M+T} \sum_{x_t \in \{0, 1\}^M} (-\sum_{i=1}^{D} \log(\sigma_i^2) - (Y_t - H x_t)^H \Sigma^{-1}(Y_t - H x_t)) q_{t+1}^{(m)}(t; i, j)
\]

where \(\gamma_{t+1}(t; i, j) = f_{t+1}(S_t = q_t, S_{t+1} = q_j|Y_t^T)\) is the conditional probability of the trellis branch \((S_t = q_t, S_{t+1} = q_j)\) given the observations \(Y_t^T\) and the current estimated parameter \(\theta^{(m)}\).

Standard EM algorithm

The posterior probability of the trellis branch \(\gamma_{t+1}(t; i, j)\) is efficiently computed using the forward and the backward variables [14, 10, 12]

\[
\alpha_{t+1}(t; i) \triangleq f_{t+1}(Y_t^T, t|S_t = q_t) \quad \beta_{t+1}(t; j) \triangleq f_{t+1}(Y_t^T|S_t = q_t)
\]

according to the following equation (where multiplicative scaling factors independent of \(t, i, j\) are omitted)

\[
\gamma_{t+1}(t; i, j) \propto \beta_{t+1}(t + 1; j) \alpha_{t+1}(t; i) b_{t+1}(Y_t^T; i, j)
\]

\[
b_{t+1}(Y_t^T; i, j) \propto \prod_{i=1}^{D} (c_t^i)^{-2} \exp(-(Y_t - H^{(m)} x_{ij})^H (\Sigma^{(m)})^{-1}(Y_t - H^{(m)} x_{ij}))
\]

The forward and backward variables are evaluated recursively (as in the classical Baum-Welch algorithm) according to

\[
\alpha_{t+1}(t + 1; i) = \frac{1}{N} \sum_{j \in \mathcal{F}(t)} \alpha_{t+1}(t; j) b_{t+1}(Y_t^T; i, j)
\]

\[
\beta_{t+1}(t; i) = \frac{1}{N} \sum_{j \in \mathcal{B}(t)} \beta_{t+1}(t + 1; j) b_{t+1}(Y_t^T; i, j)
\]

where \(\mathcal{F}(t) and \mathcal{B}(t)\) denotes the set of states connected to \(q_t\) in forward and backward directions, respectively. The reestimation function is quadratic in its argument and that the maximization step reduces to:

\[
\frac{H^{(m+1)}}{\Sigma^{(m+1)}} = R_{xy} R_x^{-1}
\]

where the matrices \(R_{xy}, R_{xy}\) and \(R_{xx}\) are respectively the array output autocorrelation matrix, the "weighted" cross-correlation matrix between the unobserved symbol transitions and the array output and the "weighted" autocorrelation matrix of the unobserved symbol transitions. They are given by:

\[
R_{xy} = \sum_{t=1}^{T} x_t^H Y_t^T
\]

\[
R_{xx} = \sum_{t=1}^{T} x_t x_t^H
\]
Detection of information symbols
Denote \( \theta^{(*)} \) the values of the parameters obtained at the end of the iterations. Given this estimate, the observation sequence \( Y_{1:T} \) and the trellis diagram of the channel, several detection methods can be designed. The optimality criterion retained here is the minimum symbol-error probability [8], which is easily implemented in this context. Minimizing the symbol-error probability amounts to choose, at time \( t \), the information symbol which maximizes the posterior probability of the symbol \( u_t \) given the observations \( Y_{1:t} \) and the channel parameters \( \theta^{(*)} \):

\[
\hat{u}_t = \arg\max_{a_t \in A} f_{\theta^{(*)}}(u_t = a_t | Y_{1:t}; \theta^{(*)})
\]

This quantity may be simply expressed in functions of \( \gamma_{\theta^{(*)}}(t); k, l \) the posterior probability of the trellis branch \( S(k) = x_{kl} \) given the observations \( Y_{1:t} \) and the channel parameters \( \theta^{(*)} \):

\[
f_{\theta^{(*)}}(u_t = a_t | Y_{1:t}; \theta^{(*)}) = \sum_{k \in S(t)} \gamma_{\theta^{(*)}}(t; k, l)
\]

where \( S(t) \) is the set of all trellis branch values \( x_{kl} \) corresponding to the input symbol \( u_t = a_t \).

Approximate EM algorithm
The parameters reestimation in (5), (6) requires at each iteration the calculation of the conditional expectations \( E_{\theta^{(n)}}(u_t | Y_{1:t}) \) and \( E_{\theta^{(n)}}(u_t, i_t | Y_{1:t}) \) for \( 1 \leq t \leq i \leq j \leq T \). The total amount of calculation, using the forward-backward algorithm is \( O(2(B + M^2)) \); it is exponential in the trellis size, prohibiting the use of such technique for large filters length \( M \). A way of reducing the total amount of calculation is to use approximate expression of the a posteriori probabilities of the symbols [16].

Let us rewrite the array output as:

\[
Y = [y_1^T, \ldots, y_T^T]^T = \sum_{k=1-M}^{T} h_k u_k + [n_1^T, \ldots, n_T^T]^T
\]

where \( h_k \) is the \((k + M)^{th}\) column of the \( DT \times (M + T) \) matrix

\[
\begin{pmatrix}
  h_M & h_{M-1} & \cdots & h_0 & 0 & 0 \\
  0 & h_M & h_{M-1} & \cdots & h_0 & 0 \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
  0 & 0 & \cdots & h_M & h_{M-1} & h_0 \\
\end{pmatrix}
\]

and \( x_{ij} = \sum_{k=1}^{M} \hat{h}_k u_k + [n_1^T, \ldots, n_T^T]^T \).

By considering in (9) \( x_{ij} \) as the total observation noise, and by assuming that it is Gaussian with zero mean and covariance matrix denoted \( K_{ij} \), the joint a posteriori distribution of \( u_i \) and \( u_j \) is expressed as:

\[
f_{\theta^{(*)}}(u_i = a_i, u_j = a_j | Y_{1:t}; \theta^{(*)}) = \frac{\exp(-z_i^2 K_{ij}^{-1} z_j)}{\exp(-z_i^2 K_{ij}^{-1} z_j) + \exp(-z_j^2 K_{ij}^{-1} z_j)}
\]}

where

\[
\rho_{ij} \stackrel{def}{=} \hat{h}_i^T K_{ij}^{-1} Y \quad \text{and} \quad \lambda_{ij} \stackrel{def}{=} \hat{h}_j^T K_{ij}^{-1} Y
\]

As shown in [16], above quantities are efficiently calculated using the matrix inversion lemma as follows:

Let us define the \( DT \times DT \) matrices:

\[
K = \sum_{k=1-M}^{T} \hat{h}_k \hat{h}_k^T \quad \text{and} \quad \Lambda = \text{diag}(\Sigma^1, \ldots, \Sigma^T)
\]

\[
K_{ij} = K - \hat{h}_i \hat{h}_j^T
\]

where \( \Sigma^T \) is the signal variance. Then applying the inverse matrix lemma we obtain:

\[
K_{ij}^{-1} = (K_{ij} - \hat{h}_i \hat{h}_j^T)^{-1}
\]

\[
= K_{ij}^{-1} + (\hat{h}_i^T K_{ij}^{-1} \hat{h}_j) / (\sigma^2 - \hat{h}_i^T K_{ij}^{-1} \hat{h}_j)
\]

\[
K_{ij}^{-1} = K_{ij}^{-1} + (\hat{h}_i^T K_{ij}^{-1} \hat{h}_j) / (\sigma^2 - \hat{h}_i^T K_{ij}^{-1} \hat{h}_j)
\]

it follows:

\[
\alpha_{ij} = \hat{h}_i^T K_{ij}^{-1} \hat{h}_j + \sigma^2 \hat{h}_i^T K_{ij}^{-1} \hat{h}_j / (\sigma^2 - \hat{h}_i^T K_{ij}^{-1} \hat{h}_j)
\]

\[
\beta_{ij} = \hat{h}_i^T K_{ij}^{-1} \hat{h}_j / (1 - \sigma^2 \hat{h}_i^T K_{ij}^{-1} \hat{h}_j)
\]

\[
\gamma_{ij} = \hat{h}_i^T K_{ij}^{-1} \hat{h}_j / (1 - \sigma^2 \hat{h}_i^T K_{ij}^{-1} \hat{h}_j)
\]

\[
\rho_{ij} = \hat{h}_i^T K_{ij}^{-1} Y + (\hat{h}_i^T K_{ij}^{-1} \hat{h}_j) / (\sigma^2 - \hat{h}_i^T K_{ij}^{-1} \hat{h}_j)
\]

\[
\lambda_{ij} = \hat{h}_i^T K_{ij}^{-1} Y / (1 - \sigma^2 \hat{h}_i^T K_{ij}^{-1} \hat{h}_j)
\]

By recursion it is calculated:

\[
K_{ij}^{-1}(-M) = \text{diag}(\Sigma^1, \ldots, \Sigma^T)
\]

\[
K_{ij}^{-1}(k) = K_{ij}^{-1}(k-1) - \frac{(K_{ij}^{-1}(k-1) \hat{h}_k \hat{h}_k^T K_{ij}^{-1}(k-1))}{(\sigma^2 - \hat{h}_i^T K_{ij}^{-1} \hat{h}_j)}
\]

\[
K_{ij}^{-1} = K_{ij}^{-1}(T)
\]

The total amount of calculation of the quantities (10) and (11), is about \( O((DT)^3) \) for the inversion of the matrix \( K \) and \( O(D(M+1)^2) \) for the calculation of \( \hat{h}_i^T K_{ij}^{-1} \hat{h}_j \) and \( \hat{h}_i^T K_{ij}^{-1} Y \) (since \( \hat{h}_i \) has only \( D(M+1) \) non-zero elements).

3. Simulations
The standard EM algorithm has been simulated for the case of independent QAM4-modulated input symbols [17]. The number of channels is \( D = 2 \) and the degree of the ISI is \( M = 2 \). The transfer function of the first channel was \( F_1(z) = 1.00 + (0.500 + 0.333)z^{-1} + (0.500 + 0.333)z^{-2} \). The transfer function of the second channel was \( F_2(z) = (0.500 + 0.800)z^{-1} - (0.333 + 0.900)z^{-2} + (0.900 - 0.333)z^{-3} \). The additive noise is a temporally and spatially white Gaussian process. For each situation, the blind subspace identification algorithm (BMU) (presented in [7]) was used for the initialization step and 100 Monte-Carlo runs were conducted. Figure 1 and 2 show the fast convergence of the EM algorithm, practically in 1 iteration for this range of parameters. Furthermore figure 2 shows that the probability of error after 1 or 2 EM-iterations is the same with this one obtained by a Viterbi-algorithm using the exact
channel parameters. In figure 3 the mean square error of the channel estimates is plotted against the SNR. It compares performances obtained after the initialization step with those obtained after 1 iteration of the EM algorithm. It shows the superiority of the derived algorithm.

4. Conclusion

We have presented the Baum-Welch version of the Expectation-Maximization algorithm to the problem of multichannel data reception in the presence of additive white noise and intersymbol interference. In contrast with the monochannel blind identification, efficient initialization procedures are available, leading to fast convergence schemes (practically in 1 or 2 iterations). Furthermore for large ISI situations a multichannel version of an approximate EM algorithm has been presented.

---

References

Analysis of a Method for Classification of Analogue Modulated Radio Signals

Peter A.J. Nagy

National Defence Research Establishment, Sweden
Box 1165, S-581 11 Linköping, Sweden
E-mail: peina@lin.foa.se
Fax: +46-13-118009

Abstract. Since the early days of radio communication the monitoring of the electromagnetic spectra has been an important task. In threat recognition signal characteristics, estimated from the intercepted radio signal, are compared to a catalogue of typical characteristics or sorting parameters. One of the most important sorting parameters is the modulation type. By analyzing the variability of amplitude and frequency of a radio signal, the modulation type can be established. In this contribution we derive expressions for the "fourth" moment of the amplitude and the variance of the instantaneous frequency. Based on these expressions and estimates of the statistical properties of message and noise signals, we establish decision regions for the modulation types. With these decision regions it is simple to build a modulation classifier. Simulations and preliminary experiments on real data show the usefulness of this method. This study is concentrated to the analogue modulation forms AM, DSB, SSB, and FM. For completeness we also consider pure carrier signals and noise.

1. Introduction

Soon after the possibility of information transmission by radio signals was discovered, the need for monitoring the electromagnetic spectrum became obvious. Typical applications for such monitoring are threat recognition, supervision of admitted radio traffic, and general supervision and control.

In, for instance, threat recognition the intercepted radio signal is analyzed, and the signal characteristics are compared against a catalogue of typical characteristics or sorting parameters. One of the most important sorting parameters is the modulation type.

In the past, the analysis of intercepted radio signals has relied on operator interpretations of measured parameters (or just by listening to the radio signal). This operator oriented approach has some pitfalls like subjective judgments. Furthermore, the increasing activity in the frequency spectrum, has made manual identification difficult and non practical. This motivates an interest in automated methods for modulation classification.

A method for automatic modulation classification must be able to cope with an complicated signal environment. Here are potential problems caused by noise, fading, man-made broadband disturbances and interfering from nearby frequency channels.

2. Previous Work

In 1985 Gadbois [1,2] introduced some "new" parameters for modulation classification based on envelope and instantaneous frequency characteristics. The envelope based parameter is denoted \( R \) and it is defined as

\[
R = \frac{\mathbb{E}[a^2(t)] - \mathbb{E}[a^2(t)]^2}{(\mathbb{E}[a^2(t)])^2},
\]

where \( a(t) \) is the envelope of the radio signal. The parameter \( R \), which measures the amplitude variability, was shown - by Gadbois and others [3] - to have distinct values for different modulated radio signals. The published analysis was carried out for message and noise signals with Gaussian distribution. The reason for using the squared amplitude instead of just the amplitude in equation (1) is that the parameter obtained in the first case is more easy to analyze.

To distinguish between constant envelope signals, three more parameters were estimated. These parameters measured the variability of the instantaneous frequency, which was estimated with a DPLL based device. The parameters used are

\[
R_4 = \frac{\mathbb{E}[\omega_2(t) - \mathbb{E}[\omega_2]]^2}{(\mathbb{E}[\omega_2])^2}, \quad R_5 = \frac{\mathbb{E}[\omega_3(t) - \mathbb{E}[\omega_3]]^2}{(\mathbb{E}[\omega_3])^2}, \quad R_6 = \frac{\mathbb{E}[\omega_4(t) - \mathbb{E}[\omega_4]]^2}{(\mathbb{E}[\omega_4])^2},
\]

where \( \omega_i \) is a smoothed version of the instantaneous frequency and \( \tilde{\omega}_i \) is a lowpass filtered version, see [1,7] for further details. Others have also contributed to the study of these and similar parameters, like Galfant [4], Ribble [5], and UTL [6].

3. The Work Presented in This Paper

In the current contribution we use Gadbois' parameter \( R \) (for the envelope), see equation (1), and the variance of the instantaneous frequency normalized with the squared sample time

\[
R_4 = \frac{\mathbb{E}[\omega_i(t) - \mathbb{E}[\omega_i]]^2}{(\mathbb{E}[\omega_i])^2},
\]

where \( \omega_i(t) \) is the instantaneous frequency. Here we normalize with the squared sample time (7) to get a bandwidth independent parameter.

We concentrate on the analogue modulation forms AM, DSB, SSB and FM (while Gadbois and others also tried to classify FSK and FSK signals). For completeness we also study pure carrier signals (CW) and noise (NOISE).

The analysis carried out is extensive compared to the analysis by Gadbois and others [3], which only considered Gaussian signals. Both noise and message signals are assumed to be non Gaussian. Typical values of the statistical properties, which describe these signals, are estimated from real signals. For radio signal acquisition we use a Steinbrecher DT-14088 C-1 based receiving system.

The modulation classifier presented here will be used together with a classifier for digitally modulated signals [8] in a system for detection and classification of special transmission types. We currently work on burst transmissions in the HF-band (1.5 MHz - 30 MHz).
4. Some Definitions

A radio signal can be written as

\[ r(t) = \Re \left[ g(t) e^{j(\omega_c t + \phi)} \right] + n(t), \quad (4) \]

where \( g(t) = \int m(t) dt \) is the complex envelope, \( \omega_c \) is the nominal center frequency, \( \Delta \omega \) is a carrier offset, \( \theta \) is a random angle uniformly distributed in the interval \([0, 2\pi]\), and \( n(t) \) is a bandpass noise process. The noise process can be written (using the Rice-representation [10]) as

\[ n(t) = x(t) \cos(\omega_d t) - y(t) \sin(\omega_d t), \quad (5) \]

and the noise variance is \( \sigma_n^2 = \sigma_x^2 = \sigma_y^2 \). The received radio signal can also be written using the Rice-representation,

\[ r(t) = \Re \left[ g(t) e^{j(\omega_c t + \theta)} \right] + \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] + n(t), \quad (6) \]

where \( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] \) is the quadrature-phase component. The in-phase and quadrature-phase components can be obtained by IQ-decomposition,

\[ r(t) = \Re \left[ g(t) e^{j(\omega_c t + \theta)} \right] \cos(\omega_d t) + \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] \sin(\omega_d t), \quad (7) \]

where \( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] \) is the Hilbert transform of \( r(t) \). We thus get

\[ \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] + x(t), \quad \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] + y(t). \quad (8) \]

In the table below the in-phase and quadrature-phase components for the studied modulation types are given.

<table>
<thead>
<tr>
<th>AM</th>
<th>( g(t) = A(1 + m(t)) )</th>
<th>( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = \Re \left[ A(1 + m(t)) \right] \cos(\omega_d t) + x(t) )</th>
<th>( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = \Re \left[ A(1 + m(t)) \right] \sin(\omega_d t) + y(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSB</td>
<td>( g(t) = Am(t) )</td>
<td>( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = \Re \left[ Am(t) \right] \cos(\omega_d t + \theta) + x(t) )</td>
<td>( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = \Re \left[ Am(t) \right] \sin(\omega_d t + \theta) + y(t) )</td>
</tr>
<tr>
<td>SSB</td>
<td>( g(t) = A(\tilde{m}(t) \pm j\tilde{n}(t)) )</td>
<td>( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = \Re \left[ A(\tilde{m}(t) \pm j\tilde{n}(t)) \right] \cos(\omega_d t + \theta) + x(t) )</td>
<td>( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = \Re \left[ A(\tilde{m}(t) \pm j\tilde{n}(t)) \right] \sin(\omega_d t + \theta) + y(t) )</td>
</tr>
<tr>
<td>FM</td>
<td>( g(t) = A \exp \left( j(D_0 \int \tilde{m}(t) dt) \right) )</td>
<td>( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = A \cos(\omega_d t + D_0 \int \tilde{m}(t) dt + \theta) + x(t) )</td>
<td>( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = A \sin(\omega_d t + D_0 \int \tilde{m}(t) dt + \theta) + y(t) )</td>
</tr>
<tr>
<td>CW</td>
<td>( g(t) = A )</td>
<td>( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = A \cos(\omega_d t + \theta) + x(t) )</td>
<td>( \Re \left[ g(t) e^{j(\omega_d t + \theta)} \right] = A \sin(\omega_d t + \theta) + y(t) )</td>
</tr>
</tbody>
</table>

5. The Analysis

The most important issue for this analysis is the assumptions made about the statistical properties of the message and the noise signals.

5.1. The Noise Signals

There are models that give a good description of the noise in the HF-band. The HF-band noise is strongly non Gaussian with typical impulsive characteristics. These models are unfortunately a bit too complicated to be used in this analysis. What we have to do instead is to find a few statistical parameters that characterize the noise appropriately. The important aspects of the noise that the parameters must catch, is that the noise is non-Gaussian and nearly white. The whiteness of the noise signal is assured by the signal processing applied. The received noise signal is after the bandpass filter in the receiver, a bandpass process. This bandpass process can be split into two lowpass processes (the in-phase and quadrature-phase components) by IQ-decomposition. By proper declination of the sampled data, the noise process becomes uncorrelated and thus white.

A parameter encountered in the analysis that will describe the degree of non Gaussianness of the noise is the kurtosis. In this context we define the kurtosis as \( \langle \text{Kurtosis} \rangle = \text{var}^4 \).\( \text{var}^2 \)

\[ n_t = \text{var}^4(\text{Kurtosis}) = \text{var}^4. \quad (9) \]

For a Gaussian signal we have \( n_t = 3 \). The more impulsive the noise becomes, the larger \( n_t \) will be.

The noise bandwidth can be expressed with the RMS bandwidth [11] defined by

\[ \text{RMS Bandwidth} = \frac{\text{var}(\omega)}{\text{var}(\theta)} = \frac{1}{\text{var}(\theta)} \int \omega^2 \text{PSD}(\omega) d\omega = \frac{1}{\text{var}(\theta)} \int \omega^2 \text{PSD}(\omega) d\omega = \frac{D_D}{T^2}. \quad (10) \]

The normalized parameter \( D_D \) (we call it here the RMS-bandwidth factor) is used for simplicity. For white noise we have that \( D_D = 1 \). We can also define \( D_D \) as

\[ D_D = \frac{\text{var}(\omega)}{\text{var}(\theta)} = \frac{1}{\text{var}(\theta)} \int \omega^2 \text{PSD}(\omega) d\omega = \frac{1}{\text{var}(\theta)} \int \omega^2 \text{PSD}(\omega) d\omega = \frac{T^2}{\text{var}(\theta)}. \quad (11) \]

5.2. The Data Signal

For the message signal the characteristics we need are the variance \( \text{var}(\tilde{m}) \) (for AM and FM), the kurtosis \( n_t \), and the RMS bandwidth factor \( D_D \).

For SSB signals we will also need some more parameters \( \text{HD}_\text{m}, M_\text{m}, M_1, M_2 \). These parameters are defined as

\[ \text{HD}_\text{m} = \text{var}(\tilde{m}(t)) \int \frac{d\tilde{m}(t)}{dt} \frac{T}{\text{var}(\tilde{m})}, \quad (12) \]

\[ M_\text{m} = \text{var}(\tilde{m}(t)) \int \frac{d\tilde{m}(t)}{dt} \frac{T}{\text{var}(\tilde{m})}, \quad (13) \]

\[ M_1 = \text{var}(\tilde{m}(t)) \int \frac{d\tilde{m}(t)}{dt} \frac{T}{\text{var}(\tilde{m})}, \quad (14) \]

\[ M_2 = \text{var}(\tilde{m}(t)) \int \frac{d\tilde{m}(t)}{dt} \frac{T}{\text{var}(\tilde{m})}, \quad (15) \]

where \( \tilde{m}(t) \) is the Hilbert transform of \( m(t) \) and \( T \) is the sample time. The parameter \( \text{HD}_\text{m} \) is also a measure for the bandwidth. For a white noise signal we have that \( \text{HD}_\text{m} = 1/2 \). Furthermore, \( M_\text{m}, M_1, M_2 \) is related to both the sample distribution and the bandwidth of the signal in a rather complicated way.
5.3. The Results

The two parameters to compute are

\[
R = \frac{\mathbb{E}[\sigma^2(t) - \mathbb{E}[\sigma^2(t)]]^2}{\mathbb{E}[\sigma^2(t)]^2} = \frac{\mathbb{E}[\sigma^2(t)] - \mathbb{E}[\sigma^2(t)]^2}{\mathbb{E}[\sigma^2(t)]^2} \tag{16}
\]

and

\[
R_a = \mathbb{E}[(\omega(\omega) - \mathbb{E}[(\omega)])^2] \tag{17}
\]

In the last expression we defined the instantaneous frequency as

\[
\omega(t) = \Im \left[ \frac{dr(t)/dt}{r(t)} \right] = \frac{\gamma(t)(\gamma(t) - \gamma(t))}{\gamma(t) + \gamma(t)} \tag{18}
\]

where \(r(t) = \gamma(t) + i\gamma(t)\). In the table below the computed parameters are presented.

| AM | \[
R = \left( \frac{(\alpha - 1)\sigma^2 + 4\sigma^2}{\Gamma + \frac{1}{\alpha}} \right)^2 + \frac{(1 + \sigma^2)^2}{\Gamma + \frac{1}{\alpha}} \right)^2
\]

| DSB | \[
R = \left( \frac{(\alpha - 1)\sigma^2 + 4\sigma^2}{\Gamma + \frac{1}{\alpha}} \right)^2 + \frac{(1 + \sigma^2)^2}{\Gamma + \frac{1}{\alpha}} \right)^2
\]

| SSB | \[
R = \left( \frac{(\alpha - 1)\sigma^2 + 4\sigma^2}{\Gamma + \frac{1}{\alpha}} \right)^2 + \frac{(1 + \sigma^2)^2}{\Gamma + \frac{1}{\alpha}} \right)^2
\]

| FM | \[
R = \left( \frac{(\alpha - 1)\sigma^2 + 4\sigma^2}{\Gamma + \frac{1}{\alpha}} \right)^2 + \frac{(1 + \sigma^2)^2}{\Gamma + \frac{1}{\alpha}} \right)^2
\]

| CW | \[
R = \left( \frac{(\alpha - 1)\sigma^2 + 4\sigma^2}{\Gamma + \frac{1}{\alpha}} \right)^2 + \frac{(1 + \sigma^2)^2}{\Gamma + \frac{1}{\alpha}} \right)^2
\]

| NOISE | \[
R = \left( \frac{(\alpha - 1)\sigma^2 + 4\sigma^2}{\Gamma + \frac{1}{\alpha}} \right)^2 + \frac{(1 + \sigma^2)^2}{\Gamma + \frac{1}{\alpha}} \right)^2
\]

In these expressions we have that \(\text{CNR} = \text{Carrier to Noise Ratio}\) = \(\gamma^2/2\alpha^2\) and \(\text{MNR} = \text{Message to Noise Ratio}\) = \(\gamma^2/2\alpha^2\). For SSB modulation we have \(\text{MNR} = \gamma^2/2\alpha^2\).

6. The Classifier

The modulation classifier can now be designed based on the expressions presented in Section 5 and some realistic values for the used signal characteristics.

6.1. Decision Regions

The simplest method for finding the needed decision regions is to simulate different signal characteristics and use the expressions from Section 5. One can then locate the regions in the \(R / R_a\) space that is covered by the different modulation types. The different decision regions are determined so minimal classification errors will be generated. The decision regions defined in this way are realistic for all modulation types except for SSB modulated signals, which will have a too broad decision area. The reason for this is that all simulated combinations of the signal characteristics are not physical possible.

In the table below we present typical values of the parameters used for the simulations. These values are based on real signals.

<table>
<thead>
<tr>
<th>The message signal</th>
<th>(\sigma^2)</th>
<th>(0.01 - 0.10)</th>
<th>For AM and FM signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(m_b)</td>
<td>(2.5 - 8.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D_m^p)</td>
<td>(0.01 - 0.20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(H_b)</td>
<td>(0.12 - 0.35)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M_b)</td>
<td>(5 - 25)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M_b)</td>
<td>(1 - 6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M_b)</td>
<td>(0.3 - 18)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The noise signal</th>
<th>(n_b)</th>
<th>(3 - 9)</th>
<th>After some pre-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D_m^p)</td>
<td>(0.22 - 0.36)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Carrier offset</th>
<th>(\Delta\omega)</th>
<th>(-0.2 - 0.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\Delta\omega)</td>
<td>(0.22 - 0.36)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FM</th>
<th>(D_m^p)</th>
<th>(2)</th>
</tr>
</thead>
</table>

Based on these parameters and \(\text{CNR}\) or \(\text{MNR}\) values in the interval \([10 \text{ dB} - 30 \text{ dB}]\) we get the decision regions shown in figure 1.

![Figure 1: The decision regions for the classifier.](image)

A potential problem with the derived decision regions is that the noise region is inside the SSB region. There is thus a need for some other characteristics that will aid in differentiation between SSB and noise signals. A unique property for SSB signals that can be used for this purpose is that the SSB signals have single sided power spectra (only one of the side bands is used). We thus introduce the parameter \(\overline{R_a}\), which corresponds to the mean of the instantaneous frequency. We have that

\[
\overline{R_a} = \mathbb{E}[\omega(\omega)] = \frac{\mathbb{E}[\sigma^2(t) - \gamma(\gamma - \gamma)]}{\mathbb{E}[\sigma^2(t)]} \tag{19}
\]
For noise signals, this new parameter will be zero and for SSB signals we have the expression
\[
\overline{R}_m = (1 + MN R^{-1})^{-1} (\Delta \omega T + HD_m \pi)
\]
(20)

For reasonable large \( MN R \) values we get the approximation \( \overline{R}_m \approx \Delta \omega T + HD_m \pi \). By computing the \( \overline{R}_m \) parameter and comparing it to a threshold, we can distinguish between SSB signals and noise. With the typical values of the signal characteristics presented above we can conclude that the condition \( |\overline{R}_m| > -0.2 + 0.12x = 0.2 \) holds for an SSB signal. By investigating the sign of \( \overline{R}_m \) we can decide if the SSB signal is of USB or LSB type.

6.2. The Confusion Matrix

To evaluate the studied modulation classifier, we analyzed 500 simulated AM, DSB, SSB, FM, CW, and noise signals. Each signal contains 1024 samples which approximately corresponds to 0.25 sec. of data. The following result was obtained for \( CNR=15 \) dB or \( MN R=15 \) dB, whatever applicable where the counts are doubled in the table.

<table>
<thead>
<tr>
<th>15 dB</th>
<th>AM</th>
<th>DSB</th>
<th>SSB</th>
<th>FM</th>
<th>CW</th>
<th>NOISE</th>
<th>U.K.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DSB</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SSB</td>
<td>0</td>
<td>6</td>
<td>662</td>
<td>0</td>
<td>0</td>
<td>294</td>
<td>48</td>
</tr>
<tr>
<td>FM</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CW</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NOISE</td>
<td>0</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>920</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

We see clearly in the table the confusion between SSB and noise signals. This shows the need for the suggested new feature. The table below shows the confusion matrix generated with the new parameter used for distinguishing between SSB and noise signals.

<table>
<thead>
<tr>
<th>15 dB</th>
<th>AM</th>
<th>DSB</th>
<th>SSB</th>
<th>FM</th>
<th>CW</th>
<th>NOISE</th>
<th>U.K.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DSB</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SSB</td>
<td>0</td>
<td>6</td>
<td>946</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td>FM</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CW</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NOISE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Some preliminary tests with real signals have been carried out when this paper was written. The preliminary result shows that the modulation classifier works, but there is not enough with data to compile a confusion matrix.

7. Some Practical Considerations

An important aspect for real implementations is the pre-processing. It is important to ensure that the conditions, for which the modulation classifier is designed are met by the radio signals to classify. Typical pre-processing activities are:

1. **Remove noise segments from data.** Noise segments in the data confuse the classifier since they make the data look more like noise than the modulation type the data represent.
2. **Compensate for fading.** The fading of the signal will cause an unwanted amplitude modulation.
3. **Outlier removal.** Large amplitude peaks in the signals, caused by disturbances, must be removed since they alter the statistical properties of the amplitude.

4. **Removal off interfering signals.** To be able to use the mean value of the instantaneous frequency, \( \overline{R}_m \), for differentiation between SSB and noise signals it must be assured that the noise do not contain interfering signal components that can give too large \( \overline{R}_m \) values for noise.

The analysis of the instantaneous frequency depends on a correct differentiation of the signals. The classical numerical differentiation methods all have some pitfalls. There are two aspects of differentiation that is important. Firstly, a signal and its differentiation are uncorrelated. Secondly, the differentiator should not lose bandwidth compared to a true differentiator. The solution to these problems is to do the differentiation in the frequency domain. Thus we compute the differentiation as:

\[
\frac{dn(t)}{dt} = \text{FFT}^{-1} \{ \text{FFT} \{ n(t) \} \}.
\]
(21)

8. Conclusions and Further Work

By studying the variability of the amplitude and frequency of modulated radio signals a modulation classifier was developed. Based on an extensive statistical analysis of the radio signals with realistic estimates of used statistical parameters appropriate decision regions.

The next step in this development is to work on the important pre-processing, which must be present in a system that is working with real radio signals. A neural net based classifier shall also be developed. We will also continue the analysis, so both the decision regions and the confusion matrices can be computed with the expression presented in Section 5.

References

Coexistence of Asymmetric Digital Subscriber Lines using Discrete Multitone signalling with T1 Lines

Thierry POLLET and Marc MOENECIALEY

Communication Engineering Lab., University of Ghent  
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium  
tel + -32-(0)9-2643413, fax + -32-(0)9-264 42 95  
E-mail: Thierry.Pollet@ict.rug.ac.be, Mark.Moenecialey@ict.rug.ac.be

Abstract. Asymmetric Digital Subscriber Lines (ADSL) offer a high speed digital channel from the central office to the subscriber (downstream channel) over a single unshielded twisted pair (UTP). The downstream rate can vary from 1.5-6 Mb/sec; while in the upstream direction (subscriber to the central office) a much lower data rate (64kb/sec-600kb/sec) is foreseen. However, other services coexist with ADSL, such as T1 (1.544 Mb/sec), HDSL(2x800 kb/sec), Basic rate ISDN (160 kb/sec). It can be shown that ADSL systems suffer the most from T1 interference. In this contribution, we determine the maximal achievable distance for ADSL carrying different rates, depending on the number of T1 interferers in an adjacent binder. We consider ADSL systems where the spectral occupancy of the upstream and downstream channel overlap, and others where both spectra are disjoint.

1. Introduction

Most likely, ADSL systems will use orthogonal frequency division multiplexing (OFDM) because of their lower implementation complexity and higher flexibility in terms of bandwidth allocation compared to equivalent single carrier systems [1-3]. They have also been referred to as Discrete Multitone systems (DMT). This modulation technique has been standardized for ADSL within the T1-E1 committee. In this paper, we calculate the effect of T1 carriers in adjacent binders on the achievable distance for ADSL systems.

2. DMT signalling

The DMT signal basically consists of a large number of QAM modulated carriers with a carrier frequency spacing equal to the symbol rate at each individual carrier. During one OFDM symbol period, T, the complex envelope of the signal can be expressed as:

\[ s(t) = \sum_{n=0}^{N-1} a_n e^{j2\pi nft} \]

where \( F = 1/T \) equals the carrier spacing, and \( a_n \) denotes the transmitted QAM constellation point modulating the \( n \)-th carrier.

In order to reduce the inter-symbol interference and intercarrier interference, a guard time interval of duration \( vT \) is inserted within each symbol, resulting in an increase of spacing between the carriers. The signal transmitted during each OFDM symbol becomes

\[ s(t) = s(t - \frac{v}{1-v} - vT) \]

The maximal achievable data rate corresponding to a given distance and bit error rate, can be obtained by optimizing the QAM constellation size and power for each carrier, according to the signal-to-noise-ratio (SNR) at each specific carrier. The number \( n(i) \) of bits per symbol of the \( i \)-th carrier yielding a BER of \( 10^{-3} \) is given by [4]

\[ n(i) = \log_2 \left( 1 + \frac{SNR(f_i)}{10^g} \right) \]

where \( g \) denotes the coding gain (in dB). The total bit rate \( R \) can be expressed by

\[ R = \sum_{i=0}^{N-1} n(i) F (1-S) \]

where \( S \) denotes the fraction of inserted synchronization symbols in the total symbol stream, and \( N \) the total number of carriers.

3. S-ADSL versus EC-ADSL

The signals corresponding to 1 ADSL service (upstream and downstream channel) are transmitted over 1 UTP. Interference originating from ADSL signals carried over other UTP wires within the same binder strongly differs depending on whether the spectra of the downstream and upstream channels coincide or partially overlap. In the following, we denote the systems where the spectra are disjoint by S-ADSL (Split-band), and where their spectra partly overlap by EC-ADSL (Echo Canceled), respectively.

4. Impairments

3.1 Self-NEXT and self-FEXT

The useful signal is corrupted by crosstalk: FEXT or NEXT depending on whether the disturbing signals transport data in the same or in the opposite direction of the considered one. We suppose that all wires in the binder carry ADSL signals, while the adjacent binder carries only T1 signals.

For S-ADSL, the downstream channel is not effected by the self-NEXT from the ADSL upstream channels on other wires. The only contribution from the other ADSL services is self-FEXT.

For EC-ADSL, self-NEXT is the main disturbance in the frequency range where the spectra of the upstream and downstream channel coincide, whereas in the non overlapping area, only self-FEXT is present. Figure 1. depicts the contributions of self-NEXT and self-FEXT for S-ADSL and EC-ADSL.
3.2 T1 influence

Most likely, T1 and ADSL services will not be used in the same binder, because the T1 interference is distinctive for ADSL services. Therefore, we consider only the situation where T1 and ADSL signals are present in adjacent binders. We suppose that the T1 interference originating from the adjacent binder is about 10 dB less comparing with the interference from T1 carriers within the same binder. It can easily be shown that the interference from T1 carriers in the adjacent binder (50 wires per binder) becomes:

$$S_{T1}(f) = 0.1 \times K_{next} P_{T1}(f) \frac{(1-e^{-2\alpha(f)d'})}{\alpha(f)d'}$$

where d' denotes the distance between the T1 repeaters, and q(f) is the attenuation of the UTP wire. P_{T1}(f) is the transmitted T1 power spectrum. K_{next} is the NEXT coupling constant.

3.3 ECHO

At each receiver side of the UTP, an echo from the co-located transmitter disturbs the useful signal. In case the upstream channel and downstream channel spectrally overlap, the echo can only be appropriately removed by an adaptive echo canceller. A good canceller reduces the residual echo power several tens of dB below the useful signal power. The echo power density (Psd_echo) is assumed to be uniform over the useful signal band.

In case the spectra do not overlap, a fixed filter is sufficient to eliminate the echo. However, the spectral spacing between the upstream and downstream channel has to be large enough due to practical filter restrictions.

3.4 AWGN

The AWGN is the sum of the environmental noise and the thermal noise; its power spectral density N0 is conservatively estimated to be -140 dBm/Hz.

4. SNR as a function of frequency

The SNR at a given frequency f for the 2 considered ADSL systems is expressed as follows.

EC-ADSL (in the frequency region where there is spectral overlap between the upstream and downstream channel):

$$SNR(f) = \frac{P_{c}e^{-2\alpha(f)d}}{N_{0} + S_{T1}(f) + P_{k_{next}}K_{next}^{2\alpha} + Psd_{echo}}$$

EC-ADSL (in the frequency region without spectral overlap):

$$SNR(f) = \frac{P_{c}e^{-2\alpha(f)d}}{N_{0} + S_{T1}(f) + P_{k_{next}}K_{next}^{2\alpha} e^{-2\alpha(f)d} + Psd_{echo}}$$

S-ADSL:

$$SNR(f) = \frac{P_{c}e^{-2\alpha(f)d}}{N_{0} + S_{T1}(f) + P_{k_{ext}}^{2\pi} e^{-2\alpha(f)d} + Psd_{echo}}$$

P is the ADSL transmit power density, and d is the total length of the ADSL cable.

5. Simulation and Results

Table 1 lists the values of the different parameters used.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cable</td>
<td>26 AWG, 50 pair per binder</td>
</tr>
<tr>
<td>gain g (due to error correction)</td>
<td>4 dB</td>
</tr>
<tr>
<td>S</td>
<td>1/69</td>
</tr>
<tr>
<td>F(1-\nu)</td>
<td>4.3125 kHz</td>
</tr>
<tr>
<td>K_{next}</td>
<td>1/1.134 \text{10}^{13} Hz^{-1.5}</td>
</tr>
<tr>
<td>K_{ext}</td>
<td>10^{-6} kHz^{-1} Hz^{-2}</td>
</tr>
<tr>
<td>Repeater spacing d'</td>
<td>3 kft</td>
</tr>
<tr>
<td>T1 power</td>
<td>15 dBm</td>
</tr>
<tr>
<td>ADSL power density P</td>
<td>20 dBm/BW (BW is the used frequency bandwidth)</td>
</tr>
<tr>
<td>lowest frequency where carriers can be placed for the upstream channel (S-ADSL):</td>
<td>20 kHz</td>
</tr>
<tr>
<td>lowest frequency where carriers can be placed for the downstream channel (S-ADSL):</td>
<td>138 kHz</td>
</tr>
<tr>
<td>lowest frequency where carriers can be placed for the downstream and upstream channel (EC-ADSL):</td>
<td>20 kHz</td>
</tr>
<tr>
<td>Maximum number of bits per carrier</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 1. Parameters used during simulation

Figure 2 depicts the achievable distance for both EC-ADSL and S-ADSL as a function of the number of T1 interferers in the adjacent binder. The total bit rate R was 6 Mb/sec. For EC-ADSL, the residual echo power has been chosen 40 dB and 35 dB below the received useful signal power, respectively.

Figure 3 shows the achievable distance for different bit rates for S-ADSL as a function of the number of T1 interferers in the adjacent binder.

In order to show that T1 interference is the dominating disturbance, we have plotted in Figure 4 the achievable distance as a function of
the bit rate, considering either only AWGN, FEXT and (for EC-ADSL) NEXT. For EC-ADSL we assumed that the echo canceller could completely eliminate the echo. The upstream rates were set to 64 Kb/sec and 600Kb/sec. Note that for S-ADSL, the achievable distance does not depend on the upstream bit rate. We observe that, in the absence of T1 interference, a considerably larger distance can be achieved.

![Achievable distance on 26 AWG T1 interference from adjacent binder](image)

Figure 2. Achievable bit rates for S-ADSL and EC-ADSL.

5. Conclusions

From Figure 2, we conclude that EC-ADSL performs better than S-ADSL only if the residual echo is sufficiently suppressed by the echo canceller. Comparing figures 3 and 4 we notice that T1 interference from an adjacent binder is more detrimental than the interference originating from ADSL wires in the same binder. Finally, when T1 interference is absent, we see that if a too high upstream rate is chosen, the self-NEXT will determine the achievable distance for EC-ADSL. For S-ADSL, we conclude that the achievable distance is limited by the AWGN impairment.

![Achievable distance on 26 AWG](image)

Figure 3. Achievable bit rates for S-ADSL.

![Achievable distance on 26 AWG](image)

Figure 4. Achievable distance in the absence of T1 interference.

Acknowledgement

The first author expresses his gratitude to ALCATEL BELL for the financial support. The second author gratefully acknowledges the support from the Belgian National Fund for Scientific Research (NFWO).

References


Design of M-Channel Wavelet Transform Bases With Maximum Regularity

Hakan ÇAŁAR ¹, Oktay ALKIN³, Bülent SANKUR² Emin ANARIM²,
¹Marmara Research Center, TUBITAK, 41470, Gebze-Kocaeli, TURKEY
³Dept. of Electrical Engineering, Southern Illinois University at
   Edwardsville, Illinois 62026-1801, USA
²Dept. of Electrical Engineering, Bogazici University, 80815, Bebek, Istanbul-TURKEY
   Tel/Fax: +90(212)263 15 40 - 1455, [212] 257 50 30, E-Mail:sankur@trboun.bitnet

Abstract
This study presents a technique for orthonormal M-channel regular wavelet bases where \( M = 2^l \). This time domain design technique is based on finding a suitable low-pass subband filter \( h_0(n) \) and the remaining (M-1) bandpass filters are constructed by the shuffling operations. This leads to a set of subband filters corresponding to the wavelet bases that use the same numerical coefficient values in different shift positions, allowing very efficient numerical implementation. Regularity conditions of the M-channel wavelet bases have also been investigated vis-à-vis coding performance. Some design examples are provided.

1 Introduction
The theory of Perfect-Reconstruction Quadrature Mirror Filter (PR-QMF) banks has been studied extensively [1-9]. It has been shown by Mallat [5] that the multiresolution form of the orthonormal wavelet transform is functionally equivalent to the analysis section of a PR subband coder with regularity property. In this paper, we introduce a design technique for symmetric orthonormal M-channel wavelets with desired regularity.

Using a mother wavelet \( \psi(t) \), the Continuous Wavelet Transform (CWT) of a function \( f(t) \in L^2(\mathbb{R}) \) is given by

\[
W_f(a,b) = \langle f(t), \psi_{ab}(t) \rangle
\]

where \( \psi_{ab}(t) \) represents a family of functions obtained from \( \psi(t) \) using dilation and translation parameters \( a > 0 \) and \( b \), i.e.,

\[
\psi_{a,b}(t) = \frac{1}{|a|^{1/2}} \psi \left( \frac{t-b}{a} \right)
\]

In practical applications of the wavelet transform, its discrete form is usually preferred. For this purpose, dilation and translation parameters \( a \) and \( b \) are chosen to lie on a rectangular grid of points in the time-scale plane, i.e.,

\[
a = a_0^m, \quad b = nb_0a_0^m
\]

with \( a_0 > 1 \) and \( b_0 > 0 \). The most widely studied form of the Discrete Wavelet Transform (DWT) is for \( a_0 = 2 \) [3]. For discrete signals, a multiresolution form of (1) was first introduced by Mallat [5], and an extension to the M-channel case is relatively straightforward.

2 M-Channel Orthogonal Wavelet Transform
Consider a scaling function \( \phi(t) \) that satisfies the two-scale equation

\[
\phi(t) = \sum_n p_n^{(0)} \phi(Mt - n)
\]

where \( p_n^{(0)} \) denote the coefficients of the M-scale transform along with \( (M-1) \) wavelets \( \psi^{(i)}(t) \) that satisfy

\[
\psi^{(i)}(t) = \sum_n p_n^{(i)} \phi(Mt - n), \quad i = 1, \ldots, M-1
\]

The scaling and wavelet functions at larger scales will be defined by the indexed functions \( \phi_{mn}(t) \) and \( \psi_{mn}^{(i)}(t) \) in terms of \( \phi(t) \) and \( \psi^{(i)}(t) \) as

\[
\phi_{mn}(t) = M^{-m/2} \phi(M^{-m}t - n)
\]

and

\[
\psi_{mn}^{(i)}(t) = M^{-m/2} \psi^{(i)}(M^{-m}t - n), \quad i = 1, \ldots, M-1
\]

where parameter values \( a_0 = M \) and \( b_0 = 1 \) are used. Let the successive approximation subspaces \( V_m \) and \( W_m^{(i)} \) be defined as function subspaces spanned by \( \phi_{mn}(t) \) and \( \psi_{mn}^{(i)}(t) \) for fixed \( m \), i.e.,

\[
V_n = \text{clos}_{L^2(\mathbb{R})} \phi_{mn}; \forall n \in \mathbb{Z}
\]

\[
W_n^{(i)} = \text{clos}_{L^2(\mathbb{R})} \psi_{mn}^{(i)}; \forall n \in \mathbb{Z}
\]

As a consequence of the M-scale equation in (4), \( V_m \) subspaces have the property

\[
\ldots \subset V_1 \subset V_0 \subset V_{-1} \subset \ldots
\]
If the scaling function $\phi(t)$ and the wavelets $\{\psi_i(t); i = 1, \ldots, M - 1\}$ are chosen to satisfy the orthonormality conditions

$$
< \phi_0(t), \phi_0(t) > = \delta_{n-1} \\
< \phi_m(t), \psi_{i,j}^k(t) > = 0, \quad i = 1, 2, \ldots, M - 1 \\
< \psi_{mn}^k(t), \psi_{i,j}^k(t) > = \delta_{n-k} \delta_{m-i} 
$$

then subspaces $\{V_m\}$ form an orthogonal decomposition of the $L^2(\mathbb{R})$ function space, and are related to subspaces $\{W_m^i\}$ by

$$
V_m = V_{m+1} \oplus \left[ \bigoplus_{i=1}^{M-1} W_{m+1}^i \right] 
$$

Note that the symbol $\oplus$ denotes orthogonal sum.

For a given discrete sequence $x(n) \in l^2(\mathbb{R})$, a continuous function $f(t) \in L^2(\mathbb{R})$ can be constructed in the form

$$
f(t) = \sum_n x(n) \phi(t - n) 
$$

We can write $f(t)$ as the sum of its projections onto subspaces $V_i$ and $W_i$ as

$$
f(t) = \sum_n v_0(n) \phi_1(t) + \sum_{i=1}^{M-1} \sum_n v_i(n) \psi_i^1(t) 
$$

where the original sequence is implicitly assumed to be at scale $i=0$. Using the orthonormality of basis functions as established by (11), the coefficients $v_i(n)$ of this expansion can be expressed with inner products as

$$
v_i(n) = \begin{cases} 
< f, \phi_{i,n} > & i = 0 \\
< f, \psi_{i,n} > & i = 1, \ldots, M - 1 
\end{cases} 
$$

Using the two-scale equations (4) and (5) in (15), it can be shown that

$$
v_i(n) = \frac{1}{\sqrt{M}} \sum_k x(k) P_{MN-k}^i 
$$

which is equivalent to processing the sequence $x(n)$ with a set of linear time-invariant filters with impulse responses $h_i(n) = \sqrt{M} P_{MN-k}^i$, and downsampling filter outputs by $M$.

It can easily be shown that (11) is equivalent to

$$
\sum_n h_i(n) h_j(n - kM) = \delta_k \delta_{i-j} 
$$

## 3 Design Technique of Wavelet Bases

We will now present a technique for the design of a set of $M$ linear-phase analysis filters that are orthonormal in the sense of equation (17) so that the resulting subband coder has the perfect reconstruction property and gives us a symmetric M-channel orthonormal wavelet bases.

Let the column-vector $h_k(n)$ be defined as

$$
h_k(n) = [ h_k(0) \ h_k(1) \ \cdots \ h_k(N-1) ]^T 
$$

Also, let $W$ be the one-block shift matrix defined as

$$
W = \begin{bmatrix} 0 & I_{N-M} \\ 0 & 0 \end{bmatrix}_{N \times N} 
$$

so that we can write the orthonormality condition of (17) in compact form as

$$
h_i^T W^k h_j = \delta_k \delta_{i-j} 
$$

Assume that the first of $M$ filters, $h_0$, is available, has type-I linear phase (positive symmetry), and satisfies the orthogonality condition

$$
h_i^T W^k h_0 = \delta_k, \quad k = 0, \ldots, \frac{N}{M} - 1 
$$

to qualify as one of the filters of an $M$-band orthogonal perfect-reconstruction coder. The remaining $(M - 1)$ filters will be defined in terms of $h_0$ as

$$
h_i = B_i h_0, \quad i = 1, \ldots, M - 1 
$$

where the matrices $B_i$, $i = 1, \ldots, M - 1$ will be termed filter transformation matrices. Substituting (22) into (21), we have

$$
h_i^T C_{i,j,k} h_0 = \delta_k \delta_{i-j} 
$$

where we used the convention

$$
C_{i,j,k} = B_i^T W^k B_j 
$$

At this point, some definitions will be made. Let the square matrix $P_K$ be defined as

$$
P_K = [p_{ik}]_{K \times K}, \quad p_{ik} = (-1)^{\delta_{i-k}} 
$$

Thus, $P_K$ is a diagonal matrix with alternating $+1$'s and $-1$'s on its main diagonal. Using this, we will define the $N$-th order shuffling matrix of level $i$ as

$$
A_i = P_{N/2^i} \otimes J_{2^i} 
$$

where $J$ is the counter-identity matrix, and the symbol $\otimes$ denotes the Kronecker product of two matrices. Filter transformation matrices are constructed using ordered permutations of the shuffling matrices defined above. Let the filter transformation matrix $B_j$ be defined as

$$
B_j = \prod_{i=0}^{M/2-1} A_{i}^{r_{ji}}, \quad j = 0, \ldots, M/2 - 1 
$$

In (27), the constants $r_{ji}$ take on values from the finite set $\{0, 1\}$ based on the $i$-th least significant bit in the
binary representation of the integer \( j \). Note that (27) defines only the first \( M/2 \) filter transformation matrices needed. We will obtain the remaining \( M/2 \) matrices as

\[
B_{M-j} = P_j B_j, \quad j = 0, \ldots, M/2 - 1 \quad (28)
\]

Thus, once a suitable \( h_0 \) is found, the remaining \( (M-1) \) filter vectors are computed using (22) with the filter transformation matrices given by (27) and (28). A suitable \( h_0 \) can be found using under constraint in (23).

4 Regularity of Wavelet Bases

Without special precautions in the design of the \( M \)-band coder, the iterations on the filter coefficients leading to wavelet bases may result in fractal-type functions for \( \psi(t) \) and \( \psi^{(i)}(t) \). Regularity issues have studied in detail in [4]. The main results are summarized in the following two lemmas:

**Lemma 1** For the first \( K \) moments of the wavelet \( \psi^{(i)}(t) \) to be equal to zero, the \( z \)-domain transfer function of the corresponding subband filter must have a \( K \)-th order zero at \( z = 1 \) \[4\].

**Lemma 2** If filters \( H_i(z), i = 1, \ldots, M - 1 \) of the \( M \)-band coder have \( K \)-th order zeros at \( z = 1 \), then the lowpass filter transfer function \( H_0(z) \) has \( K \)-th order zeros at points \( z = \exp(2\pi nk/M), k = 1, \ldots, M - 1 \) \[4\].

Recall that, for the design technique proposed, the number of bands \( M \) was restricted to be an integer power of 2. As a result of Lemma 2, the lowpass filter transfer function \( H_0(z) \) must be in the form

\[
H_0(z) = \left[ \prod_{i=0}^{\log_2(M)-1} (1 + z^{-2^i}) \right]^K Q_0(z) \quad (29)
\]

where \( Q_0(z) \) is a polynomial in \( z^{-1} \).

Time-frequency localization of various 4-band PR subband filter banks solutions are tabulated in Table (II). It has been observed that increasing the number of vanishing moments on wavelet filters leads to decreasing time-frequency localization. Notice that the time localization degrades significantly more as compared to the frequency localization.

The scaling function and the three wavelet bases of a 32-tap 4 band linear phase PR design are depicted in Fig.1. In Fig.2, we compare the coding gain for a Markov-1 source model with correlation coefficients 0.95 as a function of the number of zeroes at half the Nyquist frequency, which is related to regularity order. It is clear from the figure that having a few zeroes at Nyquist frequencies is desirable but increasing the regularity will cause a degradation in frequency selectivity, and consequently a decrease in coding gain performance.

6 Conclusions

In this work, we have presented a new time-domain \( M \)-band orthogonal design technique for linear phase filter banks with PR property. These filter banks lead to very efficient computational structures. We also demonstrate that resulting filter banks can be used for construction of compactly supported orthonormal and symmetric wavelet bases having good regularity.

**Acknowledgements**

This research was supported by TUBITAK EEEAG-83 project.

**References**


Table I: Filter coefficients for 16-tap 4-band coder (only half the number of coefficients are shown due to symmetry).

<table>
<thead>
<tr>
<th>n</th>
<th>$h_0(n)$</th>
<th>$h_1(n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.030550699823</td>
<td>0.019908110237</td>
</tr>
<tr>
<td>1</td>
<td>-0.019908110237</td>
<td>-0.030550699823</td>
</tr>
<tr>
<td>2</td>
<td>-0.038475203055</td>
<td>-0.038475203055</td>
</tr>
<tr>
<td>3</td>
<td>-0.038475203055</td>
<td>-0.038475203055</td>
</tr>
<tr>
<td>4</td>
<td>-0.036706281658</td>
<td>-0.18841015419</td>
</tr>
<tr>
<td>5</td>
<td>0.168841015419</td>
<td>0.036706281658</td>
</tr>
<tr>
<td>6</td>
<td>0.409542299873</td>
<td>0.544260466366</td>
</tr>
<tr>
<td>7</td>
<td>0.544260466366</td>
<td>0.409542299873</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>$h_2(n)$</th>
<th>$h_3(n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.019908110237</td>
<td>0.030550699823</td>
</tr>
<tr>
<td>1</td>
<td>0.030550699823</td>
<td>0.019908110237</td>
</tr>
<tr>
<td>2</td>
<td>-0.03104883692</td>
<td>-0.058475205055</td>
</tr>
<tr>
<td>3</td>
<td>0.058475205055</td>
<td>0.038475203055</td>
</tr>
<tr>
<td>4</td>
<td>0.168841015419</td>
<td>-0.036706281658</td>
</tr>
<tr>
<td>5</td>
<td>0.409542299873</td>
<td>0.168841015419</td>
</tr>
<tr>
<td>6</td>
<td>0.544260466366</td>
<td>-0.409542299873</td>
</tr>
<tr>
<td>7</td>
<td>-0.409542299873</td>
<td>-0.544260466366</td>
</tr>
</tbody>
</table>

Table II: Time-frequency localization of 4-band filters with respect to the number of vanishing moments.

<table>
<thead>
<tr>
<th></th>
<th>$\sigma_f^2$</th>
<th>$\sigma_n^2$</th>
<th>$\sigma_f^2$</th>
<th>$\sigma_n^2$</th>
<th>No. of Van. Moments</th>
</tr>
</thead>
<tbody>
<tr>
<td>15T4B</td>
<td>7.47</td>
<td>1.82</td>
<td>13.59</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.47</td>
<td>1.79</td>
<td>13.37</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.46</td>
<td>1.83</td>
<td>12.61</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>24T4B</td>
<td>7.54</td>
<td>2.36</td>
<td>17.82</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.54</td>
<td>2.36</td>
<td>17.83</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.50</td>
<td>1.79</td>
<td>13.50</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>32T4B</td>
<td>7.56</td>
<td>2.99</td>
<td>22.68</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.56</td>
<td>2.86</td>
<td>21.65</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.57</td>
<td>2.51</td>
<td>19.02</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Scaling function and wavelets generated from 32-tap 4-band coder.

Figure 2: Variation of gain versus vanishing moments.
TIME-DOMAIN DESIGN OF MULTIRATE FILTER BANKS AND WAVELETS

A. AL-ADNANI, R. CHAPMAN, and T.S. DURRANI

Signal Processing Division, Dept. of Electronic and Electrical Engineering
University of Strathclyde, Scotland, Glasgow G1 1XW, UK.
email: adnani@uk.ac.strathclyde.eee.sdp

Abstract A new time-domain methodology for designing FIR multirate filter banks is proposed. The design is based on a time-domain formulation in which the perfect reconstruction conditions are formulated as a matrix system of equations. A recursive technique is proposed which uses the synthesis filters from one iteration to update the analysis filters for the next. It is shown that if the perfect reconstruction conditions are formulated in the time domain it is possible to control the system delay and produce low delay multirate filter banks and wavelets. This will prove advantageous in time critical applications.

1 Introduction

Multirate filter banks have received considerable attention in the past few years, because of their wide variety of applications in areas such as subband coding and data compression. Similarly, wavelet transforms are now accepted as a popular signal processing tool, especially for non-stationary signal analysis.

This paper will exploit the close relationship between the discrete wavelet transform and multirate filter banks, and use a time-domain approach to design both low delay wavelets and filter banks. This paper extends the time domain approach to multirate filter banks introduced in [1] and refined in [2].

2 Time-Domain Formulation

In this section a general time domain formulation will be derived which relates the analysis and synthesis filters coefficients from the z-domain perfect reconstruction (PR) conditions.

Consider the M-channel filter bank shown in Figure 1.

For two channels (M=2) the input/output relationship can be written as:

\[ \hat{X}(z) = \frac{1}{2} [H_0(z)F_0(z) + H_1(z)F_1(z)]X(z) + \frac{1}{2} [H_0(-z)F_0(z) + H_1(-z)F_1(z)]X(-z) \]

\[ = T(z)X(z) + A(z)X(-z) \]  (1)

which is a linear combination of the LTI system transfer function \( T(z) \) and the aliasing term \( A(z) \). For aliasing cancellation \( A(z) \) should be zero, and for perfect reconstruction \( T(z) \) must be a pure delay.

\[ \begin{bmatrix} \frac{1}{2} [H_0(z) & H_1(z) \] \begin{bmatrix} F_0(z) \\ F_1(z) \end{bmatrix} \right] = \begin{bmatrix} z^{-\Delta} \\ 0 \end{bmatrix} \]  (2)

Without loss of generalization let the analysis and synthesis filters be FIR with lengths \( N = 2L \) (L integer)

\[ \begin{bmatrix} h_0(0) & \ldots & h_0(N-1) \\ h_1(0) & \ldots & h_1(N-1) \end{bmatrix} \]

\[ \begin{bmatrix} f_0(0) & \ldots & f_0(N-1) \\ f_1(0) & \ldots & f_1(N-1) \end{bmatrix} \]  (3)

We can write eqn.(3) in terms of \((2 \times 1)\) block matrices as

\[ \mathbf{H} = \begin{bmatrix} H_0 & \ldots & H_{N-1} \end{bmatrix} \]

\[ \mathbf{F} = \begin{bmatrix} F_0 & \ldots & F_{N-1} \end{bmatrix} \]  (4)
From eqn.(1)

\[ H_0(z)F_0(z) + H_1(z)F_1(z) = \]

\[
\begin{bmatrix}
H_0^T & 0^T & \cdots \\
H_1^T & H_0^T & \cdots \\
\vdots & H_1^T & \ddots \\
H_{N-1}^T & \cdots & \cdots & H_{N-2}^T \\
0^T & H_{N-1}^T & \cdots & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
0^T & \cdots & \cdots & 0^T \\
\end{bmatrix}
\begin{bmatrix}
F_0 \\
F_1 \\
\vdots \\
F_{N-1} \\
\end{bmatrix}
= 
\begin{bmatrix}
1 \\
z^{-1} \\
\vdots \\
z^{-S} \\
\end{bmatrix}
\]  
(5)

where \( S = 2N - 2 \).
Magnitude and phase errors are eliminated by setting all the terms to zero except the \( z^{-\Delta} \) term

\[
\begin{bmatrix}
H_0^T & 0^T & \cdots \\
H_1^T & H_0^T & \cdots \\
\vdots & H_1^T & \ddots \\
H_{N-1}^T & \cdots & \cdots & H_{N-2}^T \\
0^T & H_{N-1}^T & \cdots & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
0^T & \cdots & \cdots & 0^T \\
\end{bmatrix}
\begin{bmatrix}
F_0 \\
F_1 \\
\vdots \\
F_{N-1} \\
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
\vdots \\
1 \\
\vdots \\
0 \\
\end{bmatrix}
\]  
(6)

where \( \Delta = N - 1 \).
From eqn.(1) Aliasing is eliminated if:

\[
(-1)^n
\begin{bmatrix}
H_0^T & 0^T & \cdots \\
H_1^T & H_0^T & \cdots \\
\vdots & H_1^T & \ddots \\
H_{N-1}^T & \cdots & \cdots & H_{N-2}^T \\
0^T & H_{N-1}^T & \cdots & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
0^T & \cdots & \cdots & 0^T \\
\end{bmatrix}
\begin{bmatrix}
F_0 \\
F_1 \\
\vdots \\
F_{N-1} \\
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
\vdots \\
0 \\
\end{bmatrix}
\]  
(7)

By examining the \( z^{(N-1)} \) term from eqn.(7) the following constraint is imposed on eqn.(6)

\[
\sum_{n=0}^{N-1} H_{N-1-2n}^T F_{2n} = \sum_{n=0}^{N-1} H_{N-1-(2n+1)}^T F_{2n+1} \]  
(8)

Then the PR conditions for the 2-Channel system can be expressed as:

\[
\begin{bmatrix}
H_0^T & 0^T & \cdots \\
H_1^T & 0^T & \cdots \\
\vdots & H_1^T & \cdots \\
H_{N-1}^T & \cdots & \cdots & H_{N-2}^T \\
0^T & H_{N-1}^T & \cdots & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
0^T & \cdots & \cdots & 0^T \\
\end{bmatrix}
\begin{bmatrix}
F_0 \\
F_1 \\
F_2 \\
\vdots \\
F_{N-2} \\
F_{N-1} \\
\end{bmatrix}
= 
\begin{bmatrix}
0^T \\
\vdots \\
J \\
\vdots \\
0^T \\
\end{bmatrix}
\]  
(9)

where \( J \) is the antidiagonal identity matrix.
This system can be expressed in a more compact form in terms of \((2 \times 2)\) submatrices \( P_j \) and \( Q_j (j = 0, \ldots, L - 1)\)

\[
\begin{bmatrix}
P_0^T & 0 & \cdots & O \\
P_1^T & P_0^T & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
P_{L-1}^T & P_{L-2} & \cdots & P_0^T \\
O & \cdots & \cdots & O \\
\vdots & \vdots & \ddots & \vdots \\
O & \cdots & \cdots & O \\
\end{bmatrix}
\begin{bmatrix}
Q_0 \\
Q_1 \\
\vdots \\
Q_{L-1} \\
\end{bmatrix}
= 
\begin{bmatrix}
0^T \\
\vdots \\
J \\
\vdots \\
0^T \\
\end{bmatrix}
\]  
(10)

For a maximally decimated M-channel system the submatrices \( P, Q, O \) and \( J \) are \((M \times M)\) matrices.
The system of equations

\[
As = B
\]  
(11)

represents the necessary and sufficient conditions for PR.
The system is over determined, and is only solvable for arbitrary filters when \( N = 2 \). From eqn.(6),(9) it is observed that the position of \( J \) is directly proportional to the system delay \( \Delta \). This formulation has been independently derived by Nayebi[1] using a complete time-domain analysis of the system.

3 Wavelets and Filter Banks

Wavelets transforms are closely related to tree structured two channel filter banks, and hence to multiresolution analysis [3].
Such a tree structure gives rise to nonuniform filter bandwidths and nonuniform decimation ratios [4].
The properties required for a discrete signal to be fully represented by a series of scaled wavelets is equivalent to the design constraints imposed on the two channel filter bank.
obtain perfect reconstruction.
In this paper we are concerned with designing a new form of wavelets termed low delay wavelets using the time-domain formulation in eqn.(10). It can be verified that the classical wavelets [4] satisfy these conditions with \( \Delta = \frac{N}{2} - 1 \).
Low delay wavelets can be achieved by noting that the position of the submatrix \( J \) within the matrix \( \mathbf{B} \) in eqn.(10) is directly proportional to the system delay [1]. If \( J \) is equispaced at row position \( (N - 1) \) then the delay is \( (N - 1) \), which can lead to orthonormal and bi-orthogonal linear phase wavelets. Solving eqn.(10) when \( J \) is moved away from the centre of \( \mathbf{B} \) leads to the new low delay wavelets.

4 Design Algorithm

In [1] eqn.(10) is solved using an iterative approach in which the matrix \( \mathbf{A} \) is upgraded during each iteration using a constrained conjugate gradient technique. Frequency domain constraints are needed to ensure that the filters satisfy the original design criteria. This is a computationally intensive procedure.
In this paper a simpler recursive procedure is proposed to design FIR multirate filter banks and wavelets, based upon the properties of the linear system in eq.(11).
The system in eqn.(11) is overdetermined for \( N > M \), and usually does not have an exact solution. Thus we strive to minimize

\[
|| \mathbf{A} \mathbf{s} - \mathbf{B} ||_p
\]

for a suitable value of \( p \).
The matrix \( \mathbf{A} \) is a symmetric block Toeplitz matrix with full column rank. Under these conditions there exists a unique Least square (LS) [5] solution to the synthesis matrix \( \mathbf{s} \) by solving the symmetric positive definite linear system

\[
\mathbf{A}^T \mathbf{A} \mathbf{s}_{\text{LS}} = \mathbf{A}^T \mathbf{B}
\]

We will exploit this fact and the fact that for a perfect reconstruction system, the analysis and synthesis filters are interrelated bandpass filters. Therefore we use an iterated algorithm to update the matrix \( \mathbf{A} \) from the coefficients of the matrix \( \mathbf{s} \) using a formula that ensures convergence such that

\[
\epsilon_{\text{LS}} = || \mathbf{A} \mathbf{s} - \mathbf{B} ||_2
\]

is minimized.
The design flowgraph is shown in Figure 2. The quality of the starting filters is crucial to the design, so to ensure that the filters remain good band pass filters we constrain the initial analysis filters to have mirror image symmetry with respect to \( \omega = \pi / 2 \)

\[
|H_k(e^{j\omega})| = |H_{M-k-1}(e^{j\omega+\pi})|
\]

This method allows one to trade off frequency-domain and time-domain properties of the filter bank, thus optimizing the design for a given application.

5 Design Examples

Example 1: A five channel system is designed from the initial analysis filters obtained using the Parks-McClellan program [6] with \( N = 55 \) and \( \Delta = N - 1 \). Figure 3 shows the designed analysis filters responses. The design algorithm was stopped when the reconstruction error reached \( 1e^{-5} \).
Example 2: Low delay wavelets were designed, using initial filters using the same program, with \( N = 32 \) and \( \Delta = 15 \). Figure 4 shows the frequency responses of the wavelets with five resolution levels. Figure 5 shows the impulse responses of the low delay wavelets.
6 Conclusions

This paper has derived the time-domain equations (10, 11) for perfect reconstruction of FIR filter banks from the classical z-domain formulation. A simpler solution methodology for eqn. (10) has been outlined based upon exploiting the inherent dependence between the analysis and synthesis filters. There is a link between multirate systems and the discrete wavelet transform. By shifting the position of the matrix J in eqn. (10), it becomes possible to reduce the inherent delay. When J is away from the centre low delay multirate systems and wavelets result. Low delay wavelets will find applications in speech processing, communications and other time critical applications.

References


A New Method for Efficient Convolution in Frequency Domain by Non-Uniform Partitioning

Gerard P.M. EGGLEMERS and Piet C.W. SOMMEN

Dept. of Electrical Eng., Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, Tel./Fax: +31 40 472366 / +31 40 455674, E-mail: G.P.M.Egglemers@ele.tue.nl

Abstract. A direct implementation of a filter (convolution) of length \( N \) in time domain requires \( N \) multiplications per sample. For large filters (thousands of coefficients) this implies a large computational complexity. By using block-processing techniques and Fast Fourier Transforms (FFT's), filtering operations can be performed in frequency domain. If the processing delay and thus the block length \( B \) are bound to a maximum, computational complexity can be reduced by partitioning of the filter into several smaller filters (of equal length). For small block lengths computational complexity can still be to large for implementation on a Digital Signal Processor (DSP). A solution is to partition the filter into unequall length subfilters, which results in much smaller computational complexity for small block lengths. Filters of 4000 coefficients, the number of real multiplications can be reduced by a factor 16.5 using non-uniform partitioning.

1 Introduction

1.1 Problem description

The implementation of a length \( N \) convolution (a filter) directly in time domain costs \( N \) multiplications per sample. In some applications, such as acoustic echo cancellation, filters are needed that have several thousands of coefficients, so time domain filtering results in a very large computational complexity.

By using block-processing techniques with block length \( B \) and Fast Fourier Transforms (FFT's), filtering is performed in frequency domain \([2, 3]\). In practical situations, like in the earlier mentioned acoustic echo canceller, processing delay and thus block length \( B \) is bound to a maximum that is much smaller than the number of filter coefficients \( N \). The above mentioned method then results in a huge computational complexity. Partitioning of the filter into smaller equal-length subfilters can reduce this computational complexity. For the example of acoustic echo cancellation the Partitioned Block Frequency Domain Adaptive Filter (PBFDADF) \([4, 5, 6]\) can be used.

The computational complexity of the filtering operations in that algorithm still is quite large, and in fact determines the overall complexity \([7]\). Therefore here a new algorithm for frequency domain convolution is introduced, that uses a non-uniform partitioning of the filter into sub-filters.

1.2 Notation

Signals are denoted by lower case characters, and constants by upper case characters. A vector is obtained by underlining a bold-case character, lower case for time domain, and upper case for frequency domain. A matrix is denoted by a caligraphic upper case character (such as \( X \)) or bold face upper case character (such as \( I \)). For matrices and vectors the dimension is put in superscript (e.g. a \( B \times Q \) matrix \( X \) is denoted as \( X^{B \times Q} \), for a square matrix the second dimension is omitted). A subscript \( i \) to a scalar, vector or matrix, such as \( w_i \), denotes the \( i \)th version of that scalar, vector or matrix. The \( k \)th element of a vector is obtained by putting the vector between brackets with as subscript \( k \), for example \((w_i)_k\). A time index \( k \) is denoted by appending \([k]\), \((\cdot)^t\) denotes the transpose and \((\cdot)^\ast\) the complex conjugate. Finally, the symbol \( T \) denotes the intersample distance (1/T is the sample rate).

2 Filtering in time domain

2.1 Uniform partitioning

A filter \( W \) performs a convolution of the signal \( z[k] \) and its coefficients \( w_0 \ldots w_{N-1} \):

\[
\hat{y}[k] = \sum_{i=0}^{N-1} z[k-i] \cdot w_i
\]

\[
= (z^N[k])^t \cdot w^N
\]

(1)

with:

\[
z^N[k] = (z[k-N+1], \ldots, z[k])^t
\]

\[
w^N = (w_{N-1}, \ldots, w_0)^t
\]

(2)

(3)

Computing this convolution directly requires \( N \) multiplications per sample. To reduce computational complexity, the filter can be partitioned into several sub-filters of equal length that are individually transformed to frequency domain. In \([4, 5, 6]\) it is shown that a delay-line in frequency-domain can be used by proper choice of the block-length \( B \) and the sub-filter length (partition factor) \( Q \). Doing this, the convolution is partly performed in time and partly in frequency domain. This partitioning can be described in the following equation (for simplicity it is assumed that \( N/Q \) is integer):

\[
\hat{y}[k] = \sum_{i=0}^{B-1} (z^Q[k - iQ])^t \cdot w_i^Q
\]

(4)
\[ X^Q[k] = (z[k - Q + 1] \cdots z[k])^t \]
\[ W^Q = (w_{iQ + Q-1} \cdots w_{1Q})^t \]

The above equations can be implemented efficiently in frequency domain with the help of block processing and FFTs. The efficiency of block processing depends strongly on the block- and subfilterlength: a larger block length results in a more efficient implementation (with a larger subfilterlength). The processing delay however, depends linearly on the block length. In practice this means that the block length is bounded to a (small) maximum using uniform partitioning. This can lead to an unacceptably large computational complexity.

From eq. (4) it follows that the i'th sub-filter needs vector \( X^Q[k - iQ] \). This vector is iQ sample times earlier available than the result of the convolution is needed. When block computation is used in combination with a non-uniform partitioning, a larger block-length could be used here as computational can be started earlier. A larger block-length in combination with a larger filter length results in a more efficient implementation.

### 2.2 Non-uniform partitioning

Here we introduce a new approach to partition the filter into sub-filters of not necessarily equal length. The goal is to reduce complexity by using a larger sub-filter length (and a larger block-length for block-processing) where that is possible. We propose to partition the filter into G sub-sets of sub-filters. The j'th sub-set consists of \( g_j \) sub-filters, each of length \( Q_j \) (with for \( 1 \leq j < G : Q_j < Q_{j-1} \)). This non-uniform partitioning is depicted in figure 1 and can be described as:

\[ \hat{\epsilon}[k] = \sum_{j=0}^{G-1} \sum_{i=0}^{g_j-1} (X^Q[k - S_{j,i}])^t \cdot W^Q_{j,i} \]
\[ = \sum_{j=0}^{G-1} \hat{\epsilon}_j[k] \]

with:

\[ \hat{\epsilon}_j[k] = \sum_{i=0}^{g_j-1} (X^Q[k - S_{j,i}])^t \cdot W^Q_{j,i} \]
\[ S_{j,i} = \begin{cases} i \cdot Q_j + \sum_{a=0}^{j-1} g_a \cdot Q_a & \text{for } 1 \leq j < G \\ i \cdot Q_0 & \text{if } j = 0 \end{cases} \]
\[ X^Q_{j,i}[k] = (z[k - Q_j + 1] \cdots z[k])^t \]
\[ W^Q_{j,i} = (w_{iQ + Q_j - 1} \cdots w_{1Q})^t \]

### 2.3 Block-based computation

To calculate all \( \hat{\epsilon}_j[k] \) efficiently in frequency domain, eq.(8) is transformed to a block-based equation. It is assumed that the computation of the j'th block equation with block-length \( B_j \) requires \( D_j \cdot T \) seconds (\( T \) is the intersample distance). At time \((kB_j - D_j) \cdot T \) computation of the vector (block) \( \hat{\epsilon}_j^B[kB_j] \) is started, so at time \( kB_j \cdot T \) the computation is ready. Since the maximum allowable delay is defined as \( D_{\text{max}} \cdot T \) seconds, the first element (the "oldest" element) of \( \hat{\epsilon}_j^B[kB_j] \) must have index \( kB_j - D_{\text{max}} \). The following block-based equations can now be constructed for all \( j, 0 \leq j < G \):

\[ \hat{\epsilon}_j^B[kB_j] = \sum_{i=0}^{g_j-1} \left( (X^Q_{j,i}[kB_j - D_{\text{max}}] \cdots (X^Q_{j,i}[kB_j - D_{\text{max}} - S_{j,i}])^t \right) \cdot W^Q_{j,i} \]
\[ = \sum_{i=0}^{g_j-1} (X^Q_{j,i}[kB_j + B_j - D_{\text{max}} - S_{j,i} - 1])^t \cdot W^Q_{j,i} \]

To simplify notation, it is defined that:

\[ \tau_j = B_j + D_{\text{max}} + S_{j,0} + 1 \]

This implies that for all \( i, j \):

\[ X^B_{j,i}[kB_j + B_i - D_{\text{max}} - S_{j,i} - 1] = X^B_{j,i}[kB_j - \tau_j - iQ_j] \]

To obtain a causal realisation of eq.(12), all time indices on the right hand side must be smaller than or equal to \( kB_j - D_j \). Knowing that \( iQ_j > 0 \) (if \( j > 0 \)), the next set of conditions must be true for \( 0 \leq j < G \):

\[ \tau_j \geq D_j \]
\[ B_j \leq D_{\text{max}} + S_{j,0} - D_i + 1 \]
\[ B_j \leq D_{\text{max}} + 1 - D_i + \sum_{a=0}^{j-1} g_a Q_a \]

The block-length \( B_0 \) is bound to \( D_{\text{max}} + D_0 + 1 \), the other block-lengths can be maximized to minimize computational complexity.

### 3 Transformation to frequency domain

All block-equations are transformed to frequency domain with Fast Fourier Transforms (FFT's) using an overlap save method [1, 4, 5]. For every sub-set of filters one FFT and one IFFT (of length \( M_j \) for the j'th sub-set) are needed. For \( 0 \leq j < G \) this is denoted as:

\[ \hat{\epsilon}_j^B[kB_j] = \begin{pmatrix} \hat{\epsilon}_j^B[kB_j] & \cdots & \hat{\epsilon}_j^B[kB_j] \end{pmatrix} (F^{M_j})^{-1} \]

\[ = \begin{pmatrix} \sum_{i=0}^{g_j-1} (X^M_{j,i}[kB_j] \otimes W^M_{j,i}) \end{pmatrix} \]

\[ = F^{M_j} \left( (Y^Q_{j,i} \otimes W^Q_{j,i})^t 0 \ldots 0 \right) \]

\[ X^M_{j,0}[kB_j] = F^{M_j} \cdot X^M_{j}[kB_j - \tau_j] \]
Figure 1: Partitioning of coefficient vector

\[ \mathbf{x}^{M_j}[kB_j] = \mathbf{X}^{M_j}_{j,r}[(k - \frac{Q_j}{B_j})B_j] \] (19)

\( I^Q \) denotes the \( Q \times Q \) identity matrix and \( J^Q \) the \( Q \times Q \) mirror matrix with the \((p,q)\)'th element \((J^Q)_{p,q}\) equal to 1 for \( p + q = Q - 1 \). \( 0 \) is an all zero matrix of appropriate dimension. The \( M_j \times M_j \) Fourier matrix is denoted by \( \mathbf{F}^{M_j} \). The symbol \( \otimes \) is defined as the elementwise product of two vectors (resulting in a vector). If we assume that \( Q_j/B_j = g_j \) is an integer (if \( g_j > 1 \)), then a delay-line can be used for eq. (19). The new algorithm is depicted in figure 2 and 3. The boxes

![Figure 2: Filter part for NU-PBFDAF](image)

\( q_j \Delta' \) represent \( q_j \) one-sample delays (of \( B_j \cdot T \) seconds). The boxes \( r_j \Delta' \) are \( r_j \) one-sample delays (of \( T \) seconds). The blocks \( \text{Subfil}_j \) of figure 2 are depicted in figure 3.

### 4 Complexity

#### 4.1 Fast Fourier Transforms

The \( j \)'th set of sub-filters needs 1 FFT and 1 IFFT per \( B_j \) samples for the filter input and output signals. The number of real multiplications needed for an FFT (and an IFFT) of length \( M_j \) is equal to \( f(M_j) \). The average number of real multiplications per sample \( \Phi(j) \) for the \( j \)'th sub-set then equals:

\[ \Phi(j) = \frac{2}{B_j} \cdot f(M_j) + \begin{cases} \frac{2g_j(M_j-1)}{g_j} & \text{if } M_j > 1 \\ 0 & \text{if } M_j = 1 \end{cases} \] (20)

To get insight into the advantage of using more than one partition factor, an example is introduced here. The number of real multiplications for an FFT (with real-valued input) used here, is given by:

\[ f(M) = \begin{cases} \frac{3}{M} \log_2 M & \text{\( M \) integer and } M \geq 8 \\ \frac{1}{M} \log_2 M & \text{\( M \) integer and } 2 \leq M < 8 \\ \text{undefined} & \text{else} \end{cases} \] (21)

The delay \( D_j \) of the \( j \)'th sub-set depends on the FFT length \( M_j \). As only the first sub-filter of each sub-set is critical in time, \( D_j \) is assumed to be equal to the number of sample intervals required for the computation of two real FFTs and one elementwise vector multiplication. Using a sample frequency of 8000 Hertz, and implementing the FFTs on the TMS320C30 signal processor yields the delays of table 1.
4.2 Non-uniform partitioned filter

Here an example of a non-uniform partitioned filter is given. The number of coefficients for the filter is chosen to be at least 4000. The maximum allowable delay is 0.5 milliseconds at a sample frequency of 8000 Hertz, which equals 4 sample intervals. For an acoustic echo canceller these numbers are quite realistic. In table 2 the average number of real multiplications per sample is given as a function of the number of different sub-filter lengths. For all \( j, 0 \leq j < G \), the parameters \( B_j, Q_j, g_j \) and \( M_j \) are chosen to minimize \( \Phi_G \), with:

\[
\Phi_G = \sum_{j=0}^{G-1} \Phi(j)
\]  
(22)

From table 2 it can be seen that \( \Phi_G \) can be reduced by a factor \( \Phi_1/\Phi_2 = 16.5 \) here. In figure 4 the dependency of \( \Phi_G \) on \( D_{\text{max}} \) and \( N \) is depicted. This figure 4 shows that the introduced non-uniform partitioning concept is useful when the maximum allowable delay (in sample intervals) \( D_{\text{max}} \) is smaller than \( N/8 \).

5 Conclusions

The introduced new method for fast real time convolution in frequency domain by using a non-uniform partitioning reduces the required number of real multiplies compared to a traditional approach using uniform partitioning. For implementation on a DSP this new method thus reduces the number of instructions needed.

References


A SPECTRAL FACTORIZATION APPROACH TO 3-CHANNEL ANALYSIS/SYNTHESIS SYSTEMS

A. AL-ADNANI and R. CHAPMAN

Signal Processing Division, Dept. of Electronic and Electrical Engineering
University of Strathclyde, Scotland, Glasgow G1 1XW, UK.
email: adnan@uk.ac.strathclyde.eee.spd

Abstract A new Spectral Factorization methodology for designing 3-channel based FIR multirate filter banks is proposed. This method extends the well known two-channel based design, in which all the filters are derived from a single prototype filter without resorting to optimization. Therefore the method is computationally very simple and it now becomes possible to design multirate filter banks with powers of 2 or 3 number of channels without utilizing numerical procedures.

1 Introduction

One of the most commonly used multirate filter banks is a cascaded tree structure of two channel prototype systems. This obviously can implement any filter bank with a power of two number of channels. One of the advantages of such cascaded structures is that the prototype filter bank can be determined without the need for numerical optimization [1][2]. Complex optimization procedures are generally required to design analysis/synthesis systems whose orders are not powers of two [3]. This paper proposes a new technique which enables tree structures of three channel systems to be obtained, without resorting to optimization. The proposed method is based upon the application of standard digital filter techniques to design a low-pass filter, from which the required bandpass and high-pass filters are derived subject to symmetry constraints. This method produces filter banks in which the magnitude and phase distortions are completely eliminated, and with aliasing errors comparable to other techniques [4]. The virtue of the method outlined for the first time in this paper is its inherent simplicity, and it now becomes possible to design any analysis/synthesis filter bank with orders of powers of 2 or 3 number of channels without resorting to any complex procedures.

2 3-Channel System

Consider the maximally decimated 3-channel system shown in Figure 1. The overall transfer function can be written as a linear combination of a linear time varying term $A(z)$, and a linear time invariant term $T(z)$ where

$$A(z) = \frac{1}{3} \sum_{k=0}^{2} \sum_{k=0}^{2} H_k(z)F_k(z), \quad W = e^{-j2\pi/3}$$

and

$$T(z) = \frac{1}{3} \sum_{k=0}^{2} G_k(z)$$

(2)

Aliasing is eliminated if $A(z)$ is equal to zero, phase error is eliminated if $T(z)$ has linear phase, and magnitude error is eliminated if $T(z)$ is all-pass. The system has perfect reconstruction (PR) if $T(z)$ is a pure delay such that

$$T(z) = \frac{1}{3} \sum_{k=0}^{2} G_k(z) = z^{-(N-1)}$$

(3)

where $N$ is the analysis and synthesis filter lengths. Eq. (3) is satisfied if the filters $G_k(z)$ are $2N - 1$ zero-phase third-band filters with impulse responses constrained to have every third term equal to zero and with the centre term equal to $1/3$ [5] i.e.

$$g_k(N - 1) = \frac{1}{3}, \quad 0 \leq n \leq 2N - 1$$

(4)

3 Product Filter Design

To design the product filters $G_k(z)$ to satisfy eq.(4) and have good frequency responses we impose a pair-wise frequency mirror image symmetry condition on the product filters $G_k(z)$. The procedure starts by designing the low-pass filter $G_0(z)$ subject to the conditions in (4). The pair-wise mirror image symmetry states that the frequency response of the filters are symmetric about $\pi / 2$, therefore

$$G_k(z) = G_{2-k}(-z), \quad k = 0, 1, 2$$

(5)

and the high-pass filter $G_2(z)$ can be obtained by

$$g_2(n) = (-1)^n g_0(n)$$

(6)
The composite bandpass filter $G_1(z)$ can be obtained from eq.(3) by

$$G_1(z) = z^{-(N-1)} - (G_0(z) + G_2(z)) \quad (7)$$

The product filters are now said to be strictly complementary [3].

4 Spectral Factorization

The second stage of the procedure is to impose constraints on $H_k(z)$ and $F_k(z)$ to ensure aliasing is eliminated. By setting the synthesis filters as time reversed versions of the analysis filters viz

$$F_k(z) = z^{-(N-1)}H_k(z^{-1}) \quad (8)$$

it is seen that eq.(1) can be minimized. Then from eq.(2) the product filters must have the form

$$G_k(z) = z^{-(N-1)}H_k(z^{-1})H_k(z) \quad (9)$$

Substituting in eq.(3) results in

$$\frac{1}{3} \sum_{k=0}^{2} H_k(z^{-1})H_k(z) = 1 \quad (10)$$

It is observed that the analysis filters satisfy the power complementary property

$$\frac{1}{3} \sum_{k=0}^{2} |H_k(e^{j\omega})|^2 = 1 \quad (11)$$

which is not necessarily sufficient for perfect reconstruction for more than two channels [6] [3]. However with a suitable choice of filters aliasing can be minimized to an acceptable level.

Since $H_k(z)$ and $F_k(z)$ must be FIR filters, then we can write

$$G_k(z) = C \prod_{m=1}^{2N-2} (z - z_m)(z^{-1} - z_m) \quad (12)$$

where C is a constant. This means that for every zero of $G_k(z)$ at $re^{j\phi}$ there must be a zero at $(1/r)e^{-j\phi}$, and the zeros on the unit circle must have a multiplicity of two. Then for each zero pair in $G_k(z)$ one of the pair is included in the analysis filter while the other is included in the synthesis filter. The analysis and synthesis filters always have identical magnitude responses, which is exactly the square root of the magnitude responses of $G_k(z)$. Therefore $G_k(z)$ must be designed to be the square of the desired analysis filters.

The number of possible factorizations of $G_k(z)$ grows exponentially with the number of zeros, and if the product filters have low orders then it's difficult to get realistic analysis and synthesis filters. All the possible factors have identical magnitude responses but differing phase responses.

5 Minimum/Maximum phase Design

Because of the large number of possible factorizations, it is preferable to use factorizations where fast algorithms exist: a unique factorization is the minimum/maximum phase design [4]. The simplest way to design a third-band filter without the use of optimization, is to use a window based method to obtain $G_0(z)$.

The impulse response of the low-pass filter $G_0(z)$ has the form

$$g_0(n) = \frac{\sin(\pi n/3)}{\pi n} w(n)$$

where $g_0(n)$ has length $2N - 1$, and $w(n)$ is an appropriate window function.

In order for the product filter to be factorized into FIR filters with identical magnitude response, it has to have a positive spectrum. This is achieved by applying the following transformation on the impulse responses of the filters $G_k(z)$

$$\delta_0(n) = \begin{cases} g_0(n) & n \neq 0 \\ g_0(n) + \delta_0 & n = N - 1 \end{cases}$$

where $\delta_0$ is the maximum stopband attenuation. For a minimum/maximum phase design, the zeros inside the unit circle are allocated to the analysis filters, and the zeros outside the unit circle are allocated to the synthesis filters. The aliasing error is determined primarily by the order of the filters used, and to a smaller extent the windows used in the initial low pass filter design. To ensure that the aliasing error is minimized, we constrain the filters such that only adjacent filters overlap, and that they have opposite phase characteristics. That is we constrain $H_0(z)$ and $H_2(z)$ to be minimum phase and $H_1(z)$ to be maximum phase, which makes the aliasing error comparable with pseudo QMF methods such as [4].

6 Design Example

The illustrative example starts by designing a third-band filter $G_0(z)$ with length $2N - 1$ which is constrained to be a multiple of three.

In this example $N = 15$, and $G_0(z)$ obtained using eqn.(13) with an optimal prolate spheroidal window with transition band equal to $1\pi$. The filters $G_1(z)$ and $G_2(z)$ are designed according to eqn.(6),(7). Figure 2 shows the frequency response of the product filters $G_k(z)$, and Figure 3 shows their impulse responses. Figure 4 shows the zeros of $G_k(z)$ and Figure 5 shows the transformed zeros of $G_0(z)$ and $G_1(z)$, the zeros of $G_2(z)$ are the mirror image of $G_0(z)$. Figure 6 shows the analysis filters. Figure 7 shows a two level 3-channel tree-structure to obtain a 9-channel system, and Figure 8 shows the analysis filters obtained. The amplitude and phase errors are completely eliminated, and the aliasing error is reduced to $2.5\times10^{-4}$. 1035
7 Conclusion

It has been known for some time that multirate filter banks where the number of channels is a power of 2 can be designed utilizing standard digital filter design techniques. This paper has illustrated how a similar philosophy can be applied to 3-channel multirate systems. Standard digital filter techniques are used to obtain the prototype product filter. Symmetry conditions, the shifting of zeros from the unit circle and spectral factorization are then used to derive the required analysis and synthesis filters. This technique is computationally simple, and produces aliasing errors comparable to other techniques.

Figure 1: Maximally decimated 3-channel system

Figure 2: The product filters

Figure 3: Impulse responses of the product filters

Figure 4: Product filters zeros

Figure 5: Transformed zeros
References


OPTIMAL DYADIC FILTER BANKS FOR SUBBAND CODING AND ZONAL SAMPLING

Alfred Mertins

Hamburg University of Technology, Telecommunications Dept.
Eissendorfer Str. 40, 21071 Hamburg, Germany
e-mail: Mertins@tu-harburg.de

Abstract. The design of optimal prototype filters for multi-resolution signal decomposition is considered. Filters will be optimized with respect to the following quality criteria: a) maximum coding gain of subband coding over PCM; b) minimum approximation error for zonal sampling. Only in the special case of paraunitary systems both criteria are nearly equivalent. As an example for more general systems, the design of linear phase filters (including linear phase filters with unequal even or odd lengths) will be considered (parametrizations will be briefly shown).

1 Introduction

Recently, multi-resolution signal decomposition has been successfully applied to image and audio coding. In statistical optimization of dyadic tree-structured systems, several researchers have focused on paraunitary filter banks [1], [2], [3], [5], [6], [8]. The coding gain of general perfect reconstruction systems has been derived in [4].

Often it is interesting to reconstruct a signal only from a few subbands (zonal sampling). For the two-band paraunitary case it is known that maximum coding gain and minimum approximation error are equivalent criteria [8]. This paper shows that both criteria are completely different in general (e.g. linear phase filters).

2 PR Filter Bank

Perfect reconstruction (PR) systems are considered throughout this paper. Therefore filters must satisfy the condition (9)

\[ H_0(z)G_0(z) + H_1(z)G_1(z) = 2, \tag{1} \]

where \( H_0(z) \) and \( H_1(z) \) are the analysis filters and \( G_0(z) \) and \( G_1(z) \) are the synthesis filters.

It can easily be verified that the PR property is guaranteed if the Nyquist-condition

\[ H_1(z)G_1(z) + H_1(-z)G_1(-z) = 2 \tag{2} \]

is satisfied and the filters \( H_0(z) \) and \( G_0(z) \) are chosen as

\[ H_0(z) = z^{2 \ell - 1}G_1(-z), \]
\[ G_0(z) = -z^{-(2\ell - 1)}H_1(-z) \quad \ell \in \mathbb{Z}. \tag{3} \]

Moreover, in the paraunitary case we have \( H_0(z) = \bar{G}_0(z) \), where \( \bar{\cdot} \) denotes complex conjugation on the unit circle: \( G_0(e^{j\Omega}) = (G_0(e^{j\Omega}))^* \).

Usually, the subband signals are computed successively from a given input sequence \( x(n) = \bar{x}(n) \) as shown in Figure 1a. However, for the derivation of optimality criteria it is useful to rewrite the subband signals as (see Figure 1b)

\[ c_k(m) = \sum_{\ell} x(2^k m - \ell) a_k(\ell), \]
\[ d_k(m) = \sum_{\ell} x(2^k m - \ell) b_k(\ell), \tag{4} \]

where the filters are given by

\[ A_1(z) = H_0(z), \quad A_k(z) = \prod_{i=1}^{k-1} H_0(z^{2^i}), \]
\[ B_1(z) = H_1(z), \quad B_k(z) = H_1(z^{2^{k-1}}) \prod_{i=0}^{k-2} H_0(z^{2^i}), \]
\[ P_1(z) = G_0(z), \quad P_k(z) = \prod_{i=0}^{k-1} G_0(z^{2^i}), \]
\[ Q_1(z) = G_1(z), \quad Q_k(z) = G_1(z^{2^{k-1}}) \prod_{i=0}^{k-2} G_0(z^{2^i}). \tag{5} \]

3 Coding Gain

In order to calculate the coding gain of subband coding over PCM for an \( M \)-band decomposition, the formula of Katko and Yosada [4] can be used:

\[ G_{SBC} = \frac{1}{\prod_{k=1}^{M} (N_k \alpha_k)^{\beta_k}} \tag{6} \]

where

\[ \alpha_k = \begin{cases} \frac{1}{N_k} \|b_k\|^2, & k = 1, \ldots, M-1 \\ \frac{1}{N_k} \|b_k\|^2, & k = M \end{cases} \]
\[ \beta_k = \begin{cases} \frac{1}{N_k} \|b_k\|^2, & k = 1, \ldots, M-1 \\ \frac{1}{N_k} \|b_k\|^2, & k = M \end{cases} \tag{7} \]
The asterisk * denotes the hermitian and $||\cdot||$ denotes the norm of a vector. The vectors $\mathbf{a}_k, \mathbf{b}_k, \mathbf{p}_k$ and $\mathbf{q}_k$ contain the impulse responses $a_k(n), b_k(n), p_k(n)$ and $q_k(n)$, respectively. $R_{xx}(k)$ is the $L_k \times L_k$ autocorrelation matrix of the stationary input process $x(n)$, where $L_k$ is the length of the filters $a_k(n)$ and $b_k(n)$, respectively. The entries are $[R_{xx}(k)]_{i,j} = r_{xx}(j-i) = E\{x(n+j-i)x^*(n)\}$.

4 Zonal Sampling

In most cases where zonal sampling is applied, the high frequency bands will be dropped. If we assume that $M-1$ high frequency bands are not used for synthesis, the quality criterion is

$$\sigma^2 = E\{|x(n) - \hat{x}(n)|^2\},$$  \hspace{1cm} (9)

where $\hat{x}(n)$ has to be computed as shown in Figure 2.

In the paraunitary case, the signal space is decomposed into an orthogonal sum of subspaces, and the Parseval identity indicates that (9) is equivalent to maximizing the energy $\sigma^2 = E\{|c(n)|^2\}$. In general, the signal space will only be decomposed into a direct sum of subspaces so that the Parseval identity cannot be applied. In order to express the approximation error in terms of $A_k(z)$ and $P_k(z)$, the polyphase decomposition will be used. This method allows the decomposition of the cyclostationary process $\hat{x}(n)$ into $2^k$ stationary processes at a sampling rate reduced by the factor $2^k$.

First, the input and output processes will be decomposed into $2^k$ polyphase components forming the stationary vector processes

$$x_p(m) = [x(2^km), x(2^km+1), \ldots, x(2^km+2^k-1)]^T,$$

$$\hat{x}_p(m) = [\hat{x}(2^km), \hat{x}(2^km+1), \ldots, \hat{x}(2^km+2^k-1)]^T.$$  \hspace{1cm} (10)

Now the approximation error can be written as

$$\sigma^2 = 2^{-k}\text{trace} \{R_{xx}(0) - R_{xx}(0) - R_{xx}(0) + R_{xx}(0)\},$$  \hspace{1cm} (11)

where the correlation matrices are defined as

$$R_{xx}(\ell) = E\{x_p(m + \ell)x_p^*(m)\},$$

$$R_{xx}(\ell) = E\{\hat{x}_p(m + \ell)x_p^*(m)\} = R_{xx}(\ell),$$

$$R_{xx}(\ell) = E\{\hat{x}_p(m + \ell)\hat{x}_p^*(m)\}.$$  \hspace{1cm} (12)
The correlation matrices can be computed by applying the input-output relation

\[ \hat{x}_p(m) = \beta_k(m) \ast \alpha_k^*(m) \ast x_p(m), \tag{13} \]

where the vectors \( \alpha_k(m) \) and \( \beta_k(m) \) contain polyphase components of the impulse responses \( a_k(n) \) and \( p_k(n) \), respectively. The vectors can be viewed as vector impulse responses that have to be convolved. For causal synthesis and anti-causal analysis filters the definitions are

\[ \alpha_k^*(n) = [a_k(2^k n), a_k(2^k n-1), \ldots, a_k(2^k n-2^k + 1)], \]
\[ \beta_k^*(n) = [p_k(2^k n), p_k(2^k n+1), \ldots, p_k(2^k n+2^k - 1)]. \tag{14} \]

From (12) and (13), we then have

\[ \mathcal{R}_{x\hat{x}}(m) = \beta_k(m) \ast \alpha_k^*(m) \ast \mathcal{R}_{xx}(m), \]
\[ \mathcal{R}_{\hat{x}\hat{x}}(m) = \beta_k(m) \ast \alpha_k^*(m) \ast \mathcal{R}_{xx}(m) \ast \alpha_k(m) \ast \beta_k^*(m). \tag{15} \]

5 Parametrizations

Parametrizations allowing unconstrained optimization and leading to low-pass filters with a zero at \( z = -1 \) are used. This condition is necessary if zero mean high-pass filters and a wavelet interpretation are desired. The parametrizations are:

(i) Zous and Tewfiks parametrization for the paraunitary case [7], [9];
(ii) The linear phase lattice [1], which allows the design of linear phase filters having equal even lengths;
(iii) Filter design by solving a linear set of equations. This method allows the design of general perfect reconstruction filter banks including linear phase filters with unequal even or odd lengths.

Given an impulse response \( h_1(n) \), the corresponding impulse response \( g_1(n) \) may be computed from the linear set of equations

\[ \delta_0 = \sum_n g_1(n) h_1(2\ell - n), \tag{16} \]

which is the Nyquist-condition (2) in time-domain. Moreover, if zero mean high-pass filters are desired \( h_1(n) \) has to be chosen to satisfy \( \sum_n h_1(n) = 0 \), and the condition \( \sum_n g_1(n) = 0 \) has to be included in the system of equations (16).

6 Results

The coding gain (6) has been used as an objective function for filter optimization. Some results for AR(1) sources with \( \rho = 0.95 \), \( r_{xx}(\ell) = \rho^{|\ell|} \) are shown in Figure 3. Interestingly, linear phase filters with odd and unequal lengths lead to the best results (3 x 5 means: 3 coefficients for the analysis high-pass and 5 coefficients for the analysis low-pass filter). A selection of filter coefficients is shown in Table 1.

Alternatively, the approximation error (11) has been used as an objective function for filter optimization. Some results for AR(1) sources with \( \rho = 0.95 \) and \( \sigma^2 = 1 \) are shown in Table 2 (\( L \) is the filter length). Here the paraunitary filters lead to superior results.

The difference between both criteria can be seen by comparing the frequency responses of the filters shown in Figure 4.

7 Conclusion

The design of optimal prototype filters for multiresolution signal decomposition has been considered. For given input statistics, the filters have been optimized for maximum coding gain and minimum approximation error, respectively. The design methods presented in this paper are not restricted to paraunitary systems. The design of linear phase filters with unequal even or odd
Table 1: Coefficients of Filters with Maximum Coding Gain (M = 5)

<table>
<thead>
<tr>
<th></th>
<th>6 Coeff.</th>
<th>8 Coeff.</th>
<th>5 x 7 Coeff.</th>
<th>7 x 9 Coeff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>h₀(n)</td>
<td>h₀(n)</td>
<td>h₁(n)</td>
<td>h₀(n)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parannary</td>
<td>0.38514928</td>
<td>0.06111926</td>
<td>-0.01878374</td>
<td>-0.00620359</td>
</tr>
<tr>
<td></td>
<td>-0.79642380</td>
<td>-0.01716949</td>
<td>-0.39548565</td>
<td>-0.13061453</td>
</tr>
<tr>
<td></td>
<td>0.42843242</td>
<td>-0.08926140</td>
<td>0.82853879</td>
<td>0.32227068</td>
</tr>
<tr>
<td></td>
<td>-0.14607784</td>
<td>0.33116666</td>
<td>-0.39548565</td>
<td>0.89336344</td>
</tr>
<tr>
<td></td>
<td>-0.10647492</td>
<td>0.77027527</td>
<td>-0.01878374</td>
<td>0.32227068</td>
</tr>
<tr>
<td></td>
<td>0.05149111</td>
<td>0.01788900</td>
<td>-0.01306145</td>
<td>0.04079525</td>
</tr>
<tr>
<td></td>
<td>-0.03504678</td>
<td>-0.00620359</td>
<td>-0.00706792</td>
<td>0.004036128</td>
</tr>
<tr>
<td></td>
<td>-0.12477939</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain = 9.550 dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain = 9.603 dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain = 9.621 dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain = 9.788 dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Approximation error σ₀²

<table>
<thead>
<tr>
<th></th>
<th>6 Coeff.</th>
<th>8 Coeff.</th>
<th>5 x 7 Coeff.</th>
<th>7 x 9 Coeff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>h₀(n)</td>
<td>h₀(n)</td>
<td>h₁(n)</td>
<td>h₀(n)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parannary</td>
<td>0.0176</td>
<td>0.0171</td>
<td>0.0182</td>
<td>0.0173</td>
</tr>
<tr>
<td></td>
<td>0.0184</td>
<td>0.0184</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0442</td>
<td>0.0440</td>
<td>0.0432</td>
<td>0.0416</td>
</tr>
<tr>
<td></td>
<td>0.0437</td>
<td>0.0437</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0871</td>
<td>0.0849</td>
<td>0.0886</td>
<td>0.0882</td>
</tr>
<tr>
<td></td>
<td>0.0896</td>
<td>0.0896</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1715</td>
<td>0.1678</td>
<td>0.1745</td>
<td>0.1699</td>
</tr>
<tr>
<td></td>
<td>0.1769</td>
<td>0.1762</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: Frequency responses of 6-tap linear phase filters (AR(1), ρ = 0.95)

REFERENCES


Novel Complex FIR kth Degree Differentiator of Variable Delay

Ewa Hermanowicz

Faculty of Electronics, Technical University of Gdańsk, ul. Narutowicza 11/12, 80-952 Gdańsk, Poland, Phone +48 58 472578, Fax +48 58 416132

Abstract. An efficient optimal (in a maximally flat frequency response error sense) complex FIR kth degree differentiator of variable fractional sample delay, suitable for operation around any centre frequency, is proposed. Recursive formulas for exact values of the coefficients needed in the design are derived. The proposed differentiator is capable of calculating the kth derivative and fractionally delaying (thus interpolating) an input uniformly sampled discrete-time signal simultaneously in one stage of processing.

1. Introduction

Differentiators are useful in the processing of signals in various fields, such as digital control and robotics, digital image processing, communications, radar, sonar, biomedical applications, etc. In a number of signal processing systems, especially where automatic detection and tracking is involved, the precise values of the first and higher degree derivative of an available discrete-time uniformly sampled waveform, performed at optional instants of time not necessarily coincident with the original sampling instants, thus provided with a noninteger delay τ with respect to the sampling instants of the input data, are required.

This paper proposes an efficient design of a one-stage complex FIR digital differentiator of variable delay τ (DD(k)V/D) capable of calculating the kth degree derivative and delaying the input waveform simultaneously with very low approximation errors in a frequency band around any given centre frequency θ. The proposed design is optimal in the sense of a maximally flat frequency response error (MFE) criterion and exhibits equivalent performance, as measured by the approximation error, with at least two times greater bandwidth and twice smaller FIR filter length, as a two-stage design of a kth degree digital differentiator (DD) [1] cascaded with a complex variable delay [2].

2. Basic Concept

The frequency response of an ideal differentiator of kth degree is given by

\[ d^{(k)}(ω) = (jω)^k, \quad ω \in (-π, π), \quad k = 1, 2, ..., \]  \hspace{1cm} (1)

where \( j = \sqrt{-1} \), and the frequency response of an ideal delay of variable, generally noninteger, time delay τ is

\[ d^0(ω) = \exp(-jωτ), \quad ω \in (-π, π) \] \hspace{1cm} (2)

By multiplication of \( d^{(k)}(ω) \) (1) with \( d^0(ω) \) (2) we obtain the frequency response of an ideal differentiator of kth degree of variable delay (DD(k)V/D), as given by

\[ d^{(k)}(ω) = d^{(k)}(ω)d^0(ω) = (jω)^k \exp(-jωτ) \] \hspace{1cm} (3)

where \( D^{(k)}(ω) = D^{(k)}(ω) \) for \( τ = 0 \).

The basic concept which leads to a novel complex digital DD(k)V/D came from observing that

\[ D^{(k)}(ω) = (-1)^k \frac{d^k[d^{(0)}(ω)]}{dτ^k} \] \hspace{1cm} (4)

Thus, the frequency response of an ideal DD(k)V/D can be obtained by differentiating k times the frequency response of an ideal variable delay:

A FIR approximation, \( D^{(k)}_f(ω) \), of \( D^{(k)}(ω) \), around some angular frequency \( θ \) \( \epsilon \) \( -π, π \), may be written as

\[ D^{(k)}_f(ω) = \sum_{n=0}^{N-1} d^{(k)}(n)\exp(-jωn), \quad ω \epsilon \epsilon (-π + θ, π + θ) \] \hspace{1cm} (5)

where \( d^{(k)}(n), n = 0, 1, ..., N - 1 \) is the impulse response of length N of the kth degree DD of variable delay, i.e. DD(k)V/D. If we designate the impulse response of a FIR approximation of variable delay by \( d^{(0)}_f(n), n = 0, 1, ..., N - 1 \), then the frequency response of this approximation is expressed as

\[ D^{(0)}_f(ω) = \sum_{n=0}^{N-1} d^{(0)}_f(n)\exp(-jωn), \quad ω \epsilon \epsilon (-π + θ, π + θ) \] \hspace{1cm} (6)

By substituting (5) and (6) into (4) we obtain the following relationship between the impulse response, \( d^{(k)}_f(n) \), of DD(k)V/D, and the impulse response, \( d^{(0)}_f(n) \), of the FIR variable delay

\[ d^{(k)}_f(n) = (-1)^k \frac{d^k}[d^{(0)}_f(n)]/dτ^k, \quad n = 0, 1, ..., N - 1 \] \hspace{1cm} (7)

This relationship is used further to obtain recursive formulas for the coefficients, \( d^{(k)}_f(n) \), of FIR DD(k)V/D using the MFE criterion.

3. Recursive Formulas for the Coefficients of DD(k)V/D Based on the MFE Criterion

Maximal flatness of the DD(k)V/D frequency response approximation error
\[ E_{\omega}^{(k)}(\omega) = f_{\omega}^{(k)}(\omega) - D_{\omega}^{(k)}(\omega) \]  

(8)

at given \( \omega = \theta \) requires that

\[ d[\Delta E_{\omega}^{(k)}(\omega)]/d\omega = 0, \quad \omega = \theta, \quad i = 0, 1, \ldots, N - 1 \]  

(9)

Forcing the above conditions of the MFE criterion on the \( DD^{(k)}VD \) frequency response in the case of \( k = 0 \) gives the following solution for the coefficients \( d_{\omega}^{(0)}(n) \) in (6) (cf. [2]):

\[ d_{\omega}^{(0)}(n) = \exp[j\theta(n - \tau)]d_{\omega}^{(0)}(n), \quad n = 0, 1, \ldots, N - 1 \]  

(10)

where

\[ d_{\omega}^{(0)}(n) = \prod_{v=0}^{\Delta N-1}(\tau - v)/(n - v) \]  

(11)

is the impulse response of a FIR variable delay that is maximally flat at \( \omega = \theta \). (Here, we have designated the coefficients of the FIR variable delay by \( d_{\omega}^{(0)}(n) \), instead of \( h(n) \) used in [2], to stress that this delay belongs to the \( DD^{(k)}VD \) class under consideration, as it represents a zeroth degree differentiator of variable delay, and we have used the name: variable delay instead of TFD used in [2] for a tunable FIR delay, to avoid confusion with the same abbreviation, i.e. TFD, now commonly used for time-frequency distribution.)

By substituting (10) into (7) we obtain the following formula for the coefficients of \( DD^{(k)}VD \) whose frequency response error (8) is maximally flat in the sense of MFE criterion (Eq. 9).

\[ d_{\omega}^{(k)}(n) = \exp[j\theta(n - \tau)] \sum_{i=0}^{k}(-1)^{i}(i \partial^{i-k}d_{\omega}^{(0)}(n))/d\tau^{k} \]  

(12)

where \( \Delta = k!/(k - i)! \) and

\[ d_{\omega}^{(0)}(n) = (-1)^{k}d_{\omega}^{(0)}[d_{\omega}^{(0)}(n)]/d\tau^{k} \]  

(13)

The \( i \)th derivative, \( d_{\omega}^{(i)}(n) \), of \( d_{\omega}^{(0)}(n) \), needed for (12), can be readily obtained by using a decomposed instead of factorized form of Eq. 11, as given by

\[ d_{\omega}^{(i)}(n) = \prod_{v=0}^{N-1} \frac{P_{v}(n)}{\Pi(N - v)} \]  

(14)

The coefficients \( P_{v}(n) \) in (14) can be calculated via Horner’s scheme

\[ P_{v}(n) = S_{v}^{(N)} + nP_{v-1}(n) \]  

(15)

where \( S_{v}^{(N)} \) are called Stirling numbers of the first kind, [3].

These can be computed recursively by using

\[ S_{v}^{(N)} = S_{v}^{(N-1)} - (N - 1)S_{v-1}^{(N-1)}, \quad v = 0, 1, \ldots, N - 1 \]  

(16)

Finally, by substituting (14) into (13) we obtain the following expression for the coefficients of \( DD^{(k)}VD \) approximated around \( \omega = \theta \) in the MFE criterion sense

\[ d_{\omega}^{(k)}(n) = (-1)^{k} \prod_{v=0}^{N-1} P_{v}(n) \tau^{k} \]  

while the coefficients of \( DD^{(k)}VD \) approximated around an arbitrary \( \theta \) can be calculated by using (12).

In Appendix we present the listing of a Matlab program that will generate the impulse response \( d_{\omega}^{(k)}(n) \), \( n = 1, 2, \ldots, N, \quad k < N \) of \( DD^{(k)}VD \).

4. Performance

Fig. 1 shows the magnitude response and the net phase response together with the respective approximation error curves, as well as the net group delay response and the frequency response relative approximation error (RE) curve, \( RE(\omega) = 20 \log_{10}[E_{\omega}(\omega)/D_{\omega}(\omega)] \), for the ideal \( DD^{(k)}VD \) and for the proposed \( DD^{(k)}VD \) for \( k = 3, \quad N = 10, \quad 9 = 1 \) rad, \( \tau = \pi + \epsilon = 4.75 \), where \( \tau = (N - 1)/2 = 4.5 \) is the FIR filter transport delay and \( \epsilon = 0.25 \) is the net fractional sample delay between the original sampling instants of \( DD^{(k)}VD \) input and the time instants where the samples of the third derivative of this input signal are calculated. The coefficients \( d_{\omega}^{(3)}(n) \), \( n = 0, 1, \ldots, 9 \) of this complex FIR \( DD^{(3)}VD \) are as follows:

\[ +0.0093 \pm j0.0147 \]
\[ -0.1999 \pm j0.0086 \]
\[ +0.6870 \pm j0.8464 \]
\[ +0.8657 \pm j3.9653 \]
\[ -5.8461 \pm j6.9242 \]
\[ +8.1688 \pm j5.3112 \]
\[ -4.6756 \pm j1.3378 \]
\[ +1.0875 \pm j0.3265 \]
\[ -0.0839 \pm j0.1832 \]
\[ -0.0083 \pm j0.0153 \]

Very low relative errors are available in the frequency band centred around \( \omega = \theta \) for this low order \( DD^{(3)}VD \). For example, for \( N, k, \) and \( s \) as in Fig. 1, a frequency response accuracy better than 99.999% (RE(\omega) < -100dB) is achieved over the band of frequencies of 0.328\pi as compared with the bandwidth 0.1\pi for a respective 3rd order FIR DD of the same length, \( N = 10 \), in [1].

The most difficult point for an approximation is \( \omega = 0 \), where the phase response of the ideal \( DD^{(k)}VD \) exhibits a "jump" of \( \pi \) rad due to the change of sign of the factor \( (j\omega)^{k} \) in (3), responsible for the \( k \)th degree differentiation. This is well observed in Fig. 1 on the net phase curve, where the net phase response is \( \arg[D_{\omega}^{(k)}(\omega) \exp(j\omega \tau)] \) for the FIR \( DD^{(k)}VD \) and is \( \arg[D_{\omega}^{(k)}(\omega) \exp(j\omega \tau)] = \arg[(j\omega)^{k} \exp(-j\omega s)] \) for its ideal counterpart, \( D^{(k)}VD \), respectively. However, in Fig. 1, both the magnitude response and the group delay response are well approximated within the frequency band of interest around given \( \theta \).
5. Conclusions

Efficient complex FIR kth degree digital differentiators of variable delay, suitable for operation around any centre frequency with very low approximation errors in the frequency response have been proposed. Recursive formulas for the calculation of the exact values of the coefficients needed in the design have been derived using a maximally flat frequency response error criterion.

References

Appendix

In Fig. A.1 we have the listing of a Matlab program that will generate the coefficients $d(n) \equiv d(k)(n - 1)$, $n = 1, 2, \ldots, N$, $k < N$ of a maximally flat FIR $DD(k)VD$, according to the recursive procedure as outlined in Section 3, based on the explicit formulas (10), (11) for the coefficients $d^{(0)}(n)$, $n = 0, 1, \ldots, N - 1$ of $DD(0)VD$, i.e. of a maximally flat variable delay. Note however, that the $DD(k)VD$ coefficients can otherwise be derived directly for given $k$ from Eq. 9. This approach leads to the following matrix equation:

$$QR\tilde{g}^{(k)} = \tilde{g}^{(k)}$$  (A.1)

where $Q$ is a Vandermonde matrix independent of $k$, whose entries are:

$$q_{jk} = 1, k = 1, \ldots, N, q_{n dereg}, i = 2, \ldots, N, q_{jk} = (k - j)^{i-1}, i, k = 2, \ldots, N$$  (A.2)

$R_s$ is a diagonal Fourier matrix:

$$R_s = \text{diag}(1, \exp(-j\theta), \ldots, \exp(-j(N - 1)\theta))$$  (A.3)

$d^{(k)}(k)$ is the $DD(k)VD$ coefficients vector:

$$d^{(k)}_{s0} = [d^{(k)}(0), d^{(k)}(1), \ldots, d^{(k)}(N - 1)]^T$$  (A.4)

where $\cdot^T$ stands for transposition and

$$d^{(k)}_{s0} = [d^{(k)}(0), d^{(k)}(1), \ldots, d^{(k)}(N - 1)]^T$$  (A.5)

is a vector of an ideal $D(k)VD$ frequency response (3) derivatives over $\omega$, such that

$$d^{(k)}_{s0}(i) = j^i d^{(k)}(i)(0) / d\omega$$  (A.6)

Solving (A.1) for $\tilde{g}^{(k)}$ requires the inversion of the Vandermonde matrix $Q$. This task is numerically ill-conditioned and a Matlab program using the LU decomposition of $Q$ produces roundoff errors which make the results useless for $N > 12$. The latter can be readily observed for $k = 0$, where the symmetry condition $d^{(0)}(n) = d^{(0)}(N - n)\), $n = 0, 1, \ldots, N - 1$ is violated. Thus it is advisable for large $N$ to compute the $DD(k)VD$ coefficients via the recursive procedure as in Fig. A.1.

```matlab
function dxdVdWdVdW(N, k, ep, theta)
% DWDVD.W - Generation of the coefficients of max. flat kth degree differentiator of variable delay
% Syntax: dxdVdW(N, k, ep, theta)
% Arguments:
% N - coefficient vector length
% k - degree of differentiation
% ep - net fractional delay: ep=tau-(N-1)/2
% theta - centre frequency of approximation in rad/sa
% Output:
% dk - vector of length N, with the dk(n) coefficients

tau=(N-1)/2+ep
j=exp(-j*theta)
% Generate Stirling numbers
for m=1:n+1
s(m,1)=1;
end;
for m=2:n
s(m,n-1)=0; ...;end;
for m=n+1
for v=2:n
s(v,n)=(v-1)*s(v-1,n-1)+s(v,n-1);
end
v
end;
end
% Execute Horner's scheme
for m=1:n+1
j=0:n/2;
end;
for m=1:n
for v=2:n
j=0:n/2;
end;
end;
end;
end
% Main recursion to generate the D(k)VD coefficients
for m=1:n+1
j=0:n;
if v<n
(lv=(n-1)*j-v);
end
v
end;
end;
end
% For l=1:1:(N-1)
for m=1:n
sum(1,1)+j=0:
end;
end;
end;
end
% Generate Pascal triangle elements
for m=1:n
sum(1,1)+j=0:
end;
end;
end;
end;
end;
end;
end
% All done.
```

Fig. A.1. Matlab program for the generation of the coefficients $d(n) \equiv d(k)(n - 1)$, $n = 1, 2, \ldots, N$ of FIR of length $N$ complex digital $k$th degree differentiator of variable delay.
A New Method For Noise Cancellation

Eric BATAILLOU and Hervé RIX

Laboratoire I3S, URA 1376 du CNRS, Université de Nice-Sophia Antipolis
Bât. 4, 250 Av. A. Einstein - Sophia Antipolis 06560 VALBONNE FRANCE
Tél: (33)92.94.27.66, Fax: (33)92.94.28.98, E-Mail: bataillou@mimosa.unice.fr

Abstract: The classical noise cancellation methods make use of one or several inputs recording chiefly noise and only one input recording the signal. Here we present a novel noise cancellation method allowing to process several inputs containing the signal. This method executes a minimization under constraint to make a filtering of both noise and signal inputs. We show that the solution can be obtained both globally in the stationary cases and adaptively when signals are non stationary. The performances of our method are then evaluated and compared to the performances of the other methods by simulation.

1. Introduction

Noise cancelling is a special case of optimal filtering which gives good results in many applications. This technique is used when at least one noise input (reference input) is available. This input is filtered and subtracted to the primary input which contains both noise and signal.

The noise cancellation technique has been applied in many fields of signal processing [1]. It is often used in speech processing, for echo cancellation or in general methods of signal processing. We can also notice the Widrow's work [2], [3], using noise cancellation in antenna array processing and biomedical domain.

The classical methods of noise cancellation make use of one or several reference inputs, and only one primary input. In this paper, we present a novel method. This method is proposed for the cases where several primary inputs containing the signal are available. These different primary inputs contain not only the same useful signal delivered by different transmission channels, but also different independent or correlated noise sequences. In this case, we propose to filter both reference and primary inputs using a minimization under constraint.

2. Classical noise cancellation

The classical model of a noise cancellation technique with one noise input is shown in Figure 1.

The captor 1 receives the useful signal s emitted by the source through a transmission channel. Simultaneously, this captor receives a noise b, assumed to be uncorrelated with the signal s. This noise comes from various noise sources and is transmitted by different transmission channels. The captor 1 delivers the primary input. A second captor (captor 2) delivers a noise x assumed to be uncorrelated with signal s. Generally this noise comes from different sources. Applying a noise cancelling process needs a correlation between this noise x and noise b. Captor 2 gives the reference input. Filtering this input makes it possible to obtain an estimation of b.

Here we just recall that the classical minimization of the mean square error $\xi = \mathbf{E}[e(k)^2] = \mathbf{E}[(\hat{s}(k) - s(k))^2]$ between the estimated and the real signal leads to the well known result: $W = R_{xx}^{-1} P_{ax}$ where $R_{xx}$ is the autocorrelation matrix of $X = [x(k-N) \cdots x(k)]^T$ (N is the length of the filter W) and $P_{ax}$ is the intercorrelation vector between d and X.

![Figure 1](image_url)

Many studies have been done in the case of the classical noise cancelling [4], [5]. The performances of these methods have been abundantly studied as well in time domain as in frequency domain. The global approaches in the stationary case as well the adaptive ones in the non stationary case, and methods devoted to the performances evaluation have been described in many works such as [4], [6], [7].

3. Noise cancelling for multi-primary inputs

3.1 Introduction

The precedent works dealt with mutli-reference inputs. Only few works have been published concerning multi-primary inputs. Most of the methods make use of several parallel classical noise cancellers with one primary input and estimated signal $\hat{s}$ is obtained by summing the outputs of all the noise cancellers. Some specific methods have been carried out particularly in speech processing. For example, some of these methods are based on the automatic choice of the best primary input [8].

In this paper, we propose an algorithm which includes two adaptive filters V and W respectively for the primary inputs and the reference one. This algorithm can be
represented by the diagram in Figure 2 in the case of three primary inputs and three reference inputs.

![Diagram](image)

**Figure 2**

Captor \(d_i\) receives signal \(s_i\) emitted by a unique source of signal but transmitted by different transmission channels. This captor also records noise coming from the noise sources. Each captor \(x_j\) receives only noise assumed to be correlated with at least one noise recorded by of the captors \(d\).

The calculus of the parameters of these two filters is done by minimizing the mean power of the output signal.

In a first part, we show that the classical minimization of this mean power is unworkable due to a lack of a priori information. In a second part, we show that it is possible to estimate the filters coefficients using a minimization under constraint. We propose both a global algorithm to estimate \(V\) and \(W\) for the stationary case, and an adaptive one for the non-stationary case.

At last, the performances of our method are presented on simulation signals.

### 3.2 The Model

Figure 3 shows a general model of all the signals recorded by the captors.

![Diagram](image)

**Figure 3**

The useful signal is signal \(s\). On the reference input \(d_i\), we observe signal \(s_i\) which comes from signal \(s\) through transmission channels represented by \(G\). The noises b correlated with source \(r_k\) are also added on the primary inputs. On the reference inputs \(x\), we observe these different noise signals \(r_k\) coming from different noise sources.

Generally, sources \(r_k\) are not the unique noise sources; the useful signal is also embedded in noise signals \(n_i\) and \(n_j\) uncorrelated with the noise sources \(r_k\). These components can not be cancelled by subtraction.

We have to estimate the coefficients of the filters \(V\) and \(W\) which give an estimation of the useful signal \(s\) at the output of the noise canceller. Our investigation is limited here to the case where filter \(V\) is a vector, that is each input \(d_i\) is affected by only one coefficient \(v_i\) of filter \(V\).

### 3.3 Classical minimization

In this case, the filters \(V\) and \(W\) which minimize the mean square error \(\xi = \mathbb{E}[e(k)^2] = \mathbb{E}[(\hat{s}(k) - s(k))^2]\) are given by the following formulas:

\[
\begin{align*}
    V &= [R_{DD} - R_{DX}R_{XX}^{-1}R_{DX}]^{-1}P_{Di} \\
    W &= R_{XX}^{-1}R_{DX}V
\end{align*}
\]

where \(R_{IJ}\) is the cross correlation matrix between the signal vectors \(I\) and the signal vector \(J\), and \(P_{Di}\) represents the cross correlation vector between the signals vector \(I\) and the signal \(J\).

The problem is the knowledge of the cross correlation vector \(P_{Di}\) between the signal \(s\) and the primary input vector \(D\). The estimation of \(P_{Di}\) needs the knowledge of the filter \(G\) that is the transformation of \(s\) introduced by the transmission channels. In practice, this knowledge is not available. So this estimation of the coefficients of the filters \(V\) and \(W\) must be done through another approach.

### 3.4 Minimization under constraint

#### 3.4.1 The global solution

The minimization criterion is not the mean square error between \(s\) and \(\hat{s}\) but the mean power of the output \(\hat{s}\) that is: \(\mathbb{E}[(\hat{s}(k))^2]\).

To avoid the null solution for \(V\) and \(W\), we impose a constraint on the values of the filters coefficients. We have chosen to enforce the sum of the elements of \(V\) to be equal to one. So, \(V\) and \(W\) are solution of the minimization problem:

\[
\begin{align*}
    \min \mathbb{E}[(\hat{s}(k))^2] \\
    \text{under constraint} \quad V^TU = 1
\end{align*}
\]

with \(U = [1, \ldots, 1]^T\).

The mean power of \(\hat{s}\) can be written:

\[
\mathbb{E}[(\hat{s}(k))^2] = \mathbb{E}[(D^TV - X^TW)^2]
\]
\[ V^T R_{DD} V + W^T R_{XX} W - 2V^T R_{DX} W \]

The solution of the constraint minimization can be obtained resolving the Khun and Tucker equations:

\[
\begin{align*}
2R_{DD} V - 2R_{DX} W + \lambda U &= 0 \\
2R_{XX} W - 2R_{XD} V &= 0 \\
V^T U - 1 &= 0
\end{align*}
\]

The solution of this system is given by:

\[
\begin{align*}
\lambda^* &= \frac{2}{U^T K^{-1} U} \\
V^* &= \frac{\lambda^*}{2} K^{-1} U \\
W^* &= R_{XX}^{-1} R_{DX} V^*
\end{align*}
\] (2)

with \( K = [R_{XX} R_{XX}^{-1} R_{DX} - R_{DD}] \); this matrix is supposed to be invertible.

Here we obtain a global solution for the filter \( V \) and \( W \); in the stationary case, this solution can be obtained by estimating the cross correlation matrices over the all range of the signals.

3.4.2 The recursive solution

Resolving system 1 led to a global solution for the filters \( V \) and \( W \). A recursive solution can be obtained when signals are non stationary. The non stationary may be due either to the noise or to the variations of the transmission channels \( H_i \). This implies to update the filters coefficients. Using a "Constraint Least Mean Square" algorithm [9], the weights are given in function of the step \( k \) by:

\[
\begin{align*}
V(k+1) &= V(k) - 2\mu \left(1 - \frac{1}{N_D^2}\right) D(k+1) \hat{e}(k+1) + \frac{\mu}{N_D} \left[1 - U^T V(k)\right] \\
W(k+1) &= W(k) - 2\mu X(k+1) \hat{e}(k+1)
\end{align*}
\]

where \( N_D \) is the number of primary inputs.

4. Simulations

The following simulations have been done in order to evaluate the performances of our method of noise cancelling in case of multi-primary inputs. We also make a comparison with other methods. These methods are:

- The synchronous signal averaging of the primary inputs which will be considered as the reference method.
- The classical noise cancelling using the average of all the primary inputs as the unique primary input (Figure 1).

The simulation is described in Figure 3 with three primary inputs and three reference inputs. The signal to estimate is a simulated evoked potential (EP) embedded either in white noise or in a simulated electroencephalogram (EEG). These EEG signal results from the estimation of a five order autoregressive process on a real signal.

An example of the signals used for simulation is given in Figure 4. The inputs 1 to 3 are the primary inputs and the inputs 4 to 6 are the reference inputs.
In order to check the influence of the independent noise \( (m_t \text{ and } n_t) \) the performances of the three methods have been studied in function of the signal to noise ratio (SNR = \( \frac{\sum_i x^2(k) / \sum_i m^2_k(k)}{} \)) assuming this SNR is the same on each input. The level of the correlated noise coming from the \( \tau_i \) signals is maintained constant. The residual error variances of the output signals for the three methods are plotted on Figure 8 in function of SNR varying from 700 to 0.5 dB. The performances of our method is always the best. We can also remark that the two noise cancelling methods tend to the same performance when the SNR becomes smaller.

5. Conclusion

We have proposed a novel approach of noise cancelling allowing to take into account all the inputs containing the signal to estimate.

We showed that the optimal solution may be obtained by a constraint minimization algorithm either through a global approach or adaptive estimation in the non stationary cases.

The constraint adaptive estimation need some precautions of use which depend on the applications.

Nevertheless, when more than one primary input are available, the presented method gives better results than the classical noise cancellation techniques which transform the problem in a unique primary input problem.

6. References

AN EXAMPLE OF BLIND SOURCE IDENTIFICATION BY LINEAR PREDICTION

Eric MOISAN

Laboratoire CEPAG - ENSIEG
BP 46 - 38402 SAINT MARTIN D'HERES Cedex - FRANCE
Tel/Fax: [33] 76 82 62 54 / [33] 76 82 63 84
E-Mail: moisan@cepag.observ-gy.fr

Abstract. This paper deals with a method of source separation, in the case of a convolutive mixture of two signals. It is close to blind identification, although a small constraint is introduced in the modelization of the mixture, like in the "cocktail party" one. Thus a temporal method of source separation is studied, based on a multichannel linear predictor, which involves the computation of second order moments only. The proposed algorithm can be applied to sources which are white or colored noises, without any additive assumption. A complete simulation example validates the method. Finally, the extension to blind source separation is depicted.

1. Introduction

Blind identification is a timely field of research with a wide range of applications. It has been motivated by practical problems that involve multiple source signals and multiple sensors, and which share a common objective: separating and estimating the source signals without knowing the characteristics of the propagation channel.

Let us consider two unknown time series \( \{ v_n \} \) and \( \{ w_n \} \), which are assumed to be only uncorrelated. After propagation through a transmission channel, two other signals are measured, \( \{ x_n \} \) and \( \{ y_n \} \). The problem is to restore both sources \( \{ v_n \} \) and \( \{ w_n \} \) from their two observed linear combinations \( \{ x_n \} \) and \( \{ y_n \} \). Several methods have been recently proposed to solve this problem [1-5]. Most of them use higher order statistics. In this paper, a temporal method is presented, which only requires the computation of second order moments. Section 2 is devoted to the modelization of the mixture and shows the connection with multichannel linear prediction. The identification of a convolutive mixture of two white noises is investigated in section 3 and then, the method is generalized to colored sources in section 4. Simulation results are presented in section 5 and a remaining difficulty is finally discussed.

2. Modelization and linear prediction

In the following, time series will always be represented by lower case letters and their z-transform will be noted by upper case letters.

\[
X(z) = \sum_{n=-\infty}^{+\infty} x_n z^{-n}
\]  

(1)

2.1 Modelization

Using the z-transform for time series, without any consideration of existence, we assume that the mixture can be described as depicted in Figure 1, where the two unknown filters \( N_V \) and \( N_W \) are transversal of order q and strictly causal, i.e. for \( N_X \)

\[
N_X(z) = \sum_{k=1}^{q} n_X(k) z^{-k}
\]

(2)

This modelization is generally called a "cocktail party". Mathematically, such a mixture can be compacted

\[
\begin{bmatrix}
X \\
Y \\
\end{bmatrix} =
\begin{bmatrix}
1 & N_X \\
N_Y & 1 \\
\end{bmatrix}
\begin{bmatrix}
V \\
W \\
\end{bmatrix} = M_{mixture}
\begin{bmatrix}
V \\
W \\
\end{bmatrix}
\]

(3)

When only \( X \) and \( Y \) are observed, both sources \( V \) and \( W \) can be identified by simply inverting the unknown matrix \( M_{mixture} \).

2.2 Linear prediction

The proposed algorithm is essentially based on the linear prediction of a multichannel stationary time series. This one is obtained by the juxtaposition of the two observations \( X \) and \( Y \), assumed to be linearly independant.

Figure 1. Mixture modelization.

1050
Let p be the order of the predictor. The multichannel prediction error is given by

\[
\begin{bmatrix}
\tilde{X} \\
\tilde{Y}
\end{bmatrix} =
\begin{bmatrix}
A & -B \\
-G & F
\end{bmatrix}
\begin{bmatrix}
X \\
Y
\end{bmatrix}
= M_{\text{predictor}} \begin{bmatrix}
X \\
Y
\end{bmatrix}
\]

(4)

where

\[
A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}
\]

(5a)

\[
F(z) = 1 + \sum_{k=1}^{p} f_k z^{-k}
\]

(5b)

\[
B(z) = -\sum_{k=1}^{p} b_k z^{-k}
\]

(5c)

\[
G(z) = -\sum_{k=1}^{p} g_k z^{-k}
\]

(5d)

The task of the predictor is the determination of the order p and then, the estimation of the p matrices of coefficients which minimize the sum of the mean square errors \(\tilde{x}_n^2\) and \(\tilde{y}_n^2\). Assuming that signals and mixture are stationary, the expectation is evaluated on the finite number of samples which can be measured. In other words, the predictor has to compute simultaneously the four polynomials A, B, F, and G (order and coefficients) which generate two uncorrelated sequences white at the order two, \(\{\tilde{x}_n\}\) and \(\{\tilde{y}_n\}\), given the observations \(\{x_n\}\) and \(\{y_n\}\) (Figure 2).

A lattice algorithm is used to achieve the multichannel linear prediction, involving its good numerical behaviour and especially its order update [6]. Indeed, without any a priori information about the order, we can choose the one which minimize the sum of the mean square errors. Of course, the obtained results depend strongly on the source properties. That is what we are going to study now.

3. White noises mixture

Intuitively, if sources are white noises, the optimal case corresponds to the equality between the prediction error \(\tilde{X}\) and source V (respectively \(\tilde{Y}\) and W), because the mean square of \(\tilde{X}\) can not be less than the power of the source V. This can be achieved if the matrix \(M_{\text{predictor}}\) is a good approximation of the inverse of the mixture matrix \(M_{\text{mixture}}\), i.e.

\[
\begin{bmatrix}
A & -B \\
-G & F
\end{bmatrix} = \text{Det}^{-1} \begin{bmatrix}
1 & 0 \\
-N_X & N_Y
\end{bmatrix}
\]

(6a)

where \(\text{Det} = 1 - N_X \cdot N_Y\)

(6b)

When sources are white noises, if we are interested only in recovering them, the prediction errors themselves are good estimations of both signals V and W. But in some applications we have to estimate the mixture. Thus, it is interesting to look at the roots of the four polynomials A, B, F, and G, but it is generally impossible to factorize them in order to estimate directly the two transfers \(N_X\) and \(N_Y\). However, looking at eq.(6), we note that it is possible to approximate the inverse of the polynomial \(\text{Det}\) by computing the following product:

\[
Q = A \cdot F - B \cdot G = (1 - N_X \cdot N_Y) \cdot \text{Det}^{-2} = \text{Det}^{-1}
\]

(7)

Under these conditions, the identification of the filters \(N_X\) and \(N_Y\) becomes possible: we only have to prefilter both signals X and Y by Q, which is always stable because this is a transversal filter, generating \(X'\) and \(Y'\). Finally it is enough to apply multichannel linear prediction again to the new signals \(X'\) and \(Y'\). The four new polynomials are given by

\[
\begin{bmatrix}
A' & -B' \\
G' & F'
\end{bmatrix} = \frac{1}{Q} \begin{bmatrix}
1 & 0 \\
-N_X & N_Y
\end{bmatrix}
\]

(8)

so, if Q satisfies eq.(7)

\[
\begin{bmatrix}
A' & -B' \\
G' & F'
\end{bmatrix} = \begin{bmatrix}
1 & 0 \\
-N_X & N_Y
\end{bmatrix}
\]

(9)

It is important to note that the result given by eq.(6) can only be an approximation, because A, B, F, and G are polynomials and so they can not equal exactly \(\text{Det}^{-1}\), which is the inverse of an other polynomial. So, it is necessary to use a predictor whose behaviour is as good as possible when its order p increases: the bigger the order, the better the results.

4. Colored noises mixture

Now, let us assume that sources are autoregressive processes, generated by two uncorrelated white noises \(\varepsilon_V\) and \(\varepsilon_W\), filtered by 1/C and 1/H respectively. The identification becomes a little more difficult, because the predictor can not discern V and W, but rather their generators, \(\varepsilon_V\) and \(\varepsilon_W\), which are white noises!

\[
\begin{bmatrix}
V \\
W
\end{bmatrix} =
\begin{bmatrix}
1/C & 0 \\
0 & 1/H
\end{bmatrix}
\begin{bmatrix}
\varepsilon_V \\
\varepsilon_W
\end{bmatrix} = M_{\text{signals}} \begin{bmatrix}
\varepsilon_V \\
\varepsilon_W
\end{bmatrix}
\]

(10)
Thus, the measured signals can be written
\[ \begin{pmatrix} X \\ Y \end{pmatrix} = M_{\text{mixture}} \cdot M_{\text{signals}} \cdot \begin{pmatrix} \varepsilon_V \\ \varepsilon_W \end{pmatrix} \]  \hspace{1cm} (11)

Under these assumptions, the minimum mean square errors that the multichannel linear predictor can achieve are both white noises \( \varepsilon_V \) and \( \varepsilon_W \). This result becomes possible if the computed matrix \( M_{\text{predictor}} \) is close to the product \( M_{\text{mixture}} \cdot M_{\text{signals}} \). This reads
\[ \begin{pmatrix} A - B \\ -G \end{pmatrix} \begin{pmatrix} C \\ H \end{pmatrix} = \frac{1}{1-N_{X,Y}} \begin{pmatrix} C \cdot N_X \\ -H \cdot N_Y \end{pmatrix} \hspace{1cm} (12) \]

Again, it is necessary to prefilter both observations \( X \) and \( Y \) with the transversal filter \( Q = A \cdot F - B \cdot G \) to get four new polynomials, which satisfy
\[ \begin{pmatrix} A' - B' \\ -G' \end{pmatrix} \begin{pmatrix} C \\ H \end{pmatrix} = \frac{1}{Q(1-N_{X,Y})} \begin{pmatrix} C \cdot N_X \\ -H \cdot N_Y \end{pmatrix} \hspace{1cm} (13) \]

where
\[ Q = A \cdot F - B \cdot G = H \cdot C \cdot (1-N_{X,Y})^{-1} \hspace{1cm} (14) \]

therefore
\[ \begin{pmatrix} A' - B' \\ -G' \end{pmatrix} \begin{pmatrix} C \\ H \end{pmatrix} = \frac{1}{H^{-1}} \begin{pmatrix} C \cdot N_X \\ -H \cdot N_Y \end{pmatrix} \hspace{1cm} (15) \]

Thus, both sources \( V \) and \( W \) can be estimated by filtering the prediction errors \( \hat{X}' \) and \( \hat{Y}' \) respectively with \( F' \) and \( A' \).

5. Simulation results
Figures 3 to 9 show a complete simulation example:

The sources are two autoregressive processes defined on 5000 points. They are generated by two independent zero mean white noises respectively filtered by:
\[ C(z) = 1 + 0.2 z^{-1} + 0.7 z^{-2} \hspace{1cm} (16) \]
and
\[ H(z) = 1 + 0.2 z^{-1} + 0.8 z^{-2} \hspace{1cm} (17) \]

Figures 3 and 4 represent only the last 500 points of both sources.

According to Figure 1, the mixture is given by
\[ N_X(z) = z^{-1} + 0.9 z^{-2} \hspace{1cm} (18) \]
and
\[ N_Y(z) = z^{-1} - 0.8 z^{-2} \hspace{1cm} (19) \]

A first application of an adaptive multichannel linear predictor, with a forgetting factor equal to 0.999, leads to an optimal order \( p \) equal to 20 (Figure 5) and, simultaneously, four polynomials \( A, B, F \) and \( G \), whose complex roots are shown by Figures 6 and 7. Poles and zeros seem to be quite near for each pair, except an isolated zero, -0.45 for \( B(z) \) and 0.4 for \( G(z) \), which could approximate the exact values -0.9 and 0.8 respectively.

The two observations are then filtered by \( Q = A \cdot F - B \cdot G \), to generate \( X' \) and \( Y' \) and the predictor is applied to these new signals. This time, the optimal order \( p \) equals 15. For instance, two new prediction errors (\( \hat{X}' \) and \( \hat{Y}' \)) and four other polynomials \( A', B', F' \) and \( G' \) are obtained.

Finally, the sources are estimated as following
\[ \hat{V}(z) = \hat{X}'(z) \cdot F'(z) \hspace{1cm} (20) \]
and
\[ \hat{W}(z) = \hat{Y}'(z) \cdot A'(z) \hspace{1cm} (21) \]

Figure 8 (resp. 9) represents the mean square error between \( V \) and \( \hat{V} \) (resp. \( W \) and \( \hat{W} \)), normalized by the product of standard deviations of the source and its estimate. In this example, where both sources have similar spectrum, the results are very good, but some difficulties remain.
6. Conclusion

Compared to the others, this method of source separation is easier to use, because it requires the computation of second order moments only. However, it relies on assumptions which may be not always verified. The first one is the strict causality constraint, which can be easily circumvented by using simultaneously two multichannel predictors:

\[
\begin{bmatrix}
    z^{-1} \tilde{X}(z) \\
    \tilde{Y}(z)
\end{bmatrix} =
\begin{bmatrix}
    A & -B \\
    -\Phi & \Gamma
\end{bmatrix}
\begin{bmatrix}
    z^{-1} X(z) \\
    Y(z)
\end{bmatrix}
\tag{22}
\]

and

\[
\begin{bmatrix}
    \tilde{X}(z) \\
    z^{-1} \tilde{Y}(z)
\end{bmatrix} =
\begin{bmatrix}
    \Delta & -A \\
    -G & F
\end{bmatrix}
\begin{bmatrix}
    X(z) \\
    z^{-1} Y(z)
\end{bmatrix}
\tag{23}
\]

The second one is more critical. What happens when the two filters \(N_X\) and \(N_Y\) are such that the determinant of the mixture matrix has one or more roots outside the unit circle? This problem is not yet solved...

References


Discrete Multichannel $H_\infty$ Filtering

M.J. Grimble

Industrial Control Centre, University of Strathclyde
GLASGOW G1 1QE, SCOTLAND, UK

Summary

A solution to the standard sub-optimal discrete-time $H_\infty$ filtering problem is presented using a new polynomial systems approach. The polynomial matrix solution is closely related to the $H_\infty$ state-space based Riccati equation results which have recently become so popular. The calculation of the $H_\infty$ filter involves a $J$ spectral factorization and the solution of two coupled diophantine equations. There are several computational advantages over previous polynomial methods for calculating discrete-time $H_\infty$ filters.

1. Introduction

The design of optimal linear filters for systems with uncertain noise or disturbance models has lagged behind the development of the equivalent robust control theory. The Riccati equation approach is now accepted as a standard means of solving the sub-optimal $H_\infty$ optimal control problem [1]. The same state equation based technique may be applied to solving the sub-optimal $H_\infty$ filtering problem. The optimal $H_\infty$ filtering problem was first considered by Grimble [2] using a polynomial systems approach.

The solution of the sub-optimal $H_\infty$ filtering problem now obtained represents a polynomial equivalent of the Doyle state-space results. The filter is assumed here to be in a model based (Kalman filtering) structural from, so that the solution provides the constant gain matrix, rather than the transfer-function of the total filter. There are considerable numerical advantages to the solution of the diophantine equations when the solution involves constant matrices.

2. Discrete-Time System and Signal Source Descriptions

The multichannel discrete filtering problem is posed in terms of the standard system model:

$$x(t+1) = \bar{A}x(t) + \bar{B}_1w(t) + \bar{B}_2u(t)$$  \hspace{1cm} (1)
$$z_{10}(t) = \bar{C}_1x(t) + \bar{D}_{12}u(t)$$  \hspace{1cm} (2)
$$y_0(t) = \bar{C}_2x(t) + \bar{D}_{21}w(t)$$  \hspace{1cm} (3)

where $x(t) \in \mathbb{R}^n$, $z_{10}(t) \in \mathbb{R}^{p_0}$, $w(t) \in \mathbb{R}^q$, $u(t) \in \mathbb{R}^m$ and $y_0(t) \in \mathbb{R}^p$, denote the state, output to be estimated, disturbance and noise, control and observations signals respectively.

Also let

$$\begin{bmatrix} \bar{C}_1 \\ \bar{C}_2 \end{bmatrix} \in \mathbb{R}^{(p_0 + p)xn}$$  \hspace{1cm} (4)

The following assumptions are made:

A1. $\bar{D}_{21}\bar{D}_{21}^T > 0$

A2. $(\bar{C}_1, \bar{A})$ is detectable and $(\bar{A}, \bar{B}_1)$ is stabilizable.

Filter:

$$\begin{align*}
\hat{x}(t+1) &= \bar{A}\hat{x}(t) + K_{f2}(y_0(t) - \bar{C}_2\hat{x}(t)) + \bar{D}_2u(t) \\
\hat{z}_{10}(t) &= \bar{C}_1\hat{x}(t) + \bar{D}_2u(t)
\end{align*}$$

(5)  \hspace{1cm} (6)

where $\hat{x}(t), x(t) \in \mathbb{R}^n$ and $\hat{z}_{10}(t), z_{10}(t) \in \mathbb{R}^{p_0}$, represent the estimated state and estimated output, respectively.

2.1 Standard System Model State-Equations

Define the generalized error (using (2) and (6)) as:

$$z_1(t) = z_{10}(t) - \hat{z}_{10} = \bar{C}_1(x(t) - \hat{x}(t))$$  \hspace{1cm} (7)

The second output from the standard system model:

$$y(t) = y_0(t) - \bar{C}_2\hat{x}(t) = \bar{C}_2(x(t) - \hat{x}(t)) + \bar{D}_{21}w(t)$$  \hspace{1cm} (8)
and the second input corresponds with the filter gain output:

\[ u(t) = -K_{f2}y(t) \]  \hspace{1cm} (9)

The state \( x_\theta(t) = x(t) - \dot{x}(t) \) can describe the dynamics of the process where

\[ x(t+1) - \dot{x}(t+1) = \bar{A}(x(t) - \dot{x}(t)) + \bar{B}_{f1}w(t) - K_{f2}y(t) \]

The standard system model state equation may therefore be written as:

\[ x_\theta(t+1) = \bar{A}x_\theta(t) + \bar{B}_{f1}w(t) + u(t) \] \hspace{1cm} (10)

\[ z_\gamma(t) = \bar{C}_f x_\theta(t) \] \hspace{1cm} (11)

\[ y(t) = \bar{C}_2 x_\theta(t) + \bar{D}_{f1}w(t) \] \hspace{1cm} (12)

2.2 Standard System Model Transfer-Functions

The standard system model transfer-function, shown in Fig. 1, becomes:

\[ G(z^{-1}) = \begin{bmatrix} G_{01}(z^{-1}) & G_{02}(z^{-1}) \\ G_{11}(z^{-1}) & G_{12}(z^{-1}) \\ G_{21}(z^{-1}) & G_{22}(z^{-1}) \end{bmatrix} = \begin{bmatrix} \bar{A} & \bar{B}_1 & 1 \\ \bar{C}_1 & 0 & 0 \\ \bar{C}_2 & 0 & 0 \end{bmatrix} \] \hspace{1cm} (13)

\[ z_\theta(t) = G_{01}(z^{-1})w(t) + G_{02}(z^{-1})u(t) \] \hspace{1cm} (14)

\[ z_\gamma(t) = G_{11}(z^{-1})w(t) + G_{12}(z^{-1})u(t) \] \hspace{1cm} (15)

\[ y(t) = G_{21}(z^{-1})w(t) + G_{22}(z^{-1})u(t) \] \hspace{1cm} (16)

Define \[ \begin{bmatrix} Q & S \end{bmatrix} = \begin{bmatrix} \bar{B}_1 & \bar{B}_2 \\ \bar{D}_{f1} & \bar{D}_{21} \end{bmatrix} \] \hspace{1cm} (17)

and \[ \text{Ref } \Delta \begin{bmatrix} -z_\theta^2 & 0 \\ 0 & R \end{bmatrix} \text{ and } \text{SRef } \Delta \begin{bmatrix} 0 \end{bmatrix} \] \hspace{1cm} (18)

2.3 Linear Fractional Transformation

Substituting for \( u = -K_{f2}y \) (from (9)) into (14) and (15) obtain:

\[ u(t) = -K_{f2}(G_{21}w + G_{22}v) = -(1 + K_{f2}G_{22})^{-1}K_{f2}G_{21}w \] \hspace{1cm} (19)

\[ z_\theta = (G_{01} - G_{02}(1 + K_{f2}G_{22})^{-1}K_{f2}G_{21})w \] \hspace{1cm} (20)

\[ z_\gamma = (G_{11} - G_{12}(1 + K_{f2}G_{22})^{-1}K_{f2}G_{21})w \] \hspace{1cm} (21)

The linear fractional transformation between the exogenous input \( w \) and the generalized error \( z_\gamma \), whose \( H_\infty \) norm is to be minimized, can now be defined as:

\[ T_{z_\gamma w} = G_{11} - G_{12}(1 + K_{f2}G_{22})^{-1}K_{f2}G_{21} \] \hspace{1cm} (22)

The linear fractional transformation \( T_{z_\gamma w} \) is of course to be minimized in an \( H_\infty \) sense. The \( H_\infty \) norm of \( T_{z_\gamma w} \) can be interpreted as the maximal RMS gain, for any frequency and any type of input, of the system. The \( H_\infty \) filtering problem involves the minimization of the criterion.

\[ J_{f\infty} = \| T_{z_\gamma w} \|_{\infty} \]

3. Polynomial System Description for the Sub-Optimal \( H_\infty \) Filtering Problem

The standard system model transfer-function matrices may be written as:

\[ \begin{bmatrix} G_{01}(z^{-1}) & G_{02}(z^{-1}) \\ G_{11}(z^{-1}) & G_{12}(z^{-1}) \\ G_{21}(z^{-1}) & G_{22}(z^{-1}) \end{bmatrix} = \begin{bmatrix} \bar{C}_0 A_0^{-1} B_{f2}^{-1} & \bar{C}_0 A_0^{-1} B_{f2}^{-1} \\ \bar{C}_2 A_0^{-1} B_{f2}^{-1} & \bar{C}_2 A_0^{-1} B_{f2}^{-1} + \bar{D}_{21} & \bar{C}_2 A_0^{-1} B_{f2}^{-1} \end{bmatrix} \]

\hspace{1cm} (23)

Write the partitioned standard system transfers as:

\[ G_1 = \begin{bmatrix} G_{11} \\ G_{21} \end{bmatrix} = \begin{bmatrix} \bar{C}_1 A_0^{-1} B_{f2}^{-1} \\ \bar{C}_2 A_0^{-1} B_{f2}^{-1} + \bar{D}_{21} \end{bmatrix} \epsilon R([\bar{P}_0 + P_1]x_\gamma(z^{-1})) \] \hspace{1cm} (24)

\[ G_2 = \begin{bmatrix} G_{12} \\ G_{22} \end{bmatrix} = z^{-1} \begin{bmatrix} \bar{C}_1 \\ \bar{C}_2 \end{bmatrix} A_0^{-1} \epsilon R([\bar{P}_0 + P_1]x_\gamma(z^{-1})) \] \hspace{1cm} (25)

Hence \( G_2 \) can be written as:

\[ G_2 = C_0 A_0^{-1} \epsilon R([\bar{P}_0 + P_1]x_\gamma(z^{-1})) \] \hspace{1cm} (26)

where \( C_0 = \begin{bmatrix} \bar{C}_1 \\ \bar{C}_2 \end{bmatrix} \) \hspace{1cm} (27)

and the matrix \( G_2 \) can be written in left-coprime form as:

\[ G_2 = C_0 A_0^{-1} = A_2^{-1} C_2 \epsilon R([\bar{P}_0 + P_1]x_\gamma(z^{-1})) \] \hspace{1cm} (28)

Also note that \( G_1 \) can be written as:

\[ G_1 = A_2^{-1} C_2 \begin{bmatrix} 0 \\ \bar{D}_{21} \end{bmatrix} \epsilon R([\bar{P}_0 + P_1]x_\gamma(z^{-1})) \] \hspace{1cm} (29)

The filter \( J \) spectral-factor \( \gamma(z^{-1}) \) which enters the optimization procedure may be defined to satisfy:

\[ Y_f J_f^2 = G_{11} G_{11} + \Gamma. \] \hspace{1cm} (30)
\[ \gamma = \begin{bmatrix} -g^2 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad J = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \quad (31) \]

The spectral factor \( Y_f \) may be written as \( Y_f = A_2^{-1}D_f \), where \( D_f \) is a polynomial matrix spectral factor which satisfies:

\[ D_f \tilde{D}_f = C_2Q_2^* + A_2R_fA_2^* + C_2^*A_2^* + A_2^*C_2^* \quad (32) \]

The optimization proof using the polynomial systems approach requires the introduction of two diophantine equations. The solution \((G_o, H_o, F_o)\) with \( F_o \) of the smallest degree, is obtained from the following equations:

\[ G_oD_f^s = A_0F_o = (QG_0^s + S_0^*A_0^*)z^{-g} \quad (33) \]

\[ H_oD_f^s = C_oF_o = (R_fA_2^* + S_0^*C_2^*)z^{-g} \quad (34) \]

where \( g \) is the smallest positive integer which ensures these equations are polynomials in the indeterminate \( z^{-1} \). For later use, note that \( \deg(F_o) \leq \deg(D_f) - 1 \) and \( \deg(A_0) = 1 \) and hence \( \deg(G_o) = 0 \) and \( \deg(H_o) = 0 \).

A third equation can now be introduced which is needed for proving the stability of terms within the cost-function. Recall that \( C_0A_0^{-1} = A_2^{-1}C_2 \) then multiplying the first equation by \( A_2^{-1}C_2 \) and adding to the second gives:

\[ (A_2^{-1}C_0G_o + H_o)D_f^s = A_2^{-1}(C_2Q_2^* + C_2^*A_2^* + A_2^*C_2^* + A_2^*A_2^*)z^{-g} \quad (35) \]

Substituting from (32) and right multiplying by \( D_f^s \) gives:

\[ C_2G_o + A_2H_o = D_f^s \quad (36) \]

4. **Polynomial Matrix Solution of the Sub-Optimal \( H_\infty \) Filtering Problem**

The game problem which provides the solution of the \( H_\infty \) filtering problem is considered in this section. To determine the cost-function to be optimized in the game problem a duality is established with the results for the \( H_\infty \) state-feedback control problem.

The fictitious feedback loop which must be introduced in the game problem is illustrated in Fig. 1. Redefine the signal \( (u(t)) \) to include the fictitious feedback \( K_f r_1 \) so that,

\[ u = -K_f r_1 - K_f r_2 = -K_f \begin{bmatrix} z_1 \\ y \end{bmatrix} \]

where \( K_f = [K_{f1} \ K_{f2}] \)

Substituting into equations (15) and (16) obtain:

\[ \begin{bmatrix} z_1 \\ y \end{bmatrix} = \begin{bmatrix} G_{11} \\ G_{12} \end{bmatrix}w + \begin{bmatrix} G_{21} \\ G_{22} \end{bmatrix}u = G_1w + G_2u = (I + G_2 K_f)^{-1}G_1w \quad (37) \]

Let the *filter sensitivity function* (with fictitious feedback):

\[ M_2 \triangleq (I + G_2 K_f)^{-1} \quad (38) \]

then \( u = -M_2 G_1 w \) and hence,

\[ \begin{bmatrix} z_0 \\ z_1 \\ y \end{bmatrix} = \begin{bmatrix} G_{01} - G_{02} M_2 G_1 \\ G_{11} - G_{12} M_2 G_1 \\ G_{21} - G_{22} M_2 G_1 \end{bmatrix}w \quad (39) \]

In terms of the stochastic description the signal \( (v(t)) \) can be considered to represent white noise of zero-mean and unity-variance. The contribution to the *generalized error output* \( z_0(t) \), due to \( (v(t)) \), may easily be found as:

\[ z_2 = -G_{02}(I + K_f G_2)^{-1}K_f v = -(G_0 S_2 K_f) v \quad (40) \]

where \( S_2 \) denotes the *sensitivity function*:

\[ S_2 \triangleq (I + K_f G_2)^{-1} \quad (41) \]

The \( H_\infty \) filtering problem involves the minimization of the \( H_\infty \) norm: \( \|G_{11} - G_{12} M_2 G_1\|_\infty \), or equivalently:

\[ \|G_{11} - G_{12} M_2 G_1(G_{11} - G_{12} M_2 G_1)^*\|_\infty \quad (42) \]

and the \( H_\infty \) control problem involves the minimization of the \( H_\infty \) norm: \( \|G_{11} - G_{12} M_2 G_2\|_\infty \), or equivalently:

\[ \|G_{11} - G_{12} M_2 G_1(G_{11} - G_{12} M_2 G_2)^*\|_\infty \quad (43) \]

Comparison of (42) and (43) motivates the duality relationships [3].

The game problem which provides the solution to the control problem was considered in Grimble (1992), (11) and is defined as follows:

\[ J_{\infty} = \min_{G_{10}} \max_{G_{12}} \sup_{\|v\| = 1} \frac{1}{2} \text{trace} \left( (G_{10} - G_{11} M_2 G_{20}) \right) \]

\[ (G_{10} - G_{20} M_2 G_{11}) \]

1056
\[ -\gamma^2 \text{trace}(C_0 \bar{S}_2 G_0 \bar{G}_2^* G_0^* \bar{S}_2^*) \frac{dz}{z}^{1/2} \]  (44)

The duality relationships may be employed, together with this expression, to obtain the cost optimization problem which determines the $H_{\infty}$ filter:

\[ J_{f_{\infty}} = \left( \min_{K_{f_1} \text{ and } K_{f_2}} \max_{\alpha \in [0,1]} \frac{1}{2} \int_{\alpha^*} \text{trace}(\bar{G}_0^* \bar{M}_2 G_0 \bar{G}_1 \bar{M}_2^* G_0^* \bar{G}_1^*) \right) \]  (45)

\[ -\gamma^2 \text{trace}(K_{f_1}^* G_0^* G_0 G_0^* K_{f_1}^*) \frac{dz}{z}^{1/2} \]  (45)

The optimization must hold for all output maps $\bar{C}_0$.

5. Sub-Optimal $H_{\infty}$ Filtering Problem Results

The main results of the sub-optimal $H_{\infty}$ filtering problem are summarized below.

Theorem 5.1: Polynomial Matrix Solution of the Sub-Optimal $H_{\infty}$ Problem

Consider the model based filter (5), (6), represented in polynomial matrix form:

\[ \dot{x} = (zA_0 + K_{f_2} \bar{C}_0)^{-1}(K_{f_2} \bar{G}_0 + \bar{B}_2 w_0) \]  (46)

\[ \bar{z}_{10} = \bar{C}_2 \bar{x} + \bar{D}_1 \bar{w}_0 \]  (47)

The Kalman gain $K_{f_2}$ solution to the sub-optimal $H_{\infty}$ filtering problem may be found via spectral factorization and diophantine equation calculations, for some

\[ J \text{ Spectral factorization} \]

\[ D_f J D_f^* = C_2 Q C_2^* + A_2 R_f A_2^* + C_2 S_f A_2^* + A_2 S_f^* C_2^* \]  (48)

Diophantine equations:

The solution $(G_0, H_0, F_0)$ with $F_0$ of smallest degree, must be obtained of the coupled diophantine equations:

\[ G_0 D_f z^{-G} + A_0 F_0 = (Q C_2^* + S_f A_2^*) z^{-G} \]  (49)

\[ H_0 D_f z^{-G} - C_0 F_0 = (R_f A_2^* + S_f^* C_2^*) z^{-G} \]  (50)

Partition $G_0 = [G_1 \ C_2]$ where $G_1 \in \mathbb{P}^{n \times p}(s)$ and $G_2 \in \mathbb{P}^{n \times p}(s)$, and $H_0 = H_0^{-1}$ as:

\[ \bar{H}_0 = \begin{bmatrix} \bar{H}_{11} & \bar{H}_{12} \\ \bar{H}_{21} & \bar{H}_{22} \end{bmatrix} e^{R(p_o + p) x(p_o + p)} \]

Filter gain:

\[ K_{f_2} = (G_1 \bar{H}_{11} + G_2 \bar{H}_{21}) \]  (51)

Implied equation:

\[ C_2 \bar{G}_0 + A_2 \bar{F}_0 = D_f \]  (52)

Proof: That the game problem provided the solution of the $H_{\infty}$ filtering problem was established by duality with the $H_{\infty}$ control problem. The solution of the game problem may be established by showing stationarity and that the solution corresponds to a saddle point.

7. Concluding Remarks

The polynomial solution of the sub-optimal $H_{\infty}$ filtering problem provides useful insights into the links between the state-space and the frequency domain results. The solution is more attractive numerically than the previous equalizing polynomial solutions.

References


AN IDENTIFICATION METHOD OF FIR DIGITAL FILTERS IN FREQUENCY-DOMAIN

Ch. Servière, V. Capdevielle

CEPHAG-ENSIEG BP 46 38402 Saint-Martin d'Hères cedex FRANCE

Abstract.
This paper presents a new method of FIR filters identification when the Fourier transformation of the autocorrelation matrix of the filter input cannot be assumed to be a diagonal matrix. This method takes into account some correlation coefficients between close frequency bins which are usually neglected. The original point consists in applying a pre-filtering to data such that the previous matrix may be easily invertible. We propose a pre-processing which leads to a banded matrix, easily invertible by Gauss elimination and LU decomposition with few operations. The performances are equivalent to those obtained with time-domain methods with regard to the residual quadratic error.

1 Introduction
The identification of filters using FIR (Finite Impulse Response) filters has been widely used in many applications, including linear prediction, equalization or noise cancelling [1] [2]. They are generally implemented in time-domain and are based on the minimization of a mean-square error (MSE). This method leads to different solutions such as lattice algorithms, RLS, and also to the suboptimal LMS algorithm. In a similar way, the minimization of a mean-square error in frequency-domain leads to different solutions such as Wiener filtering or LMS [3] [4]. The frequency-domain filter performs generally similarly to a conventional adaptive transversal filter but promises a significant reduction in computation. Adaptive filtering in frequency domain can be accomplished by Fourier transformation of the input signal x(q) and usually independent weighting of the contents of each frequency bin. The two approaches are strictly equivalent if the Fourier transformation of the autocorrelation matrix of x(q),  \( \Gamma_x \), is exactly a diagonal matrix [3] [5] [6]. If the spectral density shape of x(q) has significant variations, this assumption is not verified and some loss in performance from optimum may be obtained versus a time-domain method. We propose a new method which takes into account the right structure of  \( \Gamma_x \), and necessarily some correlation coefficients between close frequency bins, which are usually neglected. The original point consists in applying a pre-filtering to data such that the matrix  \( \Gamma_x \) may be easily invertible. We propose a pre-processing which leads to a banded matrix, easily invertible by Gauss elimination and LU decomposition with few operations. We then verify on simulation results that the performances of the proposed method are equivalent to those obtained with a lattice algorithm.

2 FIR identification in frequency-domain
The FIR filters may be characterized in time-domain by N coefficients of their impulse response w=[w(0), ..., w(N-1)]T or in frequency domain by their frequency weight vector W=[W(O), ..., W(N-1)]T. Values of w are resulting from the minimization of a mean-square error at time j, \( \epsilon(j) \):

\[
\epsilon(j) = \sum_{q=0}^{i} \lambda^{|q|} (d(q) - w^T x(q))^2
\]

where d(q) is the desired response at discrete time instant q, x(q) the filter input and \( \lambda^{|q|} \) the vector \( [\lambda(q), ..., \lambda(q-N+1)] \). \( \lambda \) is an exponential weighting factor with 0<\lambda<1.

The coefficients of the impulse response w of (1) may be given by:

\[
\sum_{q=0}^{i} \lambda^{|q|} x(q)^T x(q) w = \sum_{q=0}^{i} \lambda^{|q|} d(q) x(q)
\]

Figure 1: FIR identification
Define F, the NxN Fourier matrix of element:

\[
F_{jk} = \exp(-j2\pi tk/N).
\]

Multiply by F, (2) may be changed into:

\[
\Gamma W = \sum, \text{ where the } i, \text{th}, \text{kth element of } \Gamma \text{ is :}
\]

Figure 1: FIR identification

Define F, the NxN Fourier matrix of element:

\[
F_{jk} = \exp(-j2\pi tk/N).
\]

Multiply by F, (2) may be changed into:

\[
\Gamma W = \sum, \text{ where the } i, \text{th}, \text{kth element of } \Gamma \text{ is :}
\]
\[ \Gamma_k = \sum_{q=0}^{j-1} X_q(i) X_{q+k} \text{exp}(\text{j}2\pi nk/N) \]

and the kth element of V is:

\[ V_k = \sum_{q=0}^{j-1} X_q(k) \sum_{m=0}^{N-1} D_q(m) \text{exp}(\text{j}2\pi mn/N) \]

X_q(i) and D_q(i) represent the N-point discrete Fourier transforms (DFT) of the sequences \{x(q-N+1), ..., x(q)\} and \{d(q-N+1), ..., d(q)\} at frequency bin i. W(k) represents the DFT of the sequence \{w(0), ..., w(N-1)\} at frequency bin k. \* denotes the complex conjugate.

As \( F \) is an invertible matrix, the solutions \( W \) of (2) and \( W \) of (3), respectively defined in time-domain and in frequency-domain, are equivalent at each sample j, while the two solutions are strictly relied by DFT. The elements of the matrix \( L \) represent some correlation coefficients between different frequency bins. The existence of these correlation coefficients are due to the fact that the DFT is computed on a finite number of samples \( N \).

We remark that the solution \( W \) is generally different from the solution \( H \) that minimizes a least-squares criterion in frequency-domain, for each frequency-bin v:

\[ \varepsilon(v,j) = \sum_{q=0}^{j-1} \lambda^{|q|} |D_q(v) - H(v) X_q(v)|^2 \]

It yields:

\[ H(v) = \frac{\sum_{q=0}^{j-1} \lambda^{|q|} X_q(v) \ast X_q(v)}{\sum_{q=0}^{j-1} \lambda^{|q|} |X_q(v)|^2} \]

We see from (3) and (5) that \( W(v) \) and \( H(v) \) are equivalent only under the condition that the matrix \( \Gamma \) is strictly a diagonal matrix. In practice, \( \Gamma \) is usually approximated as a diagonal matrix, under some conditions on the autocorrelation function of \( x(q) \) detailed in [5] [6]. Let us define the autocorrelation matrix of \( x(q) \), \( R_x \), in a probability context, of element \( r_i = E(x(q)x(q-i)) \). It is proved in [5] [6] that, when \( r_i \) is negligible for \( i > m \) and \( m \) is less than \( N/2 \), the Toeplitz matrix \( R_x \) can be approximated as a circulant matrix \( C_x \) in the sense that the eigenvalue distribution of \( R_x \) is asymptotically equivalent to that of the corresponding circulant matrix \( C_x \). The Fourier transformation of \( C_x \) is then a diagonal matrix. However this assumption is not strictly satisfied [7] and uncorrelation between \( X_q(i) \) and \( X_q(j) \) with \( i \) different from \( j \) is only asymptotically verified (with infinite \( N \)) [8]. In practice, according to the spectral density shape of \( x(q) \), \( \gamma(g) \), \( X_q(i) \) and \( X_q(j) \) may be significantly correlated [3] [9], even for far bins \( i \) and \( j \).

The correlation may be given by [9]:

\[ E(X_q(i) \ast X_q(k)) = \int_0^\infty \gamma(g) F\left(\frac{k}{N} - g\right) F\left(\frac{i}{N} - g\right) dg \]

where \( F\) represents the DFT of the natural window.

\[ F(g) = \sin(\pi g N) \]

Nsin(\pi g)\]

In the case of a white process, \( X_q(i) \) and \( X_q(j) \) are theoretically uncorrelated. In the case of a nonwhite process, the signal defined in the frequency-domain still has some residual correlation, due to the function \( F\). This correlation is more important when the spectral density \( \gamma(g) \) has significant variations. The equation (6) only proves that the correlation between \( X_q(i) \) and \( X_q(k) \) has a local maximum for \( i = k \). We see on simulation results [fig.2] the influence of \( \gamma(g) \) on the matrix \( \Gamma \). \( x(q) \) is here a white process, filtered by a low-pass filter of impulse response \( 0.9^k \). We show on figure 2 the module of the elements of the 15th row of the matrix \( \Gamma \), computed from the equation (6) and in figure 3 the module of the elements of the 24th row of the matrix \( \Gamma \), estimated with 50 independent data blocks of 64 samples. These two figures prove that the hypothesis of diagonal matrix is not all assumed in this case.

\[ \text{Fig. 2} \quad \text{Module of the elements of the 15th row of matrix } \Gamma \]

\[ \text{Fig. 3} \quad \text{Module of the elements of the 24th row of matrix } \Gamma \]

Using the usual assumption of a diagonal matrix \( \Gamma \), some loss in performance from optimum may of course be obtained [3]. We propose in section 3 an original method of filters identification which is implemented in the frequency domain. This method is strictly equivalent to optimum time-domain methods (Fast Kalman, FAEST, FFT or Lattice filters) in the sense that it provides a weight vector \( W \) which is the DFT of the solution \( W \) of (1). Its interest consists in an important decrease of the computational requirements, with regard to previous methods or QR algorithms [10].

3 Presentation of the proposed method:
3.1 Pre-processing of the data
First, we may use specific data windows (Hamming, Hanning, ...) in order to decrease leakage of signal components between bands with regard to the DFT band-pass filters. This is equivalent to the multiplication of system (3) by the product of two matrix (F,A). F is the NxN Fourier matrix of element F_{jk}=\exp(-j2\pi jk/N) and A is a diagonal matrix of element a(i), depending on the chosen window [a(0), a(1), ..., a(N-1)]. The resulting system is then equivalent to (3) if the matrix (F,A) is invertible. We see it is theoretically feasible only if the data window owns non zero values. Consequently, we tried with an Hamming window of expression
\[ f(i) = \alpha + (1-\alpha) \cos(2\pi i/N) \]
with \( \alpha = 0.54 \). Unfortunately, in this case, the module of the elements of the 24th row of matrix \( \Gamma' \) are still significant for bins far from the 24th bin [Fig.4], although this treatment increases the amplitude of the local maximum.

![Fig.4 Module of the elements of the 24th row of matrix \( \Gamma' \) with an Hamming window](image)

The original point of the proposed method consists in applying a pre-processing to data \( x(q) \) and \( d(q) \) such that the matrix \( \Gamma' \) may be easily invertible. Indeed the solution of (1) is unchanged if \( x(q) \) and \( d(q) \) are first filtered by an identical linear filter \( J \). Consider \( N \) linear filters \( J_i \) for \( i = 0, ..., N-1 \) which are band-pass filters of central frequency \( i \). They are chosen in order to bound leakage between bins to a limited band around \( i \). We propose for example filters \( J_i \) with the following impulse response:
\[ J_i(q) = x(q) \exp(2\pi ji/N) \]
with \( x \) close to 1. We show in figure 5 the corresponding complex gain of \( J_i(q) \) for \( x \) equal to 0.98.

![Fig.5 Complex gain of \( J_i(q) \) for \( x \) equal to 0.98](image)

(2) may be transformed into \( N \) equivalent systems in the frequency domain where \( x(q) \) and \( d(q) \) are filtered independently by each filter \( J_i \) \( i = 0, ..., N-1 \). Choose now the \( i \)th equation of each system in order to provide a new linear system:
\[(8) \quad \Gamma'' W = Y \]
(2) is then transformed in an equivalent system (8) where data are filtered by \( N \) linear filters \( J_i \). The \( i \)th, \( k \)th element of \( \Gamma'' \) is equal to:
\[(9) \quad \Gamma''_{ik} = \frac{1}{N} \sum_{q=0}^{N-1} \lambda^{j (q-k)} X_q(i) X_q(k) \exp(-j2\pi qk/N) \]
and the \( k \)th element of \( V'' \) is:
\[(10) \quad V''_k = \frac{1}{N} \sum_{q=0}^{N-1} \lambda^{j q} X_q(k) \sum_{m=0}^{N-1} D_q(m) Y(m,k) \exp(-j2\pi m/N) \]
where \( J_i(k) \) represents the DFT of \( J_i(q) \) at frequency bin \( k \). We know that the vector \( W \) is solution of (8). The structure of the complex gains \( J_i(k) \) is such that the matrix \( \Gamma'' \) may be approximated by a banded matrix. It also assumes that \( \Gamma'' \) is invertible while each equation is necessarily independent of the others. We see on simulation results [Fig.6] the module of the 15th row of matrix \( \Gamma'' \) after the pre-processing of the data. The elements of this row have been estimated with 50 averages on adjacent data blocks of 64 samples. We remark that the leakage of signal components between bands has been widely reduced. The correlation has been concentrated around the 16th element.

![Fig.6 Module of the elements of the 16th row of the matrix \( \Gamma'' \)](image)

The system (8) is easily solved by Gauss elimination and LU decomposition [11] with few operations (L and U have upper and lower bandwidth \( \rho \)).

3.2 Solution of the system
The identification of the filter consists then in solving the linear system of equations (8) \( \Gamma'' W = Y \). Its solution may be broken down in three steps, as detailed in [11].
1. Matrix decomposition \( \Gamma'' = LU \) where \( L \) and \( U \) have upper and lower bandwidth.
2. Solution of the lower triangular system \( L \chi = Y' \)
3. Solution of the upper triangular system \( U \chi = \chi \)

4 Simulation results:
We show the performances of the method on the figure 7. With the same example as previously, we solved the system (8), using a LU decomposition and considering the matrix \( \Gamma'' \) as a banded matrix with different bandwidths \( \rho \).
We then show the quadratic residual error [Fig6] in function of p. The elements of \( \Gamma^t \) and \( \Sigma^t \) have been computed on 50% overlapped data blocks of 64 samples each and 50 data blocks have been used. We remark that \( p \) may be estimated to 5 bins in order to assume the equivalence with the optimum time-domain solution (2) : the case \( p=32 \) consists in solving exactly the system (2). In fact, \( p \) may be overestimated without too much loss in performances. However, we shall see in paragraph 5 that it makes increase the computational requirements. We also remark that the usual frequency-domain algorithms which treat independently each frequency bin (case of \( p=1 \)) provide an important error on this example, when the spectral density of the input filter has significant variations.

![Fig7](image)

**Fig.7**: Residual quadratic error in function of \( p-1 \) for \( x \) equal to 0.98

We see from (9) that the statistical mean of the elements of the matrix \( \Sigma^t \cdot E(\Gamma^t i k) \) is equal to \( E[X_q(i) \cdot X_q(k)] \cdot i k \), where \( i k \) has a decreasing amplitude in function of \( i-k \) and only depends on the choice of the filters \( J_i(q) \). Consequently, the choice of \( p \) may be taken equal to the bandwidth \( B \) of the filters \( J_i \). However, the bandwidth must not be too little in order not to neglect some significant correlation coefficients \( E[X_q(i) \cdot X_q(k)] \) for \( k \) close to \( i \), due to the shape of \( F(g) \)

In this case, we may also write that, in that case, the number of significant coefficients of the impulse response of \( J_i \) will be greater than \( N \). Consequently, the autocorrelation function will be truncated in \( \Gamma^t \) and the quadratic error increases. An optimal choice, for example, seems to be \( x=0.98 \) for \( J_i \), and \( x < p \cdot 10 \) for \( N=64 \). The length of the impulse response of \( J_i \) is then approximated by 50 samples. We verified that the performances of the method are equivalent to those obtained by a time-domain method (lattice algorithm).

### 5 Computational requirements:

The main interest of this method consists in its computational advantage. The number of operations (additions and multiplications) required to produce one output data point is equal to:

\[
1.5 \log_2 N + 2p^2 + 15p - 1.3(p^3/N) - 6.5(p^2/N) - 2.1(p/N)(1/(1-r)),
\]

where \( r \) is the overlap rate.

The "overlap-save" method is included in this expression, in order to obtain a linear convolution after DFT. The computational requirements are broken down in:

1. DFT of the signals : \( 1.5 \log_2 2N \)
2. Iterative computation of the matrix \( \Gamma^t : 6p - 3(p^2/N) \)
3. Iterative computation of the vector \( \Sigma^t : 4p \)
4. Resolution of the system (8) by LU decomposition : \( 5p + 2p^2 - (4p^3/3N) - (7p^2)/(2N) - (13p)/(6N) + 1 \)

In comparison, QR algorithms require \( 6N \) [10], FFT 9N [12] and LMS 2N. The FSU RLS [13] requires \( 8N(N/L) \cdot \log_2 2N + 34(N/L) + 5.5L \). As an example, suppose \( N=64 \) and \( p=5 \). The proposed algorithm requires 168 operations (for \( r=0.25 \)), compared to 384 (QR), 576 (FFT), 128 (LMS) and 384 (FSU RLS) for \( L=16 \). In fact the FSU RLS is very efficient for great \( N \) (for example \( N=8192, L=256 \)). The computational requirements are then equal to 0.6N.

### 6 Conclusion:

This paper presents a new method of FIR filters identification when the Fourier transformation of the autocorrelation matrix of the filter input cannot be assumed to be a diagonal matrix. This method takes into account some correlation coefficients between close frequency bins which are usually neglected. The result point consists in applying a pre-filtering to data such that the previous matrix may be easily invertible. We propose a pre-processing which leads to a banded matrix, easily invertible by Gaussian elimination and LU decomposition with few operations. The performances are equivalent to those obtained with time-domain methods with regard to the residual quadratic error.

### Bibliography:

Discontinuity reconstruction from linear attenuating operators using the weak-string model

Mila Nikolova and Ali Mohammad-Djafari

Laboratoire des Signaux et Systèmes (CNRS-Supelec-UPS)
Supélec, Plateau du Moulon, 91122 Gif-sur-Yvette Cedex, France
Tel/Fax: +33 1 69 41 80 40/33 1 69 41 30 60
E-Mail: nikolova@lss.supelec.fr, djafari@lss.supelec.fr

Abstract. We consider the reconstruction of a piecewise smooth signal, observed through an attenuating operator, such as an incomplete Laplace transform. This inverse problem is very ill-posed, so a regularisation must be applied. The key question is the recovery of the breakpoints between smooth parts. The weak string nicely models this class of signals. But it happens that in presence of attenuation the relevant MAP estimator fails. Physically, the discontinuity recovery is only limited by the noise. We propose a reconstruction technique which works just at this limit and so exploits data very efficiently.

1 Introduction

Attenuation occurs in various applications: it may model the loss inside a medium while a wave propagates, or the signals received from an object moving away, etc... In this paper, we are concerned with models, which may be considered as being composed of a pure attenuation, applied on the signal of interest, and some subsequent non-attenuating, generally non-injective, linear transformation (Section 2). Note that generally it is the contribution of samples to the data which is decaying, rather than the data. The noisy Laplace transform observed on a sparse set is a typical example.

The inverse problem is “doubly” ill-posed: (a) it is underdetermined, and (b) the attenuation, although mathematically invertible, has a too large condition number. So, a regularisation must be applied. The reconstructed signals (chains) are known a priori to be only piecewise smooth, and the regularisation should take this into account. A pertinent model is the weak string, introduced by Blake & Zisserman [1]. It turns out that the relevant MAP solution fails in presence of attenuation (Section 4). We derived an attenuation-adapted reconstruction technique, which in the noiseless case completely compensates for attenuation; in presence of noise, it may happen that the adaptation can only be introduced up to some level. This technique exploits data much better than the MAP estimator (Sections 5, 6).

The solution calculation involves optimising an energy function, which presents many local minima; finding the global one is non-trivial. In a recent paper [5] the authors extended the Blake & Zisserman’s graduated non convexity (GNC) – a very effective in practice, though not theoretically convergent, deterministic global optimisation technique – in order to use it for ill-posed problems. The attenuation requires an adaptation of the GNC parameters as well (Section 7).

Simulation results are presented in Section 8 and concluding remarks in Section 9.

2 Attenuating models

The observation model is given by a linear integral equation whose discrete form relates noisy data $y \in \mathbb{C}^N$ to the original chain $x \in \mathbb{R}^M$ by:

$$y = Ax + n,$$

where $A$ is a known linear attenuating operator, which does not suppress the mean of $x$, and $n$ is the observation noise, assumed additive, white and Gaussian $N(0, \sigma^2 I)$. The observation operator $A$ can be factorised into

$$A = BU,$$

where $B$ is non-attenuating and its columns $b_i$ verify $\|b_i\|_2 = 1 \forall i$, and $U = \text{diag}(u_1, \ldots, u_M)$ is purely attenuating: $u_1 > u_2 > \ldots > u_M$. The contributions to $y$ of the samples $x[i]$ along the chain are highly decreasing.

The likelihood is $\mathcal{L}(x) = \frac{1}{\sigma^2} \|Bu - y\|^2$. Note that the Laplace transform on a regular set can be simplified to a purely attenuating observation: $B$ is the DFT operator, and as $B^{-1} = B^\dagger$, the likelihood is simply $\mathcal{L}(x) = \frac{1}{\sigma^2} \|u - B^\dagger y\|^2$; $\dagger$ means transposed conjugate.

3 The weak-string model

The weak string corresponds to a locally Gaussian, non-stationnary first-order Markov chain with a Boolean line process. It models piecewise continuous signals and its regularisation role is to locally smooth the signal while preserving abrupt transitions.
The relevant MAP energy is:
\[ E(x, l) = \|Ax - y\|^2 + \Phi(x, l), \]
\[ \Phi(x, l) = \sum_{i=1}^{M-1} \lambda^2 [x[i+1] - x[i]]^2 (1 - l[i]) + \alpha l[i], \]
where \( l = \{l[1], \ldots, l[M-1]\} \in \{0, 1\}^{M-1} \) is the line process and \((\alpha, \lambda)\) are positive constants. Let \( l[i] \) be the transition: \( l[i] = 0 \) if \( l[i] < T \), \( l[i] = 1 \) if \( l[i] \geq T \), \( T = \frac{\sqrt{\alpha}}{\lambda} \).

The prior discontinuity detection threshold. The prior energy \( \Phi \) is then a sum of interaction terms \( \Phi(x) = \sum \phi(l[i]) \) with \( \phi(l[i]) = \min\{\lambda |l[i]|^2, \alpha\} \).

The solution \( \hat{x} \), minimising the energy \( E \), shall be referred to as the MAP solution. In presence of attenuation it is quite unsatisfactory and our objective is to find an adaptation of the model parameters \((\alpha_A, \lambda_A)\) as a function of the attenuation, leading to a better reconstruction.

4 Behaviour in attenuation

A magnitude of crucial importance is the threshold amplitude \( H \) of a jump at position \( m \) in the original signal \( x^* \) above which the optimal solution \( \hat{x} \) will have a discontinuity at \( m \). \( H \) is in fact the posterior discontinuity detection threshold. Assuring a good thresholding assures a good estimation of the line process.

This threshold has been studied analytically and numerically in the case \( \lambda = I [1], [3], et al. \). But the role of \( B \) and especially of \( U \) is really determinant.

In order to avoid any possible interferences, in this section we treated the simplest noiseless attenuating model: \( y = UX^* \). For the simulations \( u_m = e^{-0.02m} \).

We established an expression for \( H \) as a function of the solution energies. Even when the solution can be written analytically, the energy expression is too complicated, so we computed thresholds numerically.

In the figures the exact optima are shown, which have been calculated using the Viterbi algorithm [4] for a very finely discretised state space.

The obtained results on the threshold behaviour in presence of attenuation remain true when \( B \) is an ill-conditioned non-attenuating operator.

4.1 The isolated step response

Let the original chain be
\[ x^*(m; h) = hY_m, \quad Y_m[i] = \begin{cases} 0 & \text{if } i < m, \ h \in R, \\ 1 & \text{if } i \geq m, \ h \in R. \end{cases} \]

The optimal MAP solution is either \( \hat{x}_c(m; h) \) - the minimiser of \( \|Ax - y\|^2 + \|\lambda Dx\|^2 \), where \( D \) is the first-order difference operator, or \( \hat{x}_d(m; h) \):

\[
\begin{cases}
\dot{x}_c = h(U + D)Y_m \\
\dot{x}_d = hY_m
\end{cases}
\]

\( \ddot{x}_c \) stands for real part and \( \ddot{x}_d \) for transposed. \( \dot{x}_c \) corresponds to \( \dot{l}_m = 0 \) and \( \dot{x}_d \) to \( \dot{l}_m = 1 \).

\( E(\dot{x}_c) \) is easy to compute and \( E(\dot{x}_d) = \alpha \). Finally, the optimal solution \( \dot{x} \) is determined by:

\[
\begin{cases}
E(\dot{x}_c) < E(\dot{x}_d) & \Rightarrow \ddot{x} = \dot{x}_c \\
E(\dot{x}_c) \geq E(\dot{x}_d) & \Rightarrow \ddot{x} = \dot{x}_d
\end{cases}
\]

\( h \) is the scale of the original chain; it is straightforward that \( x^*(m; h) = hX^*(m; 1) \) leads to \( \dot{x}_c(m; h) = h\dot{x}_c(m; 1) \) and hence \( E(\dot{x}_c(m; h)) = h^2 E(\dot{x}_c(m; 1)) \). The quantity of interest is \( H_0(m) \) - the critical value of \( h \), such that:

\[
\begin{cases}
|l[i] < H_0(m) & \Rightarrow \ddot{x} = \dot{x}_c(m; h) \\
|l[i] \geq H_0(m) & \Rightarrow \ddot{x} = hY_m
\end{cases}
\]

\( H_0(m) \) verifies \( h^2 E(\dot{x}_c(m; 1)) = E(\dot{x}_d(m; h)) \), so

\[
H_0(m) = \frac{\alpha}{\sqrt{E(\dot{x}_d(m; 1))}}.
\]

For a non-attenuating operator, \( H_0(m) \) is independent of the position \( m \) of the jump along the chain (if \( m \) is not close to the ends) [1]. For an attenuating operator, \( E(\dot{x}_c(m; 1)) \) decreases when the step is shifted to the right and the detection threshold \( H_0(m) \) increases jointly with the attenuation (see Fig. 1). That's why the MAP solution fails. In order to set \( H_0(m) \) constant we shall substitute \( \alpha_A \) for \( \alpha \):

\[
\alpha_A(m) = H_0^2 E(\dot{x}_c(m, 1)), \quad H_0 = H_0(1).
\]

The role of this \( \alpha \)-adaptation can be seen in Fig. 4.

![Figure 1](https://example.com/f1.png)

**Figure 1**: The isolated step response in attenuation. Original chain (\( \ddot{x} \)), reconstruction (\( \ddot{x} \)). (a) The attenuation function \( u_m = e^{-0.02m} \). (b) The detection threshold \( H_0(m) \). (c) The step is perfectly reconstructed when \( m \leq 42 \). (d) The same step but shifted cannot be detected any more.

4.2 Interacting discontinuities

Let the original chain be an \( \alpha \)-width gate:

\[
x^*(m; h) = h(Y_m - Y_{m+a-1}), \quad a \in N, \ h \in R.
\]

In a basically similar way we studied the detection threshold \( H_1(m; a) \) in \( x^* \), leading to a discontinuity at \( m \) in the solution \( \dot{x} \). In order to determine the global
minimum, we compared the potential optima: $\tilde{x}_c$ – continuous, $\tilde{x}_c = x^*$ – discontinuous, and $\tilde{x}_c$ – continuous at $m$, continuous at $\{m+a-1\}$. The threshold $H_1(m; a)$ is the critical value of $h$ such that the global minimum in $\tilde{x}_c$ switches to either $\tilde{x}_c$ or $\tilde{x}_c$.

When $A = I$, it is known that this threshold increases when $a$ decreases. On the other hand, for a fixed, this increase is independent of the position $m$ [1]. But in presence of attenuation, for each $a$ fixed, $H_1(m; a)$ increases with the position $m$. And the smaller $a$, the stronger the effect is. It is produced for both models – the non-adapted ($\alpha = \lambda_A, \lambda$) (Fig. 2 (a)) and the $\alpha$-adapted ($\alpha_A, \lambda$) (Fig. 2 (b)), weaker in the second case.

![Figure 2](image)

Figure 2: Intersecting discontinuity detection threshold $H_1(m; a)$ for $a = 1$ ( ), $a = 5$ ( ) and $a = 15$ ( ). (a) Non-adapted model ($\alpha = C^t, \lambda = C^t$). (b) $\alpha$-adapted model ($\alpha_A, \lambda = C^t$).

As a consequence, a gate and a shifted copy of it will be reconstructed differently (Fig. 3). The consequences for a more complicated chain can be seen in Fig. 4 (e).

![Figure 3](image)

Figure 3: Inadequacy of the $\alpha$-adaptation ($\alpha_A, \lambda$). Original chain ( ), reconstruction ( ). (a) The gate signal is well recovered. (b) The same gate but delayed cannot be recovered.

![Figure 4](image)

Figure 4: Discontinuity interaction. (a) Original chain with non-interacting discontinuities. (b) The MAP solution. (c) The $\alpha$-adaptation ($\alpha_A, \lambda = C^t$) is sufficient. (d) Original chain with interacting discontinuities. (e) The $\alpha$-adaptation fails. (f) Reconstruction with ($\alpha_A, \lambda_A$).

Under the hypothesis that $\forall m_i, (u_m - u_{m+1})$ is small enough, this effect can be suppressed by taking:

$$\lambda_A = \lambda_M.$$ (4)

The $\alpha$-adaptation depends on $\lambda$, so we must first compute $\lambda_A$ and then compute the adaptation $\alpha_A$. The result energy function, say $E_A$, reads:

$$E_A(x) = ||Ax - y||^2 + \sum_{i=1}^{M-1} \phi_i(x[i+1] - x[i]),$$ (5)

$$\phi_i(t[i]) = \min\{\lambda_A t[i]r[i], \alpha_A[i]\}.$$ (6)

($\alpha_A, \lambda_A$) defines a non-homogeneous Markov chain, which in the noise/less case positions the attenuated problem solution on the solution of the equivalent non-attenuated problem (see Fig. 4).

5 The noise limit

This energy adaptation, aimed to improve recovery of discontinuities, reduces the regularisation parameter values jointly with the rise in attenuation: in presence of noise it is limited then by the stability requirement.

We shall apply the adaptation only until a level deduced from the false alarm probability $P_{FA}$. As the noise $n$ is $\mathcal{N}(0, \sigma^2 I)$, the requirement: $P_{FA} \leq \varepsilon$ is equivalent to $P(m) \leq S$, where $S$ may be obtained numerically. The latter leads to $H_1(m; 1) \leq S/u_m, m = 1, \ldots, M-1$. $H_1(m; 1)$ has been set constant and $S/u_m$ is increasing, so the adaptation can only be accomplished as long as the inequality holds: $\{\alpha_A[1], \ldots, \alpha_A[K], \alpha_A[K], \ldots\}$ and $\{\lambda_A[1], \ldots, \lambda_A[K], \lambda_A[K], \ldots\}$.

![Figure 5](image)

Figure 5: The noise influence. (a) Original chain $x^*$ ( ), noisy observation with SNR 10 dB ( ). (b) Unlimited adaptation: false alarms appear with the attenuation rise. (c) The minimum of $E_A$ with limited adaptation. (d) Post-smoothed solution.

6 Local post-smoothing

As $\lambda_A[m]$ decreases with $m$, the continuous parts of the solution tend to be increasingly rough in presence of noise (Fig. 5 (b)). It is convenient to operate a local post-smoothing by computing the minimum of the true energy $E$ conditioned to the discontinuity set, obtained via $E_A$. Note that $E$ is then convex in $x$ and this calculation is very easy. The results in the Figs 5 (c), 7 (d) and 8 (b) have been post-smoothed.
7 GNC optimisation

For a real-world problem we are no longer able to compute the exact optimum: for a longer signal, or one having a greater variation, or when B is not diagonal, the Viterbi algorithm becomes numerically inapplicable. We then focused on the graduated non-convexity (GNC), which was developed by Blake & Zisserman [1] in order to compute the MAP estimate in the case when \( A = I \). The basic idea is the following. A family of continuously derivable functions \((F_r)_{r \in \mathbb{R}^+}\) is constructed, such that: (a) \( \lim_{r \to \infty} F_r = \mathcal{E} \) (the \( F_r \)'s are approximations of \( \mathcal{E} \)), and (b) for some \( r_0 \), \( F_{r_0} \) is convex.

Let \( \{ r_0, r_1, \ldots \} \) be the relaxation sequence. \( F_{r_0} \) has an unique minimum \( \hat{\mathbf{z}}_0 \); starting with it, a sequence of minima \( \hat{\mathbf{z}}_i (\hat{\mathbf{z}}_i = \arg \min F_{r_i}) \) is tracked by local descent in the vicinity of the previously computed minimum \( \hat{\mathbf{z}}_{i-1} \).

Practical convergence of GNC is very satisfactory [2].

When \( A = I \), the \( F_r \) are obtained by fitting a quadratic spline at \( \pm T \); dropping indexes \( i \), \( \phi_r \) is [1):

\[
\phi_r(t) = \min(\{ \lambda T^2, \alpha - \frac{1}{2} r[i - p(r)]^2, \alpha \})
\]

where \( p(r) = T\sqrt{(r + 2\lambda^2)/r} \). Clearly \( \lim_{r \to \infty} \phi_r = \phi \). Transitions in the zone of the spline are undetermined. As \( r \) evolves to larger values, they leave this zone.

The convex energy function is found by checking the positive-definiteness of its Hessian matrix. When \( A = I \), convexity occurs for \( r < \frac{1}{2} \) [1]. In a recent paper [5] the authors showed that an ill-posed problem does not allow the family \{\( F_r \)\} to admit any convex function. That's why they append a small auxiliary convex term, which is relaxed to zero afterwards: \( F_{r,s} = F_r + s \sum ((I[i])^2) \) and \( F_{r,s} \) is convex when \( s > \frac{1}{r} \), \( \forall A \).

A crucial thing in GNC, which is not expressed explicitly, is that the first convex approximation must be "as close as possible" to the true energy \( \mathcal{E} \). For small \( r \) a great part of the transitions are in the undetermined zone: let us look at the Hessian \( \mathcal{H} \) in the extreme case:

\[
\mathcal{H} = 2I \otimes (B^T B - (r - 2s) D^T D).
\]

It suggests that the deformation imposed by the convex approximation upon undetermined transition at the end of the chain is much greater than at its beginning. That's why GNC fails, as it can be observed in the Fig. 6 (a). Supposing again that the \( u_m \)'s vary slowly, we can counterbalance this effect by taking \((r_A, s_A)\):

\[
r_A = \alpha u^2 \text{ and } s_A = \beta u^2.
\]

It is clear that initial convexity conditions remain unchanged. The obtained result is seen in Fig. 6 (b).

8 Simulations

We present the reconstruction, using the proposed technique, of a 128-point chain from its 64-point low-pass filtered and noisy Laplace transform, with attenuation \( u_m = e^{-0.25m} \) and with a SNR of 20 dB.

![Figure 7: Incomplete noisy Laplace transform: SNR=20 dB. 64 data points for a 128-point signal. Original chain (...), reconstruction (...). (a) Data - real part. (b) Data - imaginary part. (c) The MAP solution, computed by GNC. (d) The reconstruction using the proposed technique.](image)

![Figure 8: The same model as in Fig. 7. (a) MAP via GNC. (b) The proposed technique result.](image)

9 Conclusion

The presented analysis explains why the weak-string MAP estimate fails when the linear observation model is attenuating. We established an attenuation-adapted processing and deduced its application domain in function of the noise variance. The obtained improvement in comparison with the MAP estimate is considerable: data is more efficiently exploited. The optimisation is carried out by GNC, whose relaxation parameters have also to be adapted to the attenuation. By way of application, we show the inversion, using the proposed technique, of a noisy and incomplete Laplace transform.

References

Autoregressive with n Exogenous Inputs Model for Evoked Potentials
Analysis: Performance Analysis

L. Capitanio, G.C. Filligoi, A. Cecchini †, F. Babiloni †, L. Fattorini †, A. Urbano †, S. Cerutti ††

Dpt. INFOCOM, Fac. Engineering, Via Eudossiana 18, 00184 ROMA, Italy Tel/FAX: +39/6/4458.5478/4873300;
E-mail: filligoi@infocom.ing.uniroma1.it; capitanio@infocom.ing.uniroma1.it and Centro Interdipartimentale Sistemi Biomedici, Palazzo
dello Stato, Universita' degli Studi "La Sapienza", Corso Vittorio Emanuele II, 244-00186 ROMA, Italy Tel/FAX: +39/6/4991.8611/6833734

† Ist. Fisiologia Umana, Fac. Medicine, Universita' degli Studi "La Sapienza", P.le Albo Moro 5, 00185 ROMA, Italy Tel/FAX:
+39/6/4991.0989/4991.0942; E-mail: neurall@icaspur.caspar.it

†† D.I.S., Fac. Engineering, Via Eudossiana 18, Universita' degli Studi "La Sapienza", 00184 ROMA, Italy, Tel/FAX:
+39/6/4458.5871/4458.5367; E-mail: cerutti@icaspur.caspar.it

Abstract. The mean properties of event related potentials (ERP) recorded on the scalp are normally obtained by means of the synchronized
average of a set of recorded sweeps. A parametric identification procedure based on an AutoRegressive with n Exogenous inputs (ARXn)
model of cerebral signals generation is able to extract a reliable estimate of the ERP contained in each sweep. By means of simulated data,
the most important model performances have been analysed. In this paper are also reported some results concerning the application of the
ARXn model to data recorded during the execution of simple motor tasks.

1. Introduction

The study of the electrical activity of the brain related to well
defined stimuli (EP, Evoked Potential) or to the accomplishment of
a specific task (ERP, Event Related Potential) has become a
useful tool for both clinical diagnosis and Central Nervous System
(CNS) functional studies [1,2]

As is well known, the major problem in these studies lies in the
unfavorable Signal-to-Noise Ratio (SNR) existing between the
useful signal and the various sources of noise. Separation of
the buried EP or ERP waveforms from the other electrical activity
is usually accomplished by means of synchronized average of
several successive responses. This standard technique is based on
the hypothesis of signal and noise additive superposition and
uncorrelation, provided that the signal is stationary and the noise
has a null mean value. However, this technique of SNR
improvement presents some limits and disadvantages: noise may
vary during recording; correlations between evoked responses and
background EEG are not well known; random phenomena,
sudden transients as well as non-stationarities in the signal
determine modifications in amplitude and latency of waves in
consecutive sweeps. Several methods have been proposed to
overcome these problems, such as latency correction techniques
[3], time-varying filters [4], Wiener filters [5]. Nevertheless, any
approach based upon averaging is unable to take into account the
variability among trials. This variability, which is important when
higher cognitive processes, like attention, memory and learning
are involved, becomes more evident when considering ERP
where the response is strongly dependent upon each trial. As a
consequence, the analysis of EP and ERP on a sweep-by-sweep
basis focused the attention of an increasing number of researchers
during the last decade [6-9].

In this paper, we shall be concerned with a parametric
 technique which is based upon a black-box model with both
deterministic and stochastic structure. The procedure describes the
single trial response as the sum of n noise contributions,
representing the EEG signal and various types of artifacts and
the ERP. The model introduced for the identification of the single
contributions is a AutoRegressive with n Exogenous inputs (ARXn).
The main concern of the paper is dedicated to the analysis of the
model performances when n varies from 1 to 3, and for several
SNR.

2. Material and methods

2.1 The ARXn model

The method described here makes a model M(Θ) (where Θ is a
suitable set of parameters) of the complex interactions existing in
the CNS (called S in the following) between the various
signals contributing to the recorded sweep. The electrical activity
y(t) recorded on the scalp has been modeled by means of a
stochastic parametric approach. In fact, the ARXn models are
concerned with those stochastic processes whose realization y(t)
can be decomposed into n (n=3) superimposed deterministic
distributions ydi(t) plus a stationary stochastic term yN(t).

\[ y(t) = \sum_{i=1}^{n} y_{di}(t) + y_N(t) \]  

(1)

Being k the discrete time domain variable, the recorded electrical
activity y(k) is considered as the sum of three deterministic
contribution, where, as an example, y1(k) may represent the
single sweep ERP (SSERP) recorded during the present trial,
y2(k) and y3(k) are signals captured by the electrodes on the
scalp and due to the propagation along the skull the vertical and
horizontal ocular artifacts, plus a random noise representing the
background EEG activity. Referring to Figure 1, each ydi(k)

Figure 1 Schematic diagram of the ARX3 model.

The output of a Linear and Time Invariant (LTI) filter \( H_i \), while \( Y_N(k) \) is a realization of a stochastic process with zero mean that can be modeled, according to the spectral factorization theory of Astrom [10], as the output of a LTI system driven by a white noise \( \xi(k) \). The relationship between the recorded signal \( y(k) \) and the above four contributions \( y_j(k) \), \( y_0(k) \), \( y_3(k) \) and \( Y_N(k) \) can be expressed in the z-domain as follows:

\[
Y(z) = \frac{1}{d_1} H_1(z)X_1(z)z^{-d_1} + H_2(z)X_2(z)z^{-d_2} + H_3(z)X_3(z)z^{-d_3} + \xi(k)
\]

Therefore, the sample path \( y(k) \) representing a realization of the process \( Y(k) \) is completely defined by \( H_0(z) \), \( H_1(z) \), \( H_2(z) \), \( H_3(z) \), \( X_1(k) \), \( X_2(k) \), \( X_3(k) \), \( \xi(k) \) and \( \eta_{th} \) and its variance \( \sigma_{th}^2 \) can be obtained by means of the following computation:

\[
\sigma_{th}^2 = \frac{1}{K} \sum_{k=1}^{K} \{ y(k) - \hat{y}(k) \}^2
\]

where \( K \) is the number of estimated samples of \( y(k) \). The series \( \eta_{th}(k) \) is obtained as a residual of the linear combination of \( x_1(k) \) and \( y(k) \) via the computed coefficients (Figure 2). The goodness of the identification is tested by means of the cumulative Anderson test [14] on the whiteness of the prediction error (95% confidence interval). The optimal orders \( n_a,n_b1,n_b2 \) and \( n_b3 \) should be chosen balancing the need to extract as much information as possible and the constraint to minimize the complexity of each filter. Therefore, the well known stochastic criteria Final Prediction Error (FPE) of Akaike [15] takes the optimal filters lengths in correspondence to a minimum of the function:

\[
FPE(\mu) = N + \mu + 1 + \sigma^2_{\varepsilon}
\]

where \( \mu \) is the sum of the filters orders, \( \sigma^2_{\varepsilon} \) is the variance of the prediction error and \( N \) is the length of the sample \( y(k) \).

2.2 Simulated data

In order to test the potentiality of the identification procedure, the performances of the ARX3 model, with \( n_a \) ranging from 1 to 3, has been tested by means of simulated data. This is mainly due to the fact that in real cases the SSERP (i.e. the final target of the identification algorithm) is practically unknown. Simulated data were generated, according to Figure 3a by taking as stochastic...
3. Results

3.1 ARX3 performance analysis on simulated data

Given the set of signals y(k), EEG(k) and xi(k) described in subsection 2.2.2, a first test on the model skill to correctly identify orders and parameters of the filters was carried out. The filters orders investigated ranged between 2 and 19.

In order to present one of the result obtained in a unique diagram, we show in Figure 4 the data referring to an ARX2 model. The 3-D diagram has been obtained by a particular choice of the coordinate axis: the z-axis represents the FPE value (eq.(6)), the y-axis is the order nb2 of the MA component of the exogenous channel #2 and the x-axis represents both the order na of the AR process 1/A(z) and the order nb1 of the channel #1. In particular x represents the values P=na+nb1 and conversely na=TRUNC(P/19) and nb1=P mod 19. The diagram clearly shows that: FPE is strongly dependent on na and is less sensitive to nb1 and nb2; the optimization procedure is practically insensitive to na greater than 12, but values around 4 or 5 are sufficient in most applications. Similar results were obtained for ARX3 which behaved analogously.

As is well known, the accuracy of the whole model can be evaluated by testing the whiteness of the prediction error ε(k) by means of the Anderson test. In such test, the number NT of times that the autocorrelation function \( \hat{\phi}_x(k) \) of ε(k) overcomes the 5% confidence interval is counted. The performance of the ARX3 model when the number n of exogenous inputs increased from 1 to 3 have been examined calculating NT for each model.

![Figure 4. Final Prediction Error (FPE) as a function of the filters order](image)

<table>
<thead>
<tr>
<th>Config.</th>
<th>ER1</th>
<th>ER2</th>
<th>ER3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Config.</td>
<td>ER1</td>
<td>ER2</td>
<td>ER3</td>
</tr>
<tr>
<td>1</td>
<td>-15</td>
<td>-5</td>
<td>-20</td>
</tr>
<tr>
<td>2</td>
<td>-5</td>
<td>-5</td>
<td>-20</td>
</tr>
<tr>
<td>3</td>
<td>-5</td>
<td>-5</td>
<td>-30</td>
</tr>
<tr>
<td>4</td>
<td>-5</td>
<td>-20</td>
<td>-20</td>
</tr>
<tr>
<td>5</td>
<td>-15</td>
<td>-5</td>
<td>-5</td>
</tr>
<tr>
<td>6</td>
<td>-15</td>
<td>-20</td>
<td>-5</td>
</tr>
</tbody>
</table>

Table 1. Energy Ratios configurations used in the simulation.
3.2 EEG brain maps

The actual study is currently applied to the analysis of physiological signals recorded on the scalp to give the EEG brain maps. The technique used and the results herewith illustrated are successfully applied to analyze the course of the motor cortex activation during the execution of an elementary movement. Already published results [9] showed that ARX model is able to reduce the influence of ocular artifacts on those channels closer to the eyes. Moreover the procedure is clearly able to enhance the event related component of the signal, as is shown in Figure 7.

4. Conclusions

The simulation results showed that with lower energy ratio, the performances of the procedure ARX with n=1 or n=2 or n=3 are comparable, whereas with higher energy ratios ARX behaves better than ARX and ARX. On the other hand this improvement is counterbalanced by longer computation time. Concerning the real data, the procedure showed that it clearly enhanced the event related component of the signal and that ARX model is also able to remove ocular artifacts (see also [9]).

5. References

[14] G.E.P. Box, G.M. Jenkins (1976); Times series analysis forecasting and control; Holden day, S. Francisco (USA)
Selfsimilar Processes and Kalman Filtering

Olivier MAGRE and Michel GUGLIELMI

Laboratoire d'Automatique de Nantes, U.R.A. C.N.R.S. 823
Ecole Centrale de Nantes/Université de Nantes
1 rue de la Noë, 44072 NANTES CEDEX, FRANCE
Tel:(33) 40 37 16 00/Fax:(33) 40 37 25 22
E-mail magre@lan.ec-nantes.fr
members of the GR TDSI (france)

Abstract : This paper deals with the analysis of fractal signals. We propose a new model able to represent this kind of signals. The signal is considered as a solution of a linear differential time-varying equation depending of the fractal dimension. Thus, we develop a real time method intended to identify this basic parameter. This algorithm uses the well-known Extended Kalman Filtering technic. Finally, we present some results issued from simulations.

Introduction.

In various situations ( geophysical phenomenous, economic data ...), signals cannot be modeled by the means of linear classical models (AR, ARMA). These models are only useful for signals with rational spectrum and are unable to represent a wide class of signals as selfsimilar processes even if some authors have tried to use infinity order AR models [1]. The fractional brownian motion ( fBM ), first introduced by B.B Mandelbrot [2], has the basic property of $1/2$ spectrum. On time domain, fBM represents long time correlation signals.

The modelization of this signal is still an open problem. Several technics are used : middle point construction, inverse spectral definition. None of them verify all properties of fBM. In this paper, the first part is dedicated to a new approach to modelize fBM. It uses a differential equation which is linear but with time-varying parameters.

Moreover, the identification of the fractal dimension, main parameter of fBM, is crucius. The second part of this paper deals with this problem. Issued from the differential model, we design a real time algorithm to estimate the fractal parameter.

This paper is organized as following : first, we present fBM as a serie expansions. Each term of this serie is solution of a simple differential equation of first order with a time-varying parameter function of the fractal dimension. In the second part, we analyze this model. More specially, we look after the autocorrelation function computed from this model. The third paragraph is devoted to the main problem : identification of the fractal dimension by the means of an estimation problem.

After discretization of continuous model, we design an optimal filter based on the Extended Kalman Filtering. At last, we compare the performances of this algorithm to other technics of indentification.

1 Definitions.

The fBM models, first defined by B.B Mandelbrot in [2], is expressed by the equation:

$$B_H(t) = \frac{1}{\Gamma(H + \frac{1}{2})} \int_{-\infty}^{0} \int_{s}^{0} [(t-s)^{(H-\frac{1}{2})} - (-s)^{(H-\frac{1}{2})}]dB(s)$$

$$+ \int_{0}^{t} (t-s)^{(H-\frac{1}{2})}dB(s).$$

(1)

with : $dB(s)$ is a white gaussian noise, $\Gamma(\alpha)$ the Gamma function and $H$ the fractal parameter ($H \in [0,1]$).

For the negative $t$, the definition is obtained with: $B_H(0) = 0$ et $t = -t$.

These signals have a correlation function equal to:

$$\varphi_{BH}(t, t) = |t - t_1|^H \sigma^2_H$$

(2)

with :

$$\sigma^2_H = \frac{1}{\Gamma(H + \frac{1}{2})} \int_{-\infty}^{0} (1-s)^{(H-\frac{1}{2})}$$

$$- (-s)^{(H-\frac{1}{2})}ds)^2 - \frac{1}{2H}$$

and a spectrum :

$$\Phi(f) = \frac{1}{f^{2H+1}}$$

(3)
The fractal parameter $H$ linked with the fractal dimension $D$ with $D = 2H + 1$ is the basic parameter which characterizes the fbm. For instance, the fbm is for:

- $H = 0.5$ a classical Brownian motion.
- $H < 0.5$ a negative correlation fbm.
- $H > 0.5$ a positive correlation fbm.

An other equivalent expression of (1), in [4] is the Barnes and Allan one:

$$x(t) = \frac{1}{\Gamma(H + \frac{1}{2})} \int_0^t (t - \tau)^{(H - \frac{1}{2})} \omega(\tau) d\tau.$$  

Where $\omega(t)$ is a white noise centered with a variance $\sigma^2$.

The analytical development of the fractional brownian motion [7] can easily be expressed with the following serie:

$$x(t) = \frac{1}{\Gamma(H + \frac{1}{2})} \left[ \frac{1}{2} (H - \frac{1}{2}) \right] \int_0^t \omega(\tau) d\tau + ...$$

An interesting point of view is to consider only the first term of (4) and to define the following signal:

$$x_1(t) = \frac{1}{\Gamma(H + \frac{1}{2})} \int_0^t \omega(\tau) d\tau$$

This signal is a solution of the differential equation of a first order:

$$\dot{x}_1(t) - \frac{H - \frac{1}{2}}{t} x_1(t) = \frac{1}{\Gamma(H + \frac{1}{2})} (H - \frac{1}{2}) \omega(t)$$

Moreover, the correlation function is obtained by:

$$\varphi_{x_1(t), x_1(s)} = \frac{\sigma^2}{\Gamma(H + \frac{1}{2})^2} t^{(H - \frac{1}{2})}$$

The estimation of the fractal dimension $H$ is a crucial problem. This can be done by the mean of spectral analysis (for instance on the log-log graphic of the spectral density) or the maximum likelihood method [5].

We propose another solution based on the analysis of the model [8]. As classically, the identification problem of parameters in a differential equation can be solved by the means of optimal filtering. Roughly speaking, this method is the following: from equation (6), we can get a non linear state system written as:

$$\begin{bmatrix} \dot{x}(t) \\ \dot{a}(t) \end{bmatrix} = \begin{bmatrix} 0 \\ -a(t) \end{bmatrix} \frac{1}{t}$$

$$y(t) = [1, 0] \begin{bmatrix} x(t) \\ a(t) \end{bmatrix} + \omega(t)$$

with $a(t) = \frac{(H - \frac{1}{2})}{t}$

With $Z(t) = \begin{bmatrix} x(t) \\ a(t) \end{bmatrix}$ the extended state, the system becomes:

$$\begin{bmatrix} \dot{x}(t) \\ \dot{a}(t) \end{bmatrix} = \begin{bmatrix} 0 \\ - \frac{1}{t} \end{bmatrix} Z(t)$$

$$y(t) = H \begin{bmatrix} x(t) \\ a(t) \end{bmatrix} + \omega(t)$$

with $H = [1, 0]$

This is a non linear state model in which the problem of identification of $H$ is transformed in a estimation problem of $Z(t)$. To perform this estimation, we can use the well-known Extended Kalman Filtering method (EKF). We have to discrete the continuous signal $x(t)$ or to apply the continuous-discret time theory [9].

2 Modelisation and Discretization.

In this paragraph, we consider the first approach and decide to discretise the estimation problem. The discrete model must have an impulse response with the same form that the continuous one has. Moreover, the variance and the correlation functions must be the same too. Thus, the discrete signal is:

$$x_{k+1} = (1 + \frac{1}{k})^{(H - \frac{1}{2})} x_k + \frac{1}{k} \int_0^\tau (H - \frac{1}{2}) \kappa(H - \frac{1}{2}) \omega_k$$

1071
So, the variance is:

$$P_{k+1} = E[x_{k+1}x_{k+1}] = \frac{1}{\Gamma(H + \frac{1}{2})^2}\pi q(k+1)^{2H}$$

To be compared to (7).

The correlation function is:

$$\varphi_x(k, \tau) = \frac{1}{\Gamma(H + \frac{1}{2})^2}\pi q(k+1)^{2H} \prod_{i=1}^{r}(1 + \frac{1}{k+i})^{(H-\frac{1}{2})}$$

Compare this expression to (8). With the following figures, we could see the both form of the correlation functions.

![Figure 1: both discrete and continu correlation functions for H=0.2](image)

![Figure 2: both discrete and continu correlation functions for H=0.5](image)

3 Estimation of H.

The estimation of the fractal dimension H can be obtained from the tracking of $a(t)$. Or, as we realized, by the direct tracking of the constant parameter $H$ leading to an extended state: $Z(t) = \begin{pmatrix} x(t) \\ h(t) \end{pmatrix}$ and further more with discrete time $k$: $Z_k = \begin{pmatrix} x_k \\ h_k \end{pmatrix}$. We could see that the discretisation was not a trivial problem due to the fact that the equation can be instable for some values of $H$ (for $H > 0.5$). The discrete technic leads to an iterative algorithm too. At each step $k$, we have to do a prediction which gives the predicted state $Z_{k+1|k}$ from the estimation $Z_{k|k}$. The non linear dynamical equation $Z_{k+1} = f_k(Z_k)$ must be developped in Taylor expansion:

$$Z_{k+1} = f_k(Z_{k/k}) + F_k(Z_k - Z_{k/k}) + \Gamma \mu_k$$

where $Z_{k/k}$ is the estimation to the step $k$, $\mu_k$ the higher order and $F_k$ is the first derivative of $f_k$ evaluated at $Z_{k/k}$.

The well-known Extended Kalman Filtering equations lead to matrix $A_k$ and $G_k$ like:

$$Z_{k+1} = A_k Z_k + G_k \omega_k$$

$$A_k = \begin{pmatrix} (1 + \frac{1}{k})^{(Hk-\frac{1}{2})} & 0 \\ 0 & 1 \end{pmatrix}$$

$$G_k = \begin{pmatrix} \frac{1}{\Gamma(H+\frac{1}{2})^{(Hk-\frac{1}{2})}} \\ 0 \\ 1 \end{pmatrix}$$

From $f_k$, we calculate $F_k$:

$$F_k = \begin{pmatrix} \frac{1}{k}^{(Hk-\frac{1}{2})} & x_k \ln(1 + \frac{1}{k})(1 + \frac{1}{k})^{(Hk-\frac{1}{2})} \\ 0 & 1 \end{pmatrix}$$

If we neglige $\mu_k$, we can realise a Kalman Filtering on the linearised equation. If we define the innovation $\nu_k = \gamma_k - HZ_k$ and call $\Sigma_k = H^2 P_{k/k} H^T + \sigma_\nu^2 T$.

The equations of the Kalman Filtering are:

For the prediction step:

$$Z_{k+1|k} = f_k(Z_{k/k})$$

$$P_{k+1|k} = F_k P_{k/k} F_k^T + G_k \Sigma_k G_k^T$$

Where $P_{k/k}$ is the covariance matrix of errors:

$$P_{k/k} = E[(Z_k - Z_{k/k})^T(Z_k - Z_{k/k})]$$

Then, the second step of the EKF, is to compute the state $Z_{k+1|k+1}$ estimation, by including the measurement $y_{k+1}:

$$Z_{k+1|k+1} = Z_{k+1|k} + P_{k+1|k} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \Sigma_k^{-1} \nu_k$$

$$P_{k+1|k+1} = P_{k+1|k} - P_{k+1|k} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \Sigma_k^{-1} [10] P_{k+1|k}$$

This recurrent algorithm leads to a tracking convergent solution. It is well known that's impossible to prove the convergence but its behavior, based on simulation, presented in the following figures, seems to be correct.

1072
4 Graphic results.

In the final part of this paper, we present results on simulated signals. These ones are compared to those issued from a Log-Log graphic analysis of spectrum. We first synthesise the fractal signal from equation (11), and calculate his fractal dimension from a Log-Log graphic. Second time, we estimate by the Extended Kalman Discrete Filtering both fractal dimension and signal. Some graphic results are presented in figures [3] and [4] for different values of H. We present the main result of the Extended Kalman Filtering, as the tracking of the fractal parameter, and the simulated signal of our model.

5 Conclusions

We defined another way to modelize the fbm, issued to an one term reduction of the Barnes and Allan definition of the fractal signal. After a discretisation of our signal, we verified that both the variance and correlation functions were comparable in the both discrete and continu cases. Further more, we gave a discrete algorithm of EKF and present some graphic results issued from this one. The problem of identification of the fractal parameter is one of the most crucious in the nonstationary domain, in particular, with the long time dependence signals. So, we propose some further development : an estimation of the fractal parameter of signal issued from other synthesis algorithm of fbm, as the midpoint one or the spectral one. And finally, natural signals will be treated in order to estimate their fractal parameter.

References

System Theoretical Approach to Quantum Mechanical Phenomena

Yosimi MONDEN, Shinobu UEZONO and Masayasu NAGAMATSU

Department of Computer and Information Science
Faculty of Science, Shimane University
Nishikawatsu, Matsue, 690 Japan

Abstract. We will propose a system theoretical approach to one-dimensional quantum mechanical phenomena in the context of characterizing lossless physical systems based on their input-output relationships, i.e., independently of their equations of motion. To achieve this end, introducing $2 \times 2$ transfer matrices, we will first deduce their structural and axiomatic properties from the stationary Schrödinger equations for step potentials, and then classify and specify their quantum mechanical states based on the transfer matrices together with their structural and axiomatic properties. This will put one step towards our goal of constructing the axiomatic system theory for more general physical systems independently of their equations of motion.

1. Introduction

Classical and quantum mechanics are considered as established in that their equations of motion as well as elementary constituents and fundamental interactions are already well-known. In such an established mechanics, every physical phenomenon can be viewed in the context of either the direct problem of determining an input-output relationship from a given equation of motion, or the inverse problem of identifying an equation of motion from a given input-output relationship. Both problems arise from mathematics or physics and have developed as cross-disciplinary research fields with the advent of computers. For example, the direct problem is now making the third science debut as scientific visualization of modelling, simulating and visualizing physical phenomena [1] [2], while the inverse problem is enlarging its area to cover the inverse spectral problem, profile inversion and image reconstruction [3] and is combining with the cross-disciplinary research fields such as DSP and system theory [4] [5].

Both the equation of motion and the input-output relationship mentioned above provide possibly equivalent means to describe physical phenomena and systems. Usually the input-output relationship can be deduced from the equation of motion with the resulting structural and axiomatic properties which are primarily implicit in the latter. Therefore the input-output relationship is more directly related to observation and is provided with the more tractable properties, compared with the equation of motion. A couple of these properties shared with a variety of physical systems are "lattice forms" and "losslessness" corresponding to the typical structural and axiomatic ones respectively.

Lattice forms attracted wide attention in digital signal processing, when Wakita showed in 1971 under appropriate input and boundary conditions the equivalence of reflection coefficients of a transmission line and PARCOR coefficients of a linear prediction problem [6]. This relationship essentially relied on the LD(Levinson-Durbin) algorithm and readily led to lattice realization of lossless transmission lines with the help of the notion of transfer functions.

Since then, the lattice forms have been studied extensively in a variety of research fields, not only because they are the most promising canonical forms in estimation, but also because they have many other advantages of lowest computational complexity, good numerical behaviour, stability "by inspection" property and relations to physical properties such as reflection coefficients [7].

Though lattice forms associated with the LD algorithm appear in many contexts such as scattering and network theories, the theory of orthogonal polynomials and stability analysis, they are only restricted to a special type of lossless transmission lines of classical mechanics whose propagation velocities and real reflection coefficients are assumed to be independent of position and/or frequency. Therefore it is highly desirable to generalize these lattice forms without assuming the frequency or position independency and propose an system theoretical approach to lossless physical systems independently of their equations of motion [8], [9].

To attain that end, we will take a one-dimensional quantum mechanical system as a typical of lossless physical systems to deduce structural and axiomatic properties of the transfer matrices from the stationary Schrödinger equation, because quantum mechanical systems exhibit more versatile structure than classical mechanical ones as summarized below:

1. losslessness in quantum mechanics comes from conservation of probability, while it is related to conservation of energy in classical mechanics,

2. wave functions generally take complex values in quantum mechanics, while they are always real in classical mechanics,
3. "Energy" in quantum mechanics plays the same role as "frequency" in classical mechanics and takes negative values corresponding to bound states,

4. propagation velocity and reflection coefficients of quantum mechanical systems generally depend on frequency and/or position unlike those of classical mechanical ones such as sound propagation within acoustical tubes,

5. quantum mechanical systems have a variety of states according to their boundary conditions and total energies.

Section 2 is devoted to derivation of structural properties of elementary lattice forms from the stationary Schrödinger equation for step potentials. In Section 3, introducing the notion of transfer matrices to represent the input-output relationship of physical systems, we will deduce a set of axiomatic properties for the matrices, which are applied in Section 4 to classification and specification of quantum states according to their boundary conditions.

2. Schrödinger Equation & Lattice Forms

Let us consider a one-dimensional quantum mechanical system in a step potential of \((N - 1)\) finite regions

\[
V(x) = \begin{cases} 
V_0, & x < x_1 \\
V_m, & x_m \leq x < x_{m+1}, \ 0 < m < N \\
V_N, & x_N \leq x
\end{cases}
\]

governed by the stationary Schrödinger equation

\[
-\hbar^2 \frac{d^2}{dx^2} \psi + V(x)\psi = E\psi
\]

(1)

for the wave function \(\psi(x)\) belonging to the energy \(E\) [10]. Within an m-th region for \(0 \leq m \leq N\), we can easily solve (1) to obtain a local wave function

\[
\psi_m(x) = A_m \exp(ik_m x) + B_m \exp(-ik_m x)
\]

for some complex \(A_m\) and \(B_m\) where \(d_m = x_{m+1} - x_m\) and \(k_m\) is defined by

\[
k_m = \begin{cases} 
\sqrt{2M(E - V_m)/\hbar}, & E > V_m \\
1/\sqrt{2M(V_m - E)/\hbar}, & E < V_m
\end{cases} \quad m = 0, \cdots, N.
\]

Owing to the continuity condition that the wave function should be continuous and continuously differentiable at the boundaries \(x_m, m = 1, \cdots, N\), \(A_m\) and \(B_m\) of contiguous regions are linked through the following relation

\[
\begin{pmatrix} A_m \\ B_m \end{pmatrix} = Q_m \begin{pmatrix} A_{m-1} \\ B_{m-1} \end{pmatrix}
\]

\[
Q_m = (1 + \mu_m)^{-1} \begin{pmatrix} z_m & \mu_m z_m^{-1} \\ \mu_m z_m & 1 \end{pmatrix}
\]

where \(z_m = \exp(ik_m d_m)\) and the m-th reflection coefficient \(\mu_m\) is defined by

\[
\mu_m = \frac{k_m - k_{m-1}}{k_m + k_{m-1}}.
\]

These relations are in the same form as those of the acoustic tube model [8] except that the reflection coefficient \(\mu_m\) generally depends on energy \(E\) and sometimes takes complex values with \(|\mu_m| = 1\). For the sake of convenience, we will assume \(E \neq V_m, m = 0, 1, \cdots, N\) without loss of generality [11]. Then the wave function \(\psi(x)\) can be represented in terms of these \(\psi_m\) in the following way:

\[
\psi(x) = \begin{cases} 
\psi_0(x), & x < x_1 \\
\psi_m(x - x_m), & x_m \leq x < x_{m+1}, \ 0 < m < N \\
\psi_N(x - x_N), & x_N \leq x
\end{cases}
\]

Structural properties of these lattice forms readily follow from elementary calculations as shown in Table 1 and 2,

| Table 1: Structural Properties of Lattice Forms (1) |
|-----------------|-----------------|-----------------|
| \(k_{m-1}\) | \(k_m\) | \(Q_m\) | \(Q_m^*\) |
| r | r | \(\Gamma Q_m \Gamma\) | \(Q_m^* \Gamma Q_m\) |
| i | r | \(k_{m-1} \Gamma Q_m^{-1}\) | \(Q_m^{-1} \Gamma k_m\) |
| r | i | \(k_{m-1} \Gamma Q_m^{-1}\) | \(Q_m^{-1} \Gamma k_m\) |
| i | i | \(k_{m-1} \Gamma Q_m^{-1}\) | \(Q_m^{-1} \Gamma k_m\) |

| Table 2: Structural Properties of Lattice Forms (2) |
|-----------------|-----------------|-----------------|
| \(k_{m-1}\) | \(k_m\) | \(Q_m^* J Q_m\) | \(Q_m^* \Gamma J Q_m\) |
| r | r | \(k_{m-1} J\) | \(k_m \Gamma J\) |
| i | r | \(k_{m-1} J\) | \(k_m \Gamma J\) |
| r | i | \(k_{m-1} J\) | \(k_m \Gamma J\) |
| i | i | \(k_{m-1} J\) | \(k_m \Gamma J\) |

where "r" and "i" denote \(k\)'s being real and imaginary respectively, and \(2 \times 2\) matrices \(J\) and \(\Gamma\) are defined in the following manner:

\[
J = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.
\]

3. Axiomatic Properties of \(T_N\)

There are a couple of alternative methods for representing the input-output relationship of a one-dimensional physical system [12]. One is based on a scattering matrix \(S\) relating incoming waves to outgoing ones at both sides \(x\) and \(x_N\) of the finite region and the other is based on a transfer matrix \(T\) relating incoming and outgoing waves at \(x\) to those waves at \(x_N\). These \(2 \times 2\) system matrices are related with each other in the following manner:

\[
S = \begin{pmatrix} t_{11} & t_{12} \frac{1}{2} \frac{1}{2} \\ -t_{21} \frac{1}{2} \frac{1}{2} & t_{22} \end{pmatrix}
\]

\[
T = \begin{pmatrix} s_{11} & s_{12} \frac{1}{2} \frac{1}{2} \\ -s_{21} \frac{1}{2} \frac{1}{2} & s_{22} \end{pmatrix}
\]

1075
where $s_{ij}$ and $t_{ij}$ denote $(i,j)$ elements of the system matrices $S$ and $T$ respectively.

They usually enjoy a set of axiomatic properties resulting from the physical axiomatic laws such as causality, invariances and conservation laws. The most important among these is "losslessness" which, in quantum mechanics, is related to the conservation of probability

$$\frac{d}{dt} \int \psi^\dagger dv + \int j \cdot dv = 0$$

for the probability current density vector

$$j = \frac{i\hbar}{2m}(\psi \nabla \psi^* - \psi^* \nabla \psi)$$

which is reduced to

$$j_x(x_1) - j_x(x_N) = |k_0(A_0^2 - B_0^2)| - k_N(|A_N|^2 - |B_N|^2) = 0$$

in a one-dimensional stationary state.

Introducing a transfer matrix $T_N$ such that

$$\begin{pmatrix} A_N \\ B_N \end{pmatrix} = T_N \begin{pmatrix} A_0 \\ B_0 \end{pmatrix}$$

$$T_N = Q_0 \cdots Q_l$$

and using the structural properties of $Q_m$, we are led to those of $T_N$ as shown in Table 3

<table>
<thead>
<tr>
<th>$k_0$</th>
<th>$k_N$</th>
<th>$T_N$</th>
<th>$T_N^\dagger JT_N$</th>
<th>$T_N^\dagger T_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>$r$</td>
<td>$\Gamma T_N \Gamma$</td>
<td>$\frac{k_0}{k_N} \Gamma J$</td>
<td>$\frac{k_0}{k_N} \Gamma J$</td>
</tr>
<tr>
<td>$i$</td>
<td>$r$</td>
<td>$\Gamma T_N \Gamma$</td>
<td>$\frac{k_0}{k_N} \Gamma J$</td>
<td>$\frac{k_0}{k_N} \Gamma J$</td>
</tr>
<tr>
<td>$r$</td>
<td>$i$</td>
<td>$T_N \Gamma$</td>
<td>$\frac{k_0}{k_N} J$</td>
<td>$\frac{k_0}{k_N} J$</td>
</tr>
<tr>
<td>$i$</td>
<td>$i$</td>
<td>$T_N$</td>
<td>$\frac{k_0}{k_N} \Gamma J$</td>
<td>$\frac{k_0}{k_N} \Gamma J$</td>
</tr>
</tbody>
</table>

from which readily follow a set of axiomatic properties for the transfer matrix $T_N$.

axiomatic property

1. $T_N$ is paraunitary iff both $k_0$ and $k_N$ are real and equal.
2. $(T_N)_{22}$ never vanishes as far as $T_N$ is paraunitary.
3. $T_N$ is unimodular iff $k_0 = k_N$.
4. $T_N$ has real trace iff both $k_0$ and $k_N$ are either real or imaginary.

These axiomatic properties can be easily derived by using the structural properties of lattice forms and transfer matrices in the following manner:

1. It suffices to show $T_N^\dagger JT_N = J$ iff $k_0 = k_N > 0$, which is possible in Table 3 only when both $k_0$ and $k_N$ are real and equal.

2. The desired result can be easily obtained from the paraunitarity of $T_N$

$$|t_{11}|^2 - |t_{21}|^2 = 1$$

$$|t_{22}|^2 - |t_{12}|^2 = 1$$

$$t_{11}t_{22} = t_{21}t_{12}$$

for $t_{ij} = (T_N)_{ij}$, $i, j = 1, 2$.

3. Obvious from

$$\det T_N = \det Q_N \cdots \det Q_l = \frac{k_N^{-1}}{i} \cdots \frac{k_0}{i} = \frac{k_N}{k_0}.$$

4. Suffices to show $\text{tr}(T_N^\dagger T_N)$ is which is obvious from Table 3 by taking $\text{tr}(T_N^\dagger T_N) = \text{tr}T_N$ into accounts.

These axiomatic properties will be applied in the next section to classify and specify the quantum mechanical states. In the sequel, for the sake of convenience, we will always assume $V_0 = V_N = V_\infty$ as well as $E \not= V_m, m = 0, 1, \ldots, N$ without loss of generality [11].

4. Classification of Q. M. States

With the aids of the transfer matrix $T_N$ together with its axiomatic properties, we are now in a position to classify quantum mechanical states into bound or scattering states according as their wave functions are normalizable or not for a potential with finite limits

$$V_0 = \lim_{x \to -\infty} V(x)$$

$$V_N = \lim_{x \to \infty} V(x)$$

and periodic states for a periodic potential $V(x)$ with periodic distance $d \equiv x_N - x_0$ for some finite $x_0$ such that

$$V(x + d) = V(x).$$

bound state

Bound states are eigenstates of the stationary Schroedinger equation which possess normalizable eigenfunctions $\phi^{(n)}$ belonging to discrete eigenvalues $E_n$ such that

$$[T_N, B]_{ij} = 0 \quad \text{for } E_n < V_\infty.$$

The normalizability of $\phi^{(n)}$ can be easily deduced by noting $A_0 = 0$ and $B_N = 0$, i.e., $\phi^{(n)}$ decreases exponentially as $x$ goes to $-\infty$

$$\phi^{(n)}(x) = \begin{cases} B_N \exp(\kappa_0 x), & -\infty < x < x_1 \\ A_N \exp(-\kappa_N x), & x_N < x < \infty \end{cases}$$

due to the axiomatic properties 1 and 2 as well as (2).

scattering state

Scattering states in which incident harmonic waves $A_0 \exp(ik_N x)$ and $B_0 \exp(-ik_N x)$ coming in from $-\infty$ and $+\infty$ respectively are scattered into $A_N \exp(ik_N x)$ and $B_N \exp(-ik_N x)$ going out to $+\infty$ and $-\infty$ respectively, occur only when $E > V_\infty$ and possess non-normalizable eigenfunctions belonging to continuous eigenvalues (the continuous spectrum). In the case of a single incident wave where a right-moving wave $A_0 \exp(ik_N x)$ incident from $-\infty$ is partly reflected into a left-moving wave $B_0 \exp(-ik_N x)$ and partly transmitted into a right-moving one $A_N \exp(ik_N x)$, every
scattering phenomena can be described by reflection and transmission functions given by the use of $T_N$ as

\[
R = -t_{21}t_{21}^* = s_{21} \\
T = t_{11} - t_{12}t_{21}^*s_{11} = s_{11}
\]

respectively.

**Periodic state**

Owing to periodicity of a potential $V(x)$, periodic states in which $\varphi(x+d)$ should be in the same state as $\varphi(x)$ require

\[
\varphi(x+d) = \lambda \varphi(x) \\
\varphi'(x+d) = \lambda \varphi'(x)
\]

for some complex $\lambda$ such that $|\lambda| = 1$. These relations can be rewritten in terms of the local wave functions $\varphi_m$ as

\[
\varphi_N(0) = \lambda \varphi_0(0) \\
\varphi_N'(0) = \lambda \varphi_0'(0)
\]

from which directly follows

\[
\begin{pmatrix}
A_N \\
B_N
\end{pmatrix} = T_N
\begin{pmatrix}
A_0 \\
B_0
\end{pmatrix}
= \lambda
\begin{pmatrix}
A_0 \\
B_0
\end{pmatrix}.
\]

This means $\lambda$ is the eigenvalue of $T_N$ whose characteristic equation can be written as

\[\lambda^2 - \text{tr}(T_N) \lambda + \text{det}(T_N) = 0.\]

Then noting the reality of $\text{tr}(T_N)$ and the unity of $\text{det}(T_N)$ for a periodic potential due to the axiomatic property 3 and 4, we can easily obtain "allowed" energy regions as

\[\text{tr}(T_N(E)) \leq 2\]

comprising the band spectrum for periodic states.

In this way, we can replace the stationary Schrödinger equation by a set of axiomatic properties of the transfer matrices together with the structural properties of the lattice forms.

**5. Conclusions**

In this paper, starting from the stationary Schrödinger equation for a step potential and introducing the notion of transfer matrices, we have deduced structural and axiomatic properties for the matrices. And conversely, starting from the axiomatic properties of the transfer matrices, we could classify and specify the quantum mechanical states successfully as bound states with the discrete spectrum, scattering states with the continuous spectrum and periodic states with the band spectrum. It is such a system theoretical approach that will play a crucial role in generalizing the lattice forms to more general physical systems and establishing the axiomatic system theory which could replace their equations of motion [9].

**References**

Application of Genetic Algorithms for Wavelet Networks Signal Modelling

Aleš PROCHÁZKA and Vladimír SÝS

Department of Computing and Control Engineering, Prague Institute of Chemical Technology, Technická 1905, 166 28 Prague 6, Czech Republic, Tel/Fax: +42 [0] 332 4259 / 243 11082, E-Mail: prochaz@vscht.cz, sysv@vscht.cz

Abstract. The wavelet network has been presented recently as an efficient feedforward artificial neural network with wavelet transfer functions to solve approximation problems in nonlinear systems. The paper presents the fundamentals of wavelet networks and contributes to their analysis in the area of the initial estimate of network coefficients, taking into account the local influence of wavelets in the time and frequency domains. The main part of the paper is devoted to the study of an efficient wavelet network learning process based on genetic algorithms. The learning efficiency is assessed for different evolutions of genetic operator probabilities, and a comparison of the process is made with other approximation procedures. Results of modelling physiological signals both with wavelet and sigmoidal neural networks are presented in the final part of the paper.

1. Introduction

Nonlinear signals identification and modelling are basic tools in many engineering, biological and economical applications involving digital signal processing and time series classification and prediction. As given systems are non-stationary in most cases, various methods are studied for their analysis including the use of the wavelet transform [1, 2] and different adaptive methods.

This paper contributes to this analysis by presenting neural networks [3] with wavelet transfer functions [4, 5] instead of sigmoidal ones. The study of initialization and optimization of such wavelet networks [6, 7, 8] by genetic algorithms [9, 10, 11] is then introduced and used for electroencephalogram (EEG) signal approximation.

2. Fundamentals of Wavelet Networks

Wavelet networks are based upon the use of wavelet functions usually derived from an initial (mother) wavelet \( h(x) \) dilated and translated by values forming column vectors \( d = [d_1, \ldots, d_{d_1}] \) and \( t = [t_1, \ldots, t_{d_1}] \) respectively. For a given sequence of input \( \text{pattern} \) values \( \text{P} = [P_1, \ldots, P_Q] \) it is possible to define values of individual wavelets by separate rows of the matrix

\[
H_{S_1,Q} = \begin{bmatrix}
    h(1/d_1(P - t_1)) \\
    \vdots \\
    h(1/d_{d_1}(P - t_{d_1}))
\end{bmatrix} = [h(W1 \cdot P + B1)]
\]

for column vectors of the first layer coefficients in the form

\[
W1_{s_1,i} = [1/d_1, \ldots, 1/d_{d_1}]^T \\
B1_{s_1,i} = [-t_1/d_1, \ldots, -t_{d_1}/d_{d_1}]^T
\]

This set of wavelet functions can be used as a basic structure of a wavelet network as shown in Fig. 1, evaluating the

![Figure 1. A two layer wavelet network of structure 1 - S1 - 1 used for EEG signal approximation and the relation between wavelet dilation and its spectrum estimation for an initial wavelet function \( h(x) = -2 e^{-x^2/2} \) network output \( Y = [Y_1, \ldots, Y_Q] \) for a given pattern vector \( P \) and the second layer coefficients \( W2 \) in the form

\[
Y = W2 \ast H = \sum_{i=1}^{S1} W2(1,i) h(W1(i,1) \cdot P + B1(i,1))
\]

This structure represents a 1 - S1 - 1 neural network using wavelet transfer functions given for a chosen initial wavelet, as shown in Fig. 1 together with their spectra esti-
3. Initialization and Coefficient Estimation

There are various approaches to neural network and wavelet network initialization [12]. While a random choice of coefficients usually results in a very long optimization process, the use of wavelets provides an alternative [8, 7] based on their local influence as shown in Fig. 2.

The algorithm used in this paper for wavelet network initialization for modelling a given signal consists of the following steps:

- Pyramidal wavelet structure definition using $S$ couples of translation and dilation coefficients covering a given range of pattern values $P = [P_1, \ldots, P_9]$ and defining matrix $H_{S,Q}$ by Eq. (1).
- The choice of $S1 \leq S$ wavelets represented by $S1$ rows of matrix $H_{S,Q}$ having the largest contribution to reducing the approximation error.

To simplify the initialization process it is possible to apply the Gram-Schmidt decomposition of any matrix $H_{S,Q}$ into the product of the lower triangular matrix $A_{S,S}$ and the row-orthonormal matrix $O_{S,Q}$ to evaluate the network error vector between the target values $T = [T_1, \ldots, T_9]$ and the network output values, using the relation

$$E = T - W2 \ast H = T - W2 \ast A \ast O = T - V \ast O$$  (3)

for $V = W2 \ast A$. Defining the sum of squares in the form

$$E \ast E' = (T - V \ast O) \ast (T - V \ast O)'$$  (4)

it is possible to evaluate subsidiary weights

$$V = T \ast O'$$  (5)

giving the minimum of the sum of squares, with the value

$$E \ast E' = T \ast T' - V \ast V'$$  (6)

The choice of $S1$ fundamental wavelets can be combined with the initial matrix $H$ orthonormalization. In the first step, the most contributing wavelet is found giving the maximum value of vector $T + H^T$, with the corresponding row of matrix $H$ forming the first row of matrix $O$ before its normalization. In each following step that row of matrix $H$ is added to matrix $O$, which after its orthonormalization provides the maximum weight $V = T \ast O'$. Gram-Schmidt coefficients evaluated during the orthonormalization process form the lower triangular matrix $A$.

Results of the initialization process for a given EEG signal and the selection of $S1 = 7$ wavelets is shown in Fig. 2. Initial coefficients found in this way can be further optimized to minimize deviations between network output and target values in the mean square sense to achieve better results of signal approximation. Various optimization methods include genetic algorithms as well, which are especially useful for highly nonlinear systems.

4. The Use of Genetic Algorithms for Network Optimization

The principle of genetic algorithms (GA) is generally known and widely studied and used. The most famous work on the subject seems to be the Goldberg's book [10] but there are many other papers devoted to this topic published recently, providing further extensive bibliography [13, 14, 15]. We concentrate on implementation concepts here and on a brief discussion of setting parameters for the learning process.

The given time-series modelling described in the previous sections, with the one-hidden-layer perceptron-like network and wavelet activation functions, can be considered as a multi-dimensional optimization problem when one searches for both the dilations and the translations of wavelet functions which best fit the given data. These $2 \ast S1$ real parameters evolve through the GA and weight coefficients of the second layer are calculated by the least mean squares (LMS) method. The set of dilations and translations with associated weights represent the solution of the approximation task with a certain accuracy and it corresponds to an "individual" according to the GA terminology. Particular steps of the algorithm are as follows:

1. Generation of an initial population of individuals, which means that several network sets of coefficients are chosen at random in given ranges
2. Individuals' fitness evaluation resulting from deviations between given and estimated values
Table 1. Comparison of binary and Gray coding

<table>
<thead>
<tr>
<th>Integer</th>
<th>Binary code</th>
<th>Gray code</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>000</td>
<td>000</td>
</tr>
<tr>
<td>1</td>
<td>001</td>
<td>001</td>
</tr>
<tr>
<td>2</td>
<td>010</td>
<td>011</td>
</tr>
<tr>
<td>3</td>
<td>011</td>
<td>010</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>110</td>
</tr>
<tr>
<td>5</td>
<td>101</td>
<td>111</td>
</tr>
<tr>
<td>6</td>
<td>110</td>
<td>101</td>
</tr>
<tr>
<td>7</td>
<td>111</td>
<td>100</td>
</tr>
</tbody>
</table>

3. Parameter coding into (binary) strings of a predefined length in order to allow the use of genetic operators later

4. Random choice of individuals for the next generation according to their fitness values giving those networks with better results of data modelling a higher probability of being reproduced

5. Selection of random coefficient pairs to mate each other changing randomly their strings segments (cross-over)

6. Mutation operation turning over values of every bit in strings with relatively small probability

7. Coefficients of the current population conversion to find for each individual in the coded form its real value

In each generation the average error over the whole population can be evaluated as well as the best individual's error. The population evolves in described cycles until the solution meets a given accuracy.

The encoding is viewed as a problem which resists a more extensive application of the GA. Here two simple alternatives are employed:

• Binary coding
• Gray coding

with their codes compared in Tab. 4.

The probabilities with which the genetic algorithm operate seem to be critical for the efficiency of the learning process. Whitley [16] proposes increasing the mutation probability with time to sustain a population diversity while Etter [17] suggests reducing the mutation rate and augmenting the cross-over probability through generations. In this work three strategies are tested:

• Exponentially decreasing mutation and increasing cross-over probability
• Exponentially increasing mutation and decreasing cross-over probability
• Constant probabilities

Comparison of these choices is given in Tab. 4. The application of the first two alternatives to probabilities of mutation and cross-over provide very close results on contrary to their different strategies. The population average errors are lower in these cases with respect to that achieved by constant probabilities. The significance of binary and Gray coding seems to be not so important for the given data segment.

An example of the optimization process is shown in Fig. 3. This assumes an evolution of dilatation and bias coefficients for a single element only in a given range of these coefficients, implied by properties of corresponding wavelet functions and wavelet grid density. The error surface and its contour plot illustrate the problem of its global optimum search, together with selected results of a GA application.

Table 2. The mean values of population average errors achieved after 90 generations over 30 individuals

<table>
<thead>
<tr>
<th>Method</th>
<th>Binary code</th>
<th>Gray code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponentially decreasing mutation and increasing cross-over</td>
<td>0.2233</td>
<td>0.1944</td>
</tr>
<tr>
<td>Exponentially increasing mutation and decreasing cross-over</td>
<td>0.2199</td>
<td>0.2149</td>
</tr>
<tr>
<td>Constant probabilities</td>
<td>0.3243</td>
<td>0.3203</td>
</tr>
</tbody>
</table>

Figure 3. The genetic single element learning process for a given EEG signal approximation, together with the average error over the whole population of 18 solutions for the first 30 generations and the best individual error evolution.
Figure 4. Result of the EEG signal approximation by 1 - 7 - 1 wavelet network using selected wavelets before and after their coefficient optimization, compared with the sigmoidal neural network of structure 1 - 9 - 1 approximation

5. Conclusion

This paper presents some approaches to highly nonlinear signal approximation and their verification for a given EEG segment modeling. Basic results for a given number of selected wavelets, before and after their optimization, are presented in Fig. 4. In comparison with sigmoidal neural network approximation, the number of coefficients in the wavelet case can be lower to achieve a mean square error of the same order. Using the orthonormalization procedure it is moreover possible to achieve better results of approximation by adding further wavelets, orthonormal to previous ones, with no change of their subsidiary weights.

The problem of network optimization solved in the paper is based upon genetic algorithms, showing how the development of genetic operators probabilities affect the approximation error value. Both artificial neural networks and genetic algorithms were inspired from biology, and this paper presents their joint application for signal approximation as a basic tool for signal modeling, classification and prediction.

Acknowledgements

This paper has been supported by the Department of Computing and Control Engineering at the Prague Institute of Chemical Technology. Our thanks belong also to Dr. J. Jach from this Institute and to Dr. J. Smith from the South Bank University in London for their valuable comments. EEG data signals were kindly provided by Dr. V. Krajcik from the Neurological Department of the Prague Bulovka Hospital.

References

A New Spherically Invariant Joint Distribution Model for Image Signals

Frank MÜLLER and Bernd HÜRTGEN†

Institut für Elektrische Nachrichtentechnik, Rheinisch-Westfälische Technische Hochschule Aachen,
D-52056 Aachen, Germany, Tel/Fax: +49 (241) 807681 / 8888196, E-Mail: mueller@tnt.rwth-aachen.de
† Institut für Elektrische Nachrichtentechnik, Rheinisch-Westfälische Technische Hochschule Aachen,
D-52056 Aachen, Germany, Tel/Fax: +49 (241) 807677 / 8888196, E-Mail: huertgen@tnt.rwth-aachen.de

Abstract. Statistical models are useful for many purposes arising in signal processing contexts. Various types of statistical models for the joint probability density function (pdf) of image sources are known. In this contribution we confine on models based on spherically invariant random processes (SIRPs). Until now for the univariate density functions of SIRP-models either Meijer’s G-functions or the generalized Gaussian function have been used. The new model presented here uses symmetric stable distributions as univariate model. This leads to a more accurate treatment of sources with heavier tailed pdfs. Covariations and a parameter (named the characteristic exponent) describing the shape of the univariate stable distribution can be adjusted separately.

1. Introduction

Many algorithms in image processing and image coding involve some sort of statistical modelling. If a model is needed which determines the source statistics completely as is desired e.g. in information theory, random processes with specified joint distribution are a natural choice. In order to adapt the model to actual statistics, it should be parametric for most practical purposes. Additionally, the parameters should have an intuitive meaning and efficient parameter estimation procedures should exist. Often a parametric description with very few parameters (but enough to reflect the actual source statistics) is desired. The processes, this paper deals with, are of this type. Compared to Gaussian processes they involve only one additional parameter describing the univariate distribution shape.

Spherically invariant random processes (SIRPs) were first introduced by Blake and Thomas [1] as a natural generalization of Gaussian processes. Brehm [2] developed a comprehensive SIRP source model for telephone speech sources. He modelled the univariate probability density functions (pdfs) with Meijer’s G-functions. More recently, Du [3] developed a SIRP model for image sources and found a way to model the marginal pdf with the generalized Gaussian function, which is more appropriate in the case of image signals. Furthermore practical problems as e.g. parameter estimation were much easier to solve with his model compared to Brehm’s, due to the simple parametric formulation of the generalized Gaussian function.

Spherically invariant processes have the property that they are completely described by the univariate marginal density function and the linear statistical bindings (covariances or covariations) between the random variables. This allows for parametric modelling with these processes, which is desirable for most practical applications.

Most important, however, is that other tractable models (as e.g. Gaussian and/or identically independent distributed (iid)) often fail to reflect reality. For example, Du measured bivariate distributions of samples from mean-removed image blocks and found elliptically shaped pdfs [3]. Subband image statistics also can be modelled orders of magnitude better with SIRP models than with memoryless or Gaussian models [5]. In all these cases the Gaussian assumption as well as the iid assumption fails. In most applications, models with heavier tailed probability densities than the Gaussian are needed. This motivates the development of SIRP models with other marginal distributions than the Gaussian.

In this contribution a new SIRP model is developed, which differs from the existing ones in the univariate model for the marginal density. We model the marginal distribution as stable distribution [7]. The stable distribution has much thicker tails than the generalized Gaussian, which allows us to model real world phenomena including outliers more accurately with SIRPs as has been possible before.

It will be proven that a subset of all stable distributions (the so called $\alpha$-stable distributions) are valid marginal distributions of a SIRP (i.e. they lead to consistent joint pdfs, see [4]). The multivariate characteristic function of the resulting process is given.

The rest of the paper is organized as follows. After some preparations, spherically invariant random processes are defined and characterized by their most striking properties. Then the symmetric $\alpha$-stable ($\alpha$S) distribution which serves as a marginal distribution,
is recalled. Section 5 makes up the “core” of the paper, where the construction of the new model is given together with some properties. At last, we come up with some conclusions, and discuss possible future applications of the new model.

2. Preparations

Definition (characteristic function): Let \( X \) be a random variable with pdf \( f(x) \). The characteristic function (cf) of \( X \) is the function \( \phi \) defined by

\[
\phi(t) = \int_{-\infty}^{+\infty} e^{itx} f(x) dx = E e^{itX},
\]

where \( E \) denotes mathematical expectation. If \( X \) is a random vector with pdf \( f(x) \), the (multivariate) characteristic function \( \Phi(t) \) is given by

\[
\Phi(t) = E e^{t^T X},
\]

where \( t \) denotes a vector of the same dimension as \( X \).

Let \( Y \) be a vector consisting of independent Gaussian distributed random variables with unit variance and let \( A \) be a real-valued matrix of appropriate size. Then \( X = AY \) is multivariate Gaussian distributed with covariance matrix

\[
C = AA^T.
\]

The probability density of a multivariate Gaussian distributed random vector \( X \) with covariance matrix \( C \) is given by

\[
f_X(x) = (2\pi)^{-n/2}|C|^{-1/2}e^{-\frac{1}{2}x^T C x},
\]

where \( |C| \) denotes the determinant of \( C \) and \( n \) the dimension of \( X \). Its characteristic function is given by

\[
\Phi(t) = E e^{it^T X} = e^{-\frac{1}{2}t^T C t}.
\]

3. Spherically Invariant Random Processes

A SIRP is a random process defined by the property that any \( n \)-variate pdf of random variables taken from the process can be written as

\[
f(x) = \pi^{-n/2}|C|^{-1/2}g_n(s) \quad \text{with} \quad s = x^T C^{-1} x,
\]

where \( C \) denotes the covariance matrix of the process. Note, that the Gaussian process is included here as a special case:

\[
g_n(s) := 2^{-n/2}e^{-s/2},
\]

A decomposition theorem attributed to Yao [4] states the following. Each random vector \( Y \) taken from a SIRP has the same distribution as the random vector \( XZ \), where \( X \) is a Gaussian vector with the same normalized covariance as \( Y \) and \( Z \) is a positive random variable which controls the variance of the process.

We write:

\[
Y \overset{d}{=} XZ,
\]

where \( d \) denotes equal distribution.

For the density function of a SIRP then holds

\[
f_Y(x) = \int_0^\infty f_X(x,r)f_Z(r)dr.
\]

Here \( f_X(x,r) \) denotes the multivariate density function of a Gaussian process with standard deviation \( r \). \( f_Z(r) \) is an univariate density function called sigma density which controls the distribution of the variance. The resulting process is non-ergodic, as most natural processes (e.g. image and speech) are.

Another very elegant way to describe SIRPs is in terms of characteristic functions. The cf of a SIRP can always be written as

\[
\Phi(t) = h(u)
\]

with

\[
u = t^T C t.
\]

Hence, the cf of a spherically invariant process is "spherically invariant" itself (depends only on the quadratic form (11)). This description is indeed simpler than the one in terms of the pdf, since the function \( h(\cdot) \) does not depend on the dimension of the random vector any more.

The special case of a Gaussian process is given by

\[
h(u) = e^{-u/2}.
\]

4. Stable Distributions

A random variable \( X \) is stable (in the strict sense) if for any independent random variables \( Y, Z \) with the same distribution as \( X \) and any constants \( a, b \), there exists a constant \( c \) such that

\[
aY + bZ \overset{d}{=} cX.
\]

It can be shown, that stable distributions are the only possible limit distributions for sums of iid random variables [6]. This proposition is called the "generalized central limit theorem" and includes the well known "central limit theorem", which states that all limit distributions for sums of iid random variables with finite variance are Gaussian.

The most important class of stable distributions are the so called symmetric \( \alpha \)-stable or S\( \alpha \)S distributions. A random variable is S\( \alpha \)S if and only if its characteristic function is of the form

\[
\phi(t) = \exp(-\gamma|t|^{\alpha}),
\]

with \( \gamma > 0 \) and \( 0 < \alpha \leq 2 \).

\( \alpha \) is called the characteristic exponent and controls the shape of the distribution. \( \gamma \) is called the dispersion and behaves like the variance; i.e. \( \gamma \) is a measure of how the density spreads around its median. Note that for
\( \alpha = 1 \) the distribution is Cauchy and for \( \alpha = 2 \) Gaussian. In the Gaussian case \( \gamma \) equals half the variance.

An SfS variable is termed standard, if \( \gamma = 1 \). Every SfS variable with parameters \( \alpha, \gamma \) can be normalized by division through \( \gamma^{1/\alpha} \).

Only for a few parameters, closed form expressions for the density of stable random variables are known. Several power series expansions for stable densities have been reported. In particular, the standard SfS density function for \( 1 < \alpha < 2 \) is given, see [6][7]

\[
f(x) = \frac{1}{\pi \alpha} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k!} \Gamma \left( \frac{2k + 1}{\alpha} \right) x^{-2k}. \tag{15}\]

It can be shown, that the distribution tails of stable variables are algebraic for \( \alpha < 2 \). The Gaussian and the generalized Gaussian distribution in contrast have exponential tails. This on the one hand the reason for the different behaviour of stable distributions, but on the other hand this makes stable distributions suitable for modelling data which includes a considerable amount of outliers.

Definition: The random variables \( X_1, \ldots, X_n \) are said to be jointly SfS distributed if all linear combinations

\[
a_1 X_1 + \cdots + a_n X_n \tag{16}
\]

are SfS.

In the following we confine on the case \( 1 < \alpha \leq 2 \). Where not stated otherwise, this restriction holds throughout the rest of the paper.

5. The New Model
The crucial task in development of a spherically invariant joint distribution model is finding a consistent family of distributions. Consistency is guaranteed when the decomposition theorem is used. Modelling a process as the product of a variation controlling random variable and another stochastic process, yields consistent distributions. However, in this way it is not ensured, that a parametric formulation of the marginal densities can be obtained.

Here we will model the variation controlling random variable as stable distributed and concentrated on the interval \([0, \infty)\) and the process as Gaussian. In this way a spherically invariant process is obtained which follows from the decomposition theorem eq.(8). It will be shown, that this results in symmetric stable univariate with the parametric formulation eq.(14). Moreover, the resulting process consists of joint SfS random variables.

Theorem 1: Let \( X \) be an \( n \)-vector of independent normalized Gaussian random variables with \( c^T \)’s \( \exp(-t^2) \), let \( Z \) be a positive stable random variable independent of \( X \) with Laplace transform \( \exp(-t^{\gamma/2}) \) \((t > 0)\) and let \( A \) be a real-valued full rank \( n \times n \) matrix.

Then the random vector \( Y = AXZ^{1/2} \) consists of \( n \) stable random variables with a log cf

\[
\log \Phi_Y(t) = -\left( \sum_{k=1}^{n} \left( \sum_{i=1}^{n} a_{ik}^2 \right)^{\gamma/2} \right)^{\alpha/2} = -(t^TCt)^{\alpha/2} \tag{17}
\]

with

\[
C := AA^T
\]

The proof is given in Appendix A.

Particularly, the univariate characteristic function is given by

\[
\log \phi_Y(t) = -(c_{ii}t^2)^{\alpha/2} = -(c_{ii})^{\alpha/2} \cdot |t|^\alpha, \tag{18}
\]

and the bivariate characteristic function by

\[
\log \Phi_{Y_1, Y_2}(t) = -(c_{11}t_1^2 + 2c_{12}t_1t_2 + c_{22}t_2^2)^{\alpha/2}. \tag{19}
\]

A comparison between eq.(18) and eq.(14) yields

\[
\gamma_1 = (c_{ii})^{\alpha/2}. \tag{20}
\]

Theorem 2: All spherically invariant random processes with \( \alpha \)-stable univariate densities \((1 < \alpha \leq 2)\) can be represented by eq.(17).

Proof: The SfP property demands that the univariate density can be written as

\[
f(x) = cg(x^2), \tag{21}
\]

where \( c \) denotes a normalization constant.

Therefore \( f(x) = f(-x) \) is a necessary condition for a SfP. Thus, all possible \( \alpha \)-stable marginals have to be symmetric around the origin. These are exactly the SfS-distributions with characteristic function (14). Observe, that with proper adjustment of \( \alpha \) and \( C \) all characteristic functions (14) can be constructed. Since the mappings between univariate and multivariate distributions is one to one in the SfP case, the assertion is proven.

\( \Box \)

Theorem 3: Given the notations of theorem 1, the random vector \( Y \) consists of jointly SfS variables.

Proof: Let \( a \) denote an arbitrary real-valued \( n \)-vector. The linear combination \( a^TY = a^TAXZ^{1/2} \) then given by

\[
a^TY = a^TAXZ^{1/2}. \tag{22}
\]

\( a^TAX \) is a linear combination of Gaussian random vectors and therefore Gaussian as well. Thus \( a^TY \cong X'Z^{1/2} \) holds with \( X' \) being a Gaussian random variable. Application of theorem 1 with \( n = 1 \) concludes the proof.

\( \Box \)

The random vector \( Y \) has the linear regression property, i.e. the conditional expectations \( E(Y_i|Y_k) \) are linear functions of \( Y_k \) for every choice of \( i, k \). In this aspect the constructed joint distribution behaves like the Gaussian.

Numerous efficient estimation procedures are known for the parameters \( \alpha \) and \( \gamma \) of the univariate SfS-distribution are known. The parameters \( c_{ij} (i \neq j) \) may be estimated via the fractional lower order moments-approach. For details, see [7] and the references therein.
6. Summary and Discussion

In the paper a new spherically invariant distribution model has been discussed. It deviates from other SIRP-models in the employment of stable distributions for modelling the marginals. Computations may be more complicated than with second order models. However, a recent tutorial paper [7] shows, that signal processing algorithms for stable processes have been developed, which are computationally feasible with today's computer power. Using stable distributions, we gain a more realistic treatment of a big variety of sources.

The discussed model is theoretically appealing mainly because it generalizes Gaussian models in two respects. First, the spherical invariance is a natural generalization of Gaussianity since it implies the linear regression property. Second, the stable distributions are the only possible limit distributions in the generalized central limit theorem and include Gaussian distributions simply as special case.

There are many situations, in which signals can be thought of as sums of many independent identically distributed effects. Traditionally, in such cases Gaussian models are used. These models may fail when the distribution tails are thicker than the Gaussian law predicts. To solve this problem, it is common to treat a certain amount of the data (say 1 ~ 10%) as "outliers" which do not belong to the "true" distribution. From this considerations, a scientific discipline evolved, termed robust statistics or robust signal processing.

The concept of stable distributions allows now a unified treatment of "true" and "outlying" data in the context of robust signal processing. Interestingly, there are many algorithmic similarities in the methods used in robust statistics and the ones derived from stable process theory.

References


Appendix A

Proof of Theorem

**Theorem 1:** Let $X$ be an $n$-vector of independent normalized Gaussian random variables with cdf's $\exp(-t^2)$, let $Z$ be a positive stable random variable independent of $X$ with Laplace transform $\exp(-t^{a/2}) (t > 0)$ and let $A$ be a real-valued full rank $n \times n$ matrix.

Then the random vector $Y = AXZ^{1/2}$ consists of $n$ stable random variables with a log cf

$$
\log \Phi_Y(t) = - \left( \sum_{k=1}^{n} \left( \sum_{l=1}^{n} a_{lk} t_l \right)^2 \right)^{a/2} = -(t^T C t)^{a/2}
$$

with

$$
C := AA^T
$$

Proof:

$$
\Phi_Y(t) = E e^{j t^T Y} = E \exp\{j(Z^{1/2} A^T t)^T X\} = E \left\{ \exp\left( j \sum_{k=1}^{n} \left( Z^{1/2} \sum_{l=1}^{n} a_{lk} t_l \right) X_k \right) \right\} = E \left( \prod_{k=1}^{n} \exp\left( j \left( Z^{1/2} \sum_{l=1}^{n} a_{lk} t_l \right) X_k \right) \right) = E \left\{ \exp \left( - \left( Z^{1/2} \sum_{l=1}^{n} a_{lk} t_l \right)^2 \right) \right\} = \exp \left( - \left( \sum_{k=1}^{n} \left( \sum_{l=1}^{n} a_{lk} t_l \right)^2 \right)^{a/2} \right) = \exp \left( -(A^T t)^T (A^T t)^{a/2} \right)
$$

With $(A^T t)^T (A^T t) = t^T A A^T t = t^T C t$ follows the assertion.

\(\Box\)
Optimality Conditions for Truncated Kautz Series Based on Nth Order Allpass Sections

Albertus C. DEN BRINKER

Signal Processing Group, Department of Electrical Engineering, Eindhoven University of Technology
P.O. Box 513, NL-5500 MB Eindhoven, The Netherlands.
Tel.: +31 (0) 473 628, Fax: +31 (0) 448 375, E-mail: A.C.d.Brinker@ele.tue.nl.

Abstract. Kautz functions constitute a basis in $l^2$. In a practical situation, only truncated series are used and thus the convergence of such series is of importance. This convergence is a function of the parameters in the Kautz series. In this contribution, the optimality conditions for the parameters of a specific class of truncated Kautz series are established. These conditions are derived solely on the basis of the orthonormality and the polynomial character of the basis functions.

1. Introduction

Kautz functions [1,2] are a special case of orthonormal functions. These can be used in signal analysis and filter synthesis. More recently, adaptive filters on the basis of Kautz functions have been proposed [3].

In practice, a truncated Kautz series will be used and the question arises which parameters provide the best approximation to a given function given a limited number of expansion terms. This question has already been answered for a Laguerre series [4-6], which can be interpreted as a Kautz series governed by a single (real-valued) pole. The optimality condition for a Kautz series governed by a complex-conjugated pole pair is also known [6]. In this paper, these results are extended to a more general class of Kautz functions as will be discussed in Section 2. Some consequences of the orthonormality and the recurrent pole set for the derivatives of the basis functions with respect to the parameters are discussed in Section 3. These results are subsequently used to derive the optimality conditions (Section 4). A discussion concludes the paper.

2. The Considered Kautz Functions

Consider the discrete-time Kautz functions [2] generated as the impulse responses of the filter bank shown in Figure 1. The filter bank consists of a line of allpass filters $W$ with transfer function

$$W(z) = \prod_{i=1}^{N} \frac{1 - z p_i}{1 - z^{-1} p_i}.$$ (1)

All poles are distinct, and they are real-valued or occurring in complex-conjugated pairs. Furthermore, the allpass filter is stable: $|p_i| < 1$ for all $i$.

![Figure 1: A Kautz filter on the basis of Nth order allpass sections.](image)

The filter bank is tapped by filters $V$ where $V$ has one input and $N$ outputs. The vector $v_k(1)$ contains the $N$ impulse responses of this filter where

$$v_k(1) = (v_{k1}(1), \ldots, v_{kN}(1))^T.$$ (2)

The impulse responses $v_{ki}(k)$, $i = 1, \ldots, N$, are the result of an orthonormalization procedure on the sequences $\{p_i^k\}$. Furthermore, we require that all impulse responses $v_{ki}(k)$ are real-valued. Thus, the first $N$ basis functions are linear combinations of the sequences $\{p_i^k\}$ or in matrix notation

$$v_k(1) = Q \begin{pmatrix} p_1^k \\ \vdots \\ p_N^k \end{pmatrix}.$$ (3)

where $Q$ is a $N \times N$ matrix dependent on the poles (but not on the variable $k$) such that $(v_k, v_k') = I$, where $J$ is the identity matrix and $(\cdot, \cdot)$ denotes an inner product matrix containing inner products with respect to time index $k$. 

1086
Similarly to (2), the vectors \( v_m(k), m = 2, 3, \ldots \), contain \( N \) impulse responses with

\[
v_m = (v_{m1}(k), \ldots, v_{mN}(k))^t = v_{m-1}(k) * w(k),
\]

where \( w(k) \) is the inverse z-transform of \( W(z) \) and * denotes convolution. Consequently, all basis functions \( v_m(k) \) are real-valued. Furthermore, the \( m \)th vector contains impulse responses that are linear combinations of products of, at most, \( (m-1) \)th-order polynomials in \( k \) and exponential sequences \( \{p_i^k\} \). The set of functions \( v_m(k), m = 1, 2, \ldots; i = 1, 2, \ldots, N \), is an orthonormal basis in \( l^2(\mathbb{Z}_+) \) [2, 7].

3. Properties of the Basis Functions

The orthonormality of the basis functions can be expressed as

\[
(v_m, v_j) = \delta_{mj}, \tag{5}
\]

where \( \delta_{mj} \) is the Kronecker delta. From the derivative of the orthonormality condition (5) it is found that

\[
\left(\frac{\partial v_m}{\partial p_l}, v_j\right) = -\left(v_m, \frac{\partial v_j}{\partial p_l}\right) = -\left(\frac{\partial v_j}{\partial p_l}, v_m^t\right)' \tag{6}
\]

and, especially, the inner product matrix \( (\partial v_m/\partial p_l, v_m^t) \) is antisymmetrical. The \( N \times N \) matrices \( A_{mj} \) are introduced as

\[
A_{mj} = \left(\frac{\partial v_m}{\partial p_l}, v_j\right) = -(A_{jm})^t. \tag{7}
\]

For the sake of clarity it is noted that the indices \( m \) and \( j \) do not indicate entries in a matrix \( A_j \).

From (3) it is concluded that the vector \( \partial v_1/\partial p_l \) is a linear combination of products of, at most, first-order polynomials in \( k \) and exponential sequences. Thus, \( \partial v_1/\partial p_l \) can be expressed in the vectors \( v_1 \) and \( v_2 \):

\[
\frac{\partial v_1}{\partial p_l} = A_{11} v_1 + A_{12} v_2, \tag{8}
\]

and so \( A_{1j} = A_{j1} = 0 \) for \( j > 2 \). Extending this reasoning, it is found that \( A_{mj} = A_{j}^m = 0 \) for \( j > m+1 \). Consequently, the derivative of the \( m \)th vector with respect to pole \( p_l \) can be written as a combination of just three vectors:

\[
\frac{\partial v_m}{\partial p_l} = A_{m,m-1} v_{m-1} + A_{m,m} v_m + A_{m,m+1} v_{m+1} \tag{9}
\]

for \( m > 1 \).

Apart from (7), there exist relations between the \( A_{mj} \) matrices as a consequence of the recurrence relation (4). These will be considered in the remainder of this section.

From (9) we have

\[
\frac{\partial v_{m+1}}{\partial p_l} = \frac{\partial v_m}{\partial p_l} * w + v_m * \frac{\partial w}{\partial p_l}. \tag{10}
\]

For the first part of the right-hand side (9) can be substituted. The second part is equal to

\[
v_1 * \frac{\partial w}{\partial p_l} \{ * w \} \tag{11}
\]

where \( \{ * w \} \) means \( m - 1 \) times a convolution by \( w \).

Now consider (9) and (10) for \( m = 2 \)

\[
\frac{\partial v_2}{\partial p_l} = A_{21} v_1 + A_{22} v_2 + A_{23} v_3 \nonumber \quad = \frac{\partial v_1}{\partial p_l} * w + v_1 * \frac{\partial w}{\partial p_l}, \tag{12}
\]

Substitution of (8) in (12) reveals that the last term in (12) can be expressed in the vectors \( v_1 \) and \( v_2 \). To this end, the matrices \( D_{-1,0} \) and \( D_{0,0} \) are introduced:

\[
v_1 * \frac{\partial w}{\partial p_l} = D_{-1,0} v_1 + D_{0,0} v_2 + D_{0,0} v_3. \tag{13}
\]

Next, substitution of (9) for the left-hand side as well as for the first term of the right-hand side of (10) and using (11) and (13) for the last term yields

\[
A_{m+1,m} v_m + A_{m+1,m+1} v_{m+1} + A_{m+1,m+2} v_{m+2} \nonumber \quad = (A_{m,m-1} \tau + D_{-1,0}) v_m + (A_{m,m} + D_{0,0}) v_{m+1} + (A_{m,m+1} + D_{1,0}) v_{m+2}. \tag{14}
\]

Since the vectors \( v_m \) are independent in the \( k \)-domain, it can be concluded that for \( j = -1, 0, 1 \)

\[
A_{m+1,m+1,j} = A_{m,m+j} + D_j. \tag{15}
\]

Since \( A_{m,m} \) is antisymmetrical, so is \( D_{0,0} \) and

\[
A_{m,m} = A_{11} + (m - 1) D_{0,0}. \tag{16}
\]

From (15) it is also inferred that \( A_{m+1,m} = m D_{-1,0} \) and \( A_{m+1,m} = A_{12} + (m - 1) D_{1,0} \). From these two expressions together with (7) it is found that

\[
D_{i} = A_{i2}, \tag{17}
\]

\[
D_{-1} = -(A_{12})^t, \tag{18}
\]

and consequently

\[
A_{m+1,m} = -(m - 1)[(A_{12})^t], \tag{19}
\]

\[
A_{m,m+1} = m A_{21}. \tag{20}
\]

By (16), (19) and (20) the whole infinite set of matrices \( A_j \) is thus expressed in just the three matrices \( A_{11}, D_{0,0} \) and \( A_{12} \).

The matrix \( A_{12} \) (and, implicitly, the matrices \( A_{m,m-1} \) and \( A_{m,m+1} \)) is considered in more detail. Taking the derivative of (3) with respect to \( p_l \) yields

\[
\frac{\partial v_1}{\partial p_l} = \frac{\partial Q}{\partial p_l} \begin{pmatrix} p_i^k \\ \vdots \\ p_N^k \end{pmatrix} + Q \begin{pmatrix} 0 \\ kp_i^{k-1} \\ \vdots \\ 0 \end{pmatrix} \nonumber = \frac{\partial Q}{\partial p_l} Q^{-1} v_1 + q_i k p_i^{k-1}. \tag{21}
\]

where \( q_i \) stands for a vector identical to the \( i \)th column of \( Q \). The last term in (21) is a first-order polynomial in \( k \).
multiplied by an exponential sequence and can be written as a linear combination of \( v_1 \) and \( v_2 \):

\[ q_k p_k^{-1} = E_1^{(0)} v_1 + E_2^{(0)} v_2, \quad (22) \]

\( E_1^{(0)} \) and \( E_2^{(0)} \) being matrices. Since \( q_k \) does not depend on \( k \), both \( E_1^{(0)} \) and \( E_2^{(0)} \) have rank 1. Comparing (22) to (8) gives

\[
A_{11}^{(0)} = \frac{\partial Q}{\partial p_l} Q^{-1} + E_1^{(0)}, \quad (23)
\]

\[
A_{12}^{(0)} = E_2^{(0)}. \quad (24)
\]

Thus, all matrices \( A_{m,m-1}^{(0)} \) and \( A_{m,m+1}^{(0)} \) have rank 1 and can be written as a dyadic product.

### 4. Optimality Conditions

Since the vectors \( v_m(k) \) \((m = 1, 2, \ldots)\) constitute an orthonormal basis in \( l^2(Z_k) \), any causal square summable function \( h(k) \) can be expressed as

\[
h(k) = \sum_{m=1}^{\infty} c_m v_m(k). \quad (25)
\]

\( c_m \) is called the Kautz vector spectrum and can be calculated by an inner product

\[
c_m = \langle h, v_m \rangle. \quad (26)
\]

In the remainder only real-valued functions \( h(k) \) are considered, and thus the vector spectrum is real-valued as well.

Considered are truncated Kautz series according to

\[
h(k) \approx f_M(k) = \sum_{m=1}^{M} c_m v_m(k). \quad (27)
\]

Given a fixed number of terms \((M)\) we search for an optimality condition for the set of poles, i.e., by which poles do we obtain the ‘best’ possible representation in this finite series? The ‘best’ representation is defined as that which maximizes the energy in the approximating function \( f_M(k) \).

The energy \( E \) in \( h(k) \) is given by \( E = \langle h, h \rangle \) and the energy \( E_M \) in the approximating function \( f_m \) is given by \( E_M = \langle f_M, f_M \rangle = \sum_{m=1}^{M} c_m^2 c_m \). The last equality stems from the orthonormality relation. Optimality conditions are sought which apply to the set of poles that maximizes \( E_M \).

If a certain set of poles \((p_i)\) yields this maximum then

\[
\frac{\partial E_M}{\partial p_i} = 0 \quad (28)
\]

for each \( p_i \), assuming real-valued poles. The extension to complex-conjugated pole pairs can be easily established as well [8]. Naturally, (28) holds not just for the global maximum but for any other extremum of the energy function as well.

Taking the derivative of \( E_M \) with respect to \( p_i \) gives

\[
\frac{\partial E_M}{\partial p_i} = 2 \sum_{m=1}^{M} c_m \frac{\partial c_m}{\partial p_i} \quad (29)
\]

For \( \frac{\partial c_m}{\partial p_i} \) we have from (26) and (9)

\[
\frac{\partial c_m}{\partial p_i} = A_{m,m-1}^{(0)} c_{m-1} + A_{m,m+1}^{(0)} c_{m+1} + A_{m,m}^{(0)} c_m \quad (30)
\]

Substituting (30) in (29) and using (7) and (20) gives

\[
\frac{\partial E_M}{\partial p_i} = 2 c_m A_{M,M+1}^{(0)} c_{M+1} =
\]

\[
= 2 M c_m A_{M+1}^{(0)} c_{M+1} \quad (31)
\]

For real-valued poles this expression must be zero for each \( i \). Thus, if the poles are such that

\[
c_M = 0 \quad \text{or} \quad c_{M+1} = 0, \quad (32)
\]

then the energy function \( E_M \) of a Kautz series of \( M \) terms has an extremum for this particular set of poles. The same observation holds if complex-conjugated pole pairs are allowed [8]. All other solutions of (28) depend on \( Q \), e.g. the case of \( N = 2 \) and two real-valued poles [6]. It is clear that the condition (32) is the only solution of (28) if \( N = 1 \) (i.e., a Laguerre series [4-6]). In [6] the matrices \( A \) have been explicitly calculated for the case that \( N = 2 \). From these explicit expressions it can be concluded that if \( N = 2 \) and the poles \( p_1, p_2 \) constitute a complex-conjugated pair, (28) and (32) are equivalent.

### 5. Conclusion

The results presented in this paper can be easily adapted to the continuous-time Kautz-functions [1], since the analysis only uses the orthonormality and the polynomial character of the basis functions.

The derived optimality conditions were already forecast in [6]. The present study differs from [6] in that a more general result is obtained. However, [6] contains some specific results for the derivatives for the case \( N = 2 \) which can not be established within the context given here.

For a given function \( h(k) \) and specific values for \( N \) and \( M \), (31) can be used to establish the optimal poles numerically by a gradient-oriented search method. For a fixed number \( N \times M \), an increase in \( N \) results in more degrees of freedom and thus, in general, in a better approximation. On the other hand, an increase in \( N \) results in a more complicated optimization procedure. Furthermore, many functions can be associated with a single time-scale at which relevant changes occur. This implies that in many instances only a very limited number \( N \) of free poles suffices to obtain a good approximation.

### References


Nonlinear Channel Output Behaviour in the Presence of Error Correcting Codes
Abdelahad BENHALLAM
Ecole Nationale de l'Aviation Civile
7, avenue Edouard Belin, 31055 Toulouse cedex, France
E-mail: benhalla@eis.enac.dgac.fr

Abstract. It is already known that the Viterbi decoder output contains errors grouped in bursts when the signal-to-noise ratio becomes weak. Statistics of these bursts have been studied and modeled. Our interest here is in modeling the burst error process at the channel output. This should enable us to reduce execution time and implementation complexity, to characterize the evolution of the error distribution after Viterbi decoding and to design the encoder (code rate, code generator). Using digital computer simulations, we propose a statistical study of the nonlinear channel output symbols in an encoded satellite communication link. Results on the channel modeling are then presented. The usefulness of this modeling with regard to the decoder behaviour is analysed afterwards.

Keywords. Satellite communication, concatenated code, TWT nonlinearity, error burst.

1. The problem components
Considering an encoded digital communication scheme, this paper deals with statistical processing of both the channel and the decoder outputs. This should allow us to model the channel (or channel+codec), and draw up the bit error rate (BER) after decoding according to the error configuration at the channel output.

Three points constitute our purpose. First, such modeling would facilitate implementation software and reduce execution time (e.g., in a Monte Carlo simulation). Second, it is interesting to characterize the decoder performance in both BER and overflow probability (when buffers and/or stacks are used in the decoder structure). Finally, the desired encoded link performance may be obtained by a given code structure (rate, length and generator). Therefore, one can search for such a "good" code by observing error distribution at the channel or the decoder output (bearing in mind that a large number of erroneous symbols at the channel output does not necessarily conclude on "bad" encoding, since the decoding algorithm efficiency can achieve satisfactory performance).

The present study is made on both erroneous symbols packed in bursts- and correct symbols - intervals (also called waiting times) between bursts- at both the channel and the decoder outputs. Probability laws of bursts and intervals are depicted by cumulative distribution functions. In a previous paper [4], we made a similar study in the case of an uncoded communication scheme with a nonlinear channel.

The burst and interval distributions and the burst error density are analysed with respect to $E_b/N_0$ (energy per bit to noise monolateral spectral density) ratio, and code parameters.

In the next section, we define the channel and codec system components used in different simulated communication models. In section 3, digital simulation results on statistical channel output behaviour are presented. Sections 4 and 5 try to explain the effect of these error statistics on the decoder behaviour. Section 6 provides a synthesis of this work.

2. Simulated scheme models
Figure 1 shows the nonlinear model of the simulated encoded digital communication link. The coding system consists of a concatenated code where the convolutional (CV) code and the Reed-Solomon (RS) code are respectively the inner and the outer codes. Both RS and CV codes are characterized by the code generator, the code rate, denoted $r_1$ and $r_2$ resp., and the code length. Moreover, CV code complexity may be expressed through constraint length, denoted $K$, for a given rate.

Several sets of values of such parameters have been simulated. Some results of this paper are computed for the particular set of values approved by the Consultative Committee for Space Data Systems (CCSDS) : for RS code $r_1 = 23/255$, and for CV code $K=7$ and $r_2 = 1/2$. The global code rate is then $r_1 r_2 = 0.437$, and the signal-to-noise ratio (SNR) after coding is $E_b/N_0 = 3.59 \text{ dB}$.

In figure 1, the transmission filter and the repeater bandpass filters consist of Chebycheff filters, while the lowpass receive filter consists of a Butterworth filter. The repeater nonlinearity element is a travelling-wave tube (TWT) nonlinear amplifier. The additive noise (uplink component only in our case) is modeled by white gaussian noise. Error correcting code decoding consists of a Viterbi optimal algorithm for CV code [1] and a Berlekamp-Massey algorithm for RS code [2].
The "ideal" model of figure 2 may be taken as a reference for evaluating the nonlinearity effect. Results on the schemes of figures 1 and 2 are provided in sections 3 and 4. The nonlinear model block diagram of figure 3 allows us (in section 5) to analyse the error distribution effect on the sequential decoder stack overflow. More information about the sequential decoding suboptimal procedure may be found e.g. in [3].

3. Channel output behaviour
The error distribution behaviour at the channel output on an uncoded link has been analysed in [4]. To characterize the channel output error layout, and to keep homogeneous result comparisons between uncoded and encoded cases, we assumed in this previous paper that at the channel output, errors were arranged in "bursts" rather than randomly. Some of the result comparisons in the present paper are in fact based on burst and interval distributions obtained in [4]. Simulation results in the encoded case of figure 1 are presented hereafter, according to some pertinent parameters.

3.1 Constraint length effect
Cumulative distributions of bursts and intervals on the channel output are depicted on figure 4 for three values of the constraint length. As for the uncoded case, both distributions here tend towards exponential laws with a negative slope. An increase in \( K \) leads to a greater burst size distribution slope, while it does not alter (on average) that of the interval length distribution. In other words, for a given burst size \( S_b \), \( P(S \geq S_b) \) quickly increases (but less than exponentially) with the increase in \( K \). The burst size distribution slope \( \alpha_b \) therefore follows the same evolution as in the uncoded case (cf. [4]). Moreover, the interval length distribution slope \( \alpha_i \) keeps on average the same value for different \( K \) values \( (K \) is assumed < 10 when Viterbi decoding is used). This slope unicity can be useful for obtaining an information about a possible memory overflow of the sequential decoder stack. Denote \( K_o \) the particular constraint length corresponding to \( \alpha_b = \alpha_i \), the usefulness of \( K_o \) may be validated only when sequential decoder is used (figure 3) as is discussed in [4] (cf. section 5.).

3.2 Encoding effect
In section 2., we saw that the CCSDS concatenated code reduces the signal energy by 3.59 dB. The channel output statistical performance has consequently to be degraded in comparison with that of the uncoded case. However, the error distribution conserves the same behaviour. Figure 5 provides a statistical comparison, where the SNR has been augmented by 3.59 dB in the encoded case.

3.3 Eb/No effect
Figure 6 shows the distribution slope \( \alpha_b \) and \( \alpha_i \) variation according to some Eb/No values. A greater SNR value results in a lower \( \alpha_b \) and a higher \( \alpha_i \).
3.4 Nonlinearity effect and channel modeling

In figure 7, we compare statistics relating to four different channel (TWT) nonlinearities. For a given Eb/No ratio, these different TWTs lead to similar burst (resp. interval) statistics. Thus, for a given couple (K,Eb/No), we nearly obtain the same slope $\alpha_B$ for burst size (resp. $\alpha_I$ for interval length) distribution. For example, when K=7, measured slopes are about $\alpha_B \approx -0.076$ and $\alpha_I \approx -0.058$ for Eb/No=2.78 dB, and $\alpha_B \approx -0.114$ and $\alpha_I \approx -0.033$ for Eb/No=4.78 dB.

Therefore, on the basis of four used TWT nonlinearities, we can deduce (as already done for the uncoded case in [4]) the independence of burst and interval distribution from the TWT nonlinearity kind. Thus, the nonlinear channel showed on figure 1 can be modeled by an exponential law "noise generator". Such modeling has been validated by a $\chi^2_{(\nu)}$ hypothesis test.

In this way, replacing the nonlinear channel of figure 1 by a simple noise generator allows us to reduce the execution time and particularly the implementation complexity.

The consideration of processing errors at the channel output as packed in bursts (rather than isolated errors) requires characterizing the new noise generator, for a given couple (K,Eb/No), with the slope $\alpha_B$, the average interval length $L_o$, and the error density $d_o$. We can then compose the error sequence as follows. Using an exponential distribution of type $y_\alpha \cdot 10^{\alpha_B (x-x_0)}$, where $(x_0, y_\alpha)$ is a point of the distribution, $\alpha_B$ provides the burst number for a given size $S_o$. The error number in each burst is computed by $d_o$ which can be generated by a truncated normal distribution. Finally, intervals (or waiting times) are generated with a geometrically distributed variable, taking into account that an interval length must be greater than (K-2). Simulations on the model of figure 8 result in a bit error rate close to that obtained by the scheme of figure 1. For Eb/No=2 dB, obtained BER values on only one software realization are: $BER_{f_8}=0.328$ and $BER_{f_8}=0.331$.

3.5. Non-correlation of bursts

The non-correlation between burst sizes is being verified to within an 8% estimation error (figure 9), the effect of a given burst on the decoder behaviour is not altered for any burst position. Each burst is in fact followed by an interval whose length is greater than or equal to (K-1), that is a sufficient waiting time allowing the decoder to retrieve the "correct" path (a path providing an error rate lower than or equal, in average, to that of the path left when the previous burst has occurred). This fact justifies the choice of K as a construction parameter of bursts and intervals.

4. Viterbi decoder output behaviour

The burst error distribution on the Viterbi decoder output has been analyzed and modeled in [5], to directly evaluate BER at the decoder output. In this section, a brief comparison between distributions at the channel output and the decoder output is made. Results on the decoder output modeling are then presented.

The fact that the Viterbi decoder groups errors in bursts is independent of the channel structure. This is confirmed by the following results. Figure 10 shows the burst size and the interval length distributions on the Viterbi decoder output, for different couples (K,Eb/No).
Comparing figures 10 and 6 for \(Eb/No=5\) dB and \(K=7\), we can easily note the obvious appearance of "true" error bursts on the decoder output. Moreover, unlike the channel output case (fig.5), the interval length distribution at the decoder output does not present a unique slope value for different constraint lengths. This can be considered as a result of the error correction performed by the decoder. Further parameter comparisons are provided in table 1.

<table>
<thead>
<tr>
<th>65000 bits</th>
<th>Channel output</th>
<th>Viterbi decoder output</th>
</tr>
</thead>
<tbody>
<tr>
<td>(K=7, Eb/No=3.5) dB</td>
<td>(S_{max})</td>
<td>55</td>
</tr>
<tr>
<td>(L_{max})</td>
<td>85</td>
<td>451</td>
</tr>
<tr>
<td>(S)</td>
<td>4.13</td>
<td>23.28</td>
</tr>
<tr>
<td>(L)</td>
<td>14.41</td>
<td>80.64</td>
</tr>
<tr>
<td>Burst number</td>
<td>3506</td>
<td>624</td>
</tr>
<tr>
<td>Interval number</td>
<td>3507</td>
<td>625</td>
</tr>
<tr>
<td>Average density</td>
<td>.755</td>
<td>.634</td>
</tr>
</tbody>
</table>

Table 1. Statistic comparisons

Considering the Viterbi output decoder modeling, figure 11 shows that burst size distribution is easier to model than the one taken from the channel output (figures 5 and 7).

Both burst size and interval length distributions can then be modeled with nearly geometrical distributions [5]. In this manner, one can replace the link model of figure 1 by that of figure 12, where the "burst noise generator" box encloses, in this case, the convolutional encoder and the Viterbi decoder.

This modeling results in a BER value close to the one obtained with the transmission link of figure 1 : for example, for \(Eb/No=3.8\) dB, \(BER_{fig1}=.0037\) and \(BER_{fig12}=.0038\).

5. Sequential decoder behaviour

Our discussion in this section concerns the transmission model of figure 3.

In section 3, we found that the increase in constraint length \(K\) results in an increase of the burst size distribution slope, while it does not alter the interval length distribution slope. Assume that similar results have been obtained on a given set of constraint length values, when using a sequential decoder (figure 3). Let \(K_e\) be the particular value held by \(K\) when \(\alpha_s = \alpha_r\). It is easy to see that \(K_e\) augments when \(Eb/No\) ratio increases.

Otherwise, as mentioned in [4], the increase in burst size (also in constraint length) leads the sequential decoder to carry out a larger quantity of computations per decoded bit. This can cause a decoder memory (stack or input buffer) overflow.

Thus, it becomes possible to connect together the channel output burst error statistics view point and the sequential decoder overflow problem. That is, for a given \(Eb/No\) ratio, we can find a particular constraint length \(K_e\) (providing \(\alpha_s = \alpha_r\)) that must not be exceeded to ensure an acceptable overflow probability.

For a selected stack size and a \(Eb/No\) ratio, one can set a maximum tolerable overflow probability, according to the particular constraint length \(K_e\). It is thus possible to retain nearly the same overflow probability threshold whatever \(Eb/No\), provided that the constraint length takes the \(K_e\) value. This fact represents one of the consequences of the channel output analysis.

6. Conclusion

The error statistics behaviour on a nonlinear model of a practical satellite communication link is analysed and modeled. Using digital simulations, this work allowed us to emphasize the importance of the statistical characterization of the channel output, on both channel modeling and a possible prediction of the decoder behaviour.

Considering that errors on the channel output are packed in bursts permitted to lead a statistical comparison between the channel and the Viterbi decoder outputs. It also provided an evaluation of encoding, noise, and nonlinearity effects on the error configuration before and after decoding.

References


Some Weighted Objective Approaches For Sparse Deconvolution

Ignacio SANTAMARIA-CABALLERO† Aníbal R. FIGUEIRAS-VIDAL‡

†Dpto. Electrónica, ETSI Telecom-Univ. Cantabria Av. Los Castros s.n., 39005 Santander, SPAIN.
‡GPSS/DSSR, ETSI Telecom - UPM Ciudad Universitaria s.n., 28040 Madrid, SPAIN
Tel: 34-1-336-7286, Fax:34-1-336-7250
e-mail: ignacio@gtls.ssr.upm.es

Abstract. In this paper we present two new methods to perform sparse deconvolution. Basically, they use a gradient-type algorithm to minimize an objective function which consists on a weighted sum of the L2-norm of the residual and the L1-norm of the signal. Statistical information about the signal and noise can be efficiently used in our methods; for example, to derive a convergence criterion or to modify the weighted cost function. After discussing these methods, their effectiveness and robustness are illustrated by means of computer simulations using synthetic data.

1. Introduction

In many practical problems (geophysical signal processing, speech coding, synthetic aperture radar, nondestructive testing, etc.), it is often necessary to find a sparse solution to

$$z = Hx + n$$ (1)

where $H$ is an impulse matrix, $x$ the unknown signal vector, $n$ the noise, and $z$ the observations. The signal $x$ is known to be sparse, i.e., it contains many zeroes plus several comparatively large nonzero values.

A direct application of standard methods to solve (1), such as using the Moore-Penrose pseudoinverse, is not appropriate, since the ill-conditioned character of the problem prevents the obtention of the sparse solutions we are looking for. In addition, theoretical solutions to the corresponding detection plus estimation problem are cumbersome; those available, such as [1], require that the peaks of $x$ have a Gaussian distribution, a hypothesis which is not acceptable in many situations. Moreover, sometimes signal statistics are not available, so a complete analytical solution is not possible.

To overcome these limitations, many alternatives have appeared in the literature: some of them [2] combine detection and estimation tasks using an adaptively contracted selection operator. This technique is computationally attractive, but is very sensitive to the selection of the parameters involved in the method and, more important, sometimes misses small peaks in the first steps of the detection process; an irreversible mistake.

An alternative approach consists on adding a regularizing term to reduce the ill-conditioned character of the problem, similar to the method proposed by Katsaggelos et al. [3], to restore noisy-blurred images. Nevertheless, it uses a constraint operator which produces a smooth solution not adequate for our problem.

Finally, other approaches based on using an L1-norm minimization algorithm have been proposed. In particular, the method presented in [4] performs the L1-norm minimization of the residuals by means of linear programming. It was not proposed for its ability to recover sparse signals, but rather for its ability to deal with large data errors. A more appropriate approach consists of including the L1-norm of the signal in a weighted objective function [5, 6]; this approach is well suited for data driven from spiky distributions. However, it has two drawbacks: the high computational cost when the data set is large, and the sparse character of the resultant residual, which does not agree with the usual type of added noise.

In this paper we propose some new methods to avoid the problems of the previously mentioned algorithms. Specifically, we present a new objective function that consists on a weighted sum of the L2-norm of the residual and the L1-norm of the signal. Some statistical information about the signal and noise can be efficiently used in this approach to stop the algorithm. Following similar ideas, another weighted function is proposed: it includes the a priori statistical knowledge in the objective function. In the rest of the paper we discuss these approaches and illustrate their performances by means of computer simulations using synthetic data.

2. Proposed Algorithm

The first method studied in this paper consists on finding the solution $x$ that minimizes the following cost function

$$\Phi_1(x, \alpha) = (1 - \alpha)\|z - Hx\|_2 + \alpha\|x\|_1, \quad 0 \leq \alpha \leq 1$$ (2)

By doing this, we seek a minimum L1-norm solution while preserving a small L2-norm residual. The weighting parameter $\alpha$ controls the spiky character of $x$. The minimization of (2) can be accomplished by means of a gradient-type algorithm (with $\alpha$ fixed)

$$x_{k+1} = x_k + \mu(1 - \alpha)H^T(z - Hx_k) - \mu\text{sign}(x_k)$$ (3)

where the superscript $T$ denotes transpose. The iterations are carried out until a previously specified error criterion
is satisfied or a maximum number of iterations is reached. Note that parameter $\alpha$ controls the sparse character of $x$: for $\alpha=0$, the pseudoinverse solution of (1) is obtained and, as $\alpha$ is increased, sparser solutions are obtained; finally, when $\alpha=1$, the algorithm (3) converges to the trivial solution $x=0$.

On the other hand, for $\alpha=0$, $\mu$ must be less than twice the inverse of the maximum eigenvalue of $HH^T$ in order to guarantee convergence. In general, convergence depends on both $\mu$ and $\alpha$. Since $\alpha$ is a bounded parameter ($0 \leq \alpha \leq 1$), we can choose an empirical $\mu$ to ensure convergence for any $\alpha$.

A complete application of the method requires a procedure to select an optimum $\alpha$ which leads to a feasible solution. Without additional information we must choose $\alpha$ empirically. Nevertheless, as we will show later, $\alpha$ is not a critical parameter; so it can be fixed a priori for a given problem obtaining good results in a wide variety of examples. More accurate results can be obtained if we assume that some (very general) a priori knowledge about the original signal and noise is available (for example coming from a reasonable model). We will assume that estimates of the $L_2$-norm of the noise ($\hat{N}_2$) and the $L_2$-norm of the signal ($\hat{S}_1$) are available (the true values will be denoted without the symbol $\hat{}$). The estimate of the noise variance has been used in several signal deconvolution and image restoration problems [7, 8] to derive convergence criteria, to choose the regularization parameter, or to find a projection operator onto a convex set such as $C_{\alpha_2} = \{ x \mid ||z - Hx||_2 \leq \hat{N}_2 \}$. In [8] it is noted that many other constraints can be applied to the deconvolution problem. They depend upon the characteristics of the specific signal; in particular, when we are dealing with a sparse signal, to use a similar constraint over the $L_2$-norm of the signal has proven to be useful.

The proposed method starts selecting an $\alpha_{\text{max}}$ that produces a solution sparse enough. Then, the weighting parameter is iteratively decreased in fixed steps $\Delta \alpha$ until some coverage criterion is fulfilled. The criterion presented in this paper is based on the derivative of the objective function with respect to $\alpha$: each $\alpha$ yields a solution $x^\alpha$ which minimizes (2), the value of the objective function for this solution is $\Phi_1(x^\alpha, \alpha)$. Assuming that the solution $x^\alpha$ does not depend on $\alpha$, we can write

$$\frac{d\Phi_1(x^\alpha, \alpha)}{d\alpha} = -||z - Hx^\alpha||_2 + ||x^\alpha||_2,$$ (4)

when $x^\alpha$ is similar to the true series, the term $||z - Hx^\alpha||_2$ approaches the $L_2$-norm of the noise and $||x^\alpha||_2$ the $L_2$-norm of the signal. Consequently, our knowledge about $\hat{N}_2$ and $\hat{S}_1$ can be used to stop the algorithm when a near optimum $\alpha$ has been reached, i.e., we know in advance an estimate of the derivative (4) for the true signal

$$\frac{d\Phi_1(x^\alpha, \alpha)}{d\alpha} = \hat{S}_1 - \hat{N}_2,$$ (5)

thus, the final solution $x^\alpha$, and the final weighting parameter $\alpha_{\text{opt}}$, must satisfy the following criterion

$$\left| \frac{d\Phi_1(x^\alpha, \alpha_{\text{opt}})}{d\alpha} - \frac{d\Phi_1(x, \alpha)}{d\alpha} \right| < \delta,$$ (6)

where $\delta$ is a small positive constant.

It is also possible to use similar convergence criteria. For example, using only information about the $L_1$-norm of the signal: $||x^\alpha||_1 - S_1 < \delta$, or the $L_2$-norm of the noise: $||z - Hx^\alpha||_2 - \hat{N}_2 < \delta$. The three criteria provide similar performance, as it will be shown in Section 4. Summarizing, the proposed algorithm is as follows

1. Initialize $x_0 = 0_{N \times 1}$, $j=0$, $\alpha_0 = \alpha_{\text{max}}$

2. Estimate $\frac{d\Phi_1(x, \alpha)}{d\alpha} = \hat{S}_1 - \hat{N}_2$

3. for $k=0$ to $N-1$
   3.1 $x_{k+1} = x_k + \mu(1-\alpha_k)H^T(z - Hx_k) - \mu \alpha_k \text{sign}(x_k)$
   3.2 if $|\Phi_1(x_{k+1}, \alpha_j) - \Phi_1(x_k, \alpha_j)| < \epsilon$ go to 4
   end
   4. $x^\alpha = x_{k+1}$
   5. $\frac{d\Phi_1(x^\alpha, \alpha)}{d\alpha}(j+1) = \alpha_j - \mu ||z - Hx^\alpha||_2 + ||x^\alpha||_2$
   6. if $|\frac{d\Phi_1(x^\alpha, \alpha)}{d\alpha}(j+1) - \frac{d\Phi_1(x^\alpha, \alpha)}{d\alpha}(j)| < \delta$ then stop
      else $\alpha_j+1 = \alpha_j - \Delta \alpha$
      if $j = 0$ go to 3

Finally, to obtain a fully sparse signal we must apply a thresholding procedure. The method selected proceeds in three steps: first, a conservative threshold is applied; second, the amplitudes of the surviving peaks are adjusted in order to minimize the quadratic error (this step increases the magnitude of the true peaks and, conversely, it decreases the false ones); finally, a new threshold is applied that eliminates the smaller spikes. The selection of the final threshold depends on the problem. For instance, in a multipulse coding application it could be selected to obtain a fixed number of spikes in the final solution.

3. An Alternative Approach

To minimize the $L_1$-norm of the signal tends to underestimate the amplitudes of the true spikes. This effect can be reduced if we include our knowledge about $S_1$ and $N_2$ in the objective function. For example, according to

$$\Phi_2(x, \alpha) = (1-\alpha)||z - Hx||_2 - N_2^\alpha + \alpha||x||_1 - S_1||^g$$ (7)

with $0 \leq \alpha \leq 1$ and $1 \leq p, q$. This second approach has an additional advantage: it avoids the need to obtain the optimum $\alpha$, since in this case the objective function by itself forces the vector $x$ to be in a neighborhood of the true solution. Therefore, an empirical $\alpha$ can be fixed in advance achieving good results for a great number of examples. For instance, $\alpha = 0.5$ is the most obvious selection, thus giving the same importance to each term in the objective function. The possibility is to change successively between $\alpha = 0$ and $\alpha = 1$. In this way, an alternative adjustment between the residual $L_2$-norm and the signal $L_1$-norm can be made. However, the simulations indicate a superior performance of the former approach ($\alpha$ fixed) . Therefore we will consider only this procedure.

In addition to choose a weighting parameter we must
select the values $p, q$ of the objective function. Experimental evidence shows that $p=q=1$ provide the best results. The minimization of (7) is again accomplished by means of a gradient-type algorithm. As previously, for $\alpha$ fixed the parameter of the gradient algorithm $\mu$ controls the convergence. A conservative selection is needed with this approach to avoid convergence problems.

4. Simulation Results

We have selected two computer experiments with different sparse signals. The first uses a deterministic signal, the second uses randomly generated sparse signals according to a preestablished model. Specifically, we generate sparse signals with Gaussian or uniform amplitude distributions.

4.1. Experiment 1

The impulse response used in this example corresponds to the first 20 points of an ARMA filter having a zero at $z=0.6$ and two poles at $z=0.8$ \exp($\pm j5\pi/12$). The test signal $x$ is a 110 points register having nonzero values at points $x(20)=8$, $x(25)=6.845$, $x(47)=-5.4$, $x(71)=4$ and $x(95)=-3.6$. The SNR used in this example is 4 dB, and is defined as the power of $Hx$ with respect to the power of $\nu$; $\nu$ being a zero mean Gaussian white noise.

We compare the performance of the algorithms corresponding to:

A1) Cost function $\Phi_1(\hat{x}, \alpha)$ with an optimum weighting parameter chosen to fulfill (6).

A2) Cost function $\Phi_2(\hat{x}, \alpha)$ with a fixed weighting parameter.

A3) Using an adaptive threshold [2].

For the three methods we ensure convergence by selecting $\mu=0.1$. We apply the first method with $\alpha_0=0.8$, and then update this parameter in fixed steps of $\Delta \alpha=0.01$ until (6) is fulfilled. For each $\alpha$, a maximum number of 50 iterations of (3) is carried out. For the second method we have empirically selected an optimum $\alpha=0.5$ and a maximum number of iterations of 200. With respect to the selection of $\alpha$ we must remark that it is not a critical parameter: values in a neighborhood of the optimum $\alpha$ give visually satisfactory solutions.

Figure 1 shows the result obtained for one simulation applying the first algorithm after thresholding.

![Figure 1. Solution obtained with the first algorithm after thresholding. Circles depict true spikes.](image)

Table 1. Comparison of the results from the three algorithms tested in Example 1. The first row indicates the method used, the next five rows show the detection percentage for each spike, and the last row shows the averaged number of false alarms detected in each simulation.

<table>
<thead>
<tr>
<th>spike 1</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>spike 2</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>spike 3</td>
<td>99</td>
<td>98</td>
<td>96</td>
</tr>
<tr>
<td>spike 4</td>
<td>94</td>
<td>89</td>
<td>86</td>
</tr>
<tr>
<td>spike 5</td>
<td>84</td>
<td>62</td>
<td>67</td>
</tr>
<tr>
<td>average of spurious spikes</td>
<td>0.71</td>
<td>0.61</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table 1 shows the averaged results for one hundred independent simulations. Both methods have similar detection capabilities and outperform the adaptive threshold procedure but increase the number of spurious spikes.

We have tested the first algorithm applying other convergence criteria. Using only the $\hat{S}_1$ estimate or the $\hat{N}_2$ estimate, all the criteria achieve similar results. This can be explained noting that if for a solution $x^*$, $||x^*||_1$ is close to $\hat{S}_1$, then, $||x-Hx^*||_2$ will be near to $\hat{N}_2$.

Both methods use estimates of $\hat{N}_2$ and $\hat{S}_1$, therefore it is important to study their robustness against errors in the estimates. To do so, we have repeated the same one hundred simulations but in this case we introduced random errors in the estimates; i.e., we considered that $\hat{S}_1$ and $\hat{N}_2$ were two random variables uniformly distributed in the intervals $[S_1(1-c_1), S_1(1+c_1)]$ and $[N_2(1-c_2), N_2(1+c_2)]$ with $0 \leq c_1, c_2 \leq 1$. To simplify the results, we assume the same error percentage for both estimates, i.e., $c_1 = c_2 = c$. Figure 2 shows the worsening in detection percentage for the first algorithm and $c$ varying from 0 to 1. Besides, the average number of spurious spikes remains nearly constant. So we can conclude that the proposed method is remarkably robust. On the other hand, the minimization of $\Phi_2(\hat{x}, \alpha)$ is less robust against errors in $\hat{S}_1$ and $\hat{N}_1$, this is understandable because in the first approach the errors only affect the convergence criterion, while in the second approach they affect the objective function, thus modifying the algorithm itself.

4.2. Experiment 2

In this example we evaluate the performance of our algorithms using synthetic signals according to the following model: $x(k)=r(k)q(k)$, where $q(k)$ is a Bernoulli process for which $q(k)=1$ with probability $\lambda$ and $q(k)=0$ with probability $1-\lambda$; $r(k)$ is a white random process with zero mean, variance $\sigma^2$ and whose amplitudes fit a Gaussian or uniform distribution (in particular, the Gaussian distribution is often used for seismic deconvolution cases). Registers of five hundred samples were generated according to the above models (with $\lambda = 0.05$ and $\sigma^2 = 10$), and then convolved with the source wavelet described in Example 1. Finally, a zero mean Gaussian noise was added to the result.

For this example, the simulations compare the performance of the algorithms corresponding to:

B1) Cost function $\Phi_1(\hat{x}, \alpha)$ with an optimum weighting parameter chosen to fulfill (6).
B2) Cost function $\Phi_3(\hat{x}, \alpha)$ with a fixed weighting parameter.

B3) One-shot threshold detector [1].

Figure 3 shows the results obtained with the first method before thresholding. The signal has a Gaussian amplitude distribution and the SNR = 8 dB.

Table 2 shows the averaged results of 25 simulations when there is a Gaussian (Table 2.a) or a uniform (Table 2.b) amplitude distribution. The SNR for this example is 4 dB. Somewhat surprisingly, the three algorithms give better results for a uniform amplitude distribution of the sparse signal. However, this can be easily explained since for a fixed variance, data driven from a Gaussian distribution are near zero (and, therefore, are more difficult to detect) with higher probability than if they were driven from a uniform distribution. It is clear that the proposed methods outperform the one-shot threshold detector.

5. Conclusions

In this paper we have presented two new methods to recover a sparse signal from a noisy register. They use a gradient-type algorithm and incorporate statistical information about the signal and noise. The methods are computationally simple and efficient, and they achieve good performance when applied to a wide variety of examples. The first method (which uses the statistical information to derive a convergence criterion to obtain the optimum weighting parameter) is more robust than the second (which includes the statistical information in the objective function), but it is slightly more involved.

Open study lines in this weighted objective approach are: to include other regularized functionals and other approaches to adjust the regularizing parameter; as well as to modify these methods to apply them to the (dual) problem of sinusoid detection.

References


Isomorphism between Continuous-time System and Discrete-time System based on Differentiability of the Input Signal

Takahiko HORIUCHI† and Kazuo TORAICHI‡

† Doctoral Program in Engineering, University of Tsukuba, Tsukuba, Ibaraki, 305 JAPAN
Tel / Fax: +81[298]53 5052 / 53 5206, E-mail: tah@uslab.is.tsukuba.ac.jp
‡ Inst. Info. Sci. Elec., University of Tsukuba, Tsukuba, Ibaraki, 305 JAPAN
Tel / Fax: +81[298]53 5052 / 53 5206, E-mail: toraichi@uslab.is.tsukuba.ac.jp

Abstract. In order to realize a continuous-time system model in digital computers, we must construct a discrete-time system model simulating the continuous-time process in some characteristic aspect. Though many discretization methods have been proposed, they do not necessarily provide a discrete-time system in which input, state and output are identical with the sampled values of the original continuous-time system model. This paper aims at guaranteeing the isomorphism between a continuous- and a discrete-time systems with the input of piecewise polynomials (fluency system model), which were proposed by one of the authors. The isomorphism of input space had been already shown in the previous works by one of the authors. In this paper, by showing the isomorphism of the state function and output spaces, the aim will be achieved.

1 Introduction

Models of continuous-time systems have been effectively used for analysis of physical phenomena in the nature. Representation of such a continuous-time system model in digital computers demands construction of a discrete-time system model simulating the continuous-time processes in some characteristic aspect. Many discretization methods such as bilinear transformation method and time-response invariance method have been proposed [1]; however, they do not necessarily provide a discrete-time system in which the state and output as well as input are identical with the sampled values of the original continuous-time system. A discretization method yielding the identical sampled values [2], which is valid for some identification problems, can resume continuous-time signals solely at the sampling times from discrete-time signals. The isomorphic discretization that all of the input, state and output of a continuous-time system can be restored from the corresponding discrete-time system is crucial for our analysis.

Our study on the relation between continuous- and discrete-time system models led us to attempt to prove that any continuous-time system can be assigned to its isomorphic discrete-time system with appropriate approximation of the input. It is described previously that for a discrete-time system model an approximate signal space is to be specified by the signal to be handled. We have proposed a discrete-time system model called “fluency system model” [3, 4] for which piecewise polynomial functions [5] are used for the input. It is also demonstrated [6] that the isomorphism of the fluency system model is established by deriving the sampling bases for the input/output signal spaces and the state function space. The isomorphism for the input space was already derived [7]. This paper aims to guarantee the isomorphism of the fluency system model by clarifying the existence of sampling basis in the state function and output spaces as shown in Fig.1. The existence of the sampling basis is shown with definition of a coordinate transformation operator from the sampling basis to an expansion basis. The sampling basis is expressed in the form of linear combinations of the expansion basis.

2 Preliminaries

This section presents preliminaries for piecewise polynomial signal spaces which we deal with in this paper. Then, we prepare the state function space of a linear system when the input is given by a piecewise polynomial signal.
2.1 Piecewise polynomial signal space and its characteristics

In general, the set of all continuous time functions, which is the object of signal processing, can be considered, by defining the inner product

\[
(u, v)_{L_2} \triangleq \int_{-\infty}^{\infty} u(t)\overline{v(t)}dt,
\]

as the following typical Hilbert space

\[
L_2(R) \triangleq \left\{ u \mid \int_{-\infty}^{\infty} |u(t)|^2dt < +\infty \right\},
\]

where \( R \) is the set of all real numbers.

Following this definition, the signal space composed of piecewise polynomial functions of degree \((m - 1)\) is considered as a subspace of \(L_2(R)\). It is called the piecewise polynomial signal space of \((m - 1)\)-th degree, and defined as follows.

**Definition 1 (Piecewise polynomial signal space)**

Let the sampling points on the time axis be \( \{ k h \}_{k=-\infty}^{\infty}, (k h \triangleq kh, k = 0, \pm 1, \pm 2, \ldots) \). The piecewise polynomial signal space \( S^{(m)} \), which is isomorphic to the sequence space, is composed of the sample values at the above sampling points. By the system of \((m - 1)\)-th order B-spline functions

\[
\psi^{(m)}_\ell(t) \triangleq \int_{-\infty}^{\infty} \left[ \frac{\sin \pi f h}{\pi f h} \right]^m e^{j2\pi f(t-th)\frac{ mh}{2}} df,
\]

\( \ell = 0, \pm 1, \pm 2, \ldots; m = 1, 2, 3, \ldots \)

as the basis, it is defined as \([8]\)

\[
S^{(m)} \triangleq \left[ \psi^{(m)}_\ell \right]_{\ell=-\infty}^{\infty}.
\]

We shall call \( S^{(m)} \) piecewise polynomial signal space of degree \((m - 1)\).

The B-spline basis satisfies

\[
u(t) = \sum_{k=-\infty}^{\infty} \lambda(k)\psi^{(m)}_k(t), \quad \lambda = \ldots, \lambda[-1], \lambda[0], \lambda[1], \ldots^T,
\]

for any signal \( u \) in the piecewise polynomial signal space. Let \( \varphi_{S^{(m)}} \) be the mapping which transforms \( u \) into \( \lambda \). Then \( \varphi_{S^{(m)}} \) satisfies

\[
\lambda = \varphi_{S^{(m)}}(u).
\]

Next, we prepare the well known characteristics of B-spline basis.

**Property 1** The system of piecewise polynomial functions can be represented as \((m - 1)\)-th order piecewise polynomial, which is \((m - 2)\) times continuous differentiable.

\[
\psi^{(m)}_\ell(t) = h^{-m} \sum_{p=0}^{m} (-1)^p (t - (\ell + p)h)^{\ell-p-1} / p!(m-p)!, \quad \ell = 0, \pm 1, \pm 2, \ldots, \quad (7)
\]

\[
(t-a)_+^{m-1} = \begin{cases} (t-a)^{m-1}, & t > a \\ 0, & t \leq a. \end{cases} \quad (8)
\]

The B-spline basis have the property as follows:

\[
\psi^{(m)}_\ell(t) = 0, \quad t \notin (\ell h, (\ell + m)h), \quad \ell = 0, \pm 1, \pm 2, \ldots \quad (9)
\]

And by Eq.(3) the B-spline basis have also the following property.

**Property 2**

\[
\psi^{(m)}_\ell(t) = \psi^{(m)}_\ell(t-\ell h), \quad \ell = 0, \pm 1, \pm 2, \ldots \quad (10)
\]

\[
\psi^{(m)}_\ell(t) = \psi^{(m)}_\ell((2\ell + m)h - t), \quad \ell = 0, \pm 1, \pm 2, \ldots \quad (11)
\]

This means that B-spline basis \( \psi^{(m)}_\ell \) has the shift-invariant property and symmetric with respect to the middle point. \( t = (2\ell + m)h)/2 \).

2.2 State function space

We shall deal with the linear system \( S \)

\[
\begin{align*}
\dot{z}(t) &= Fz(t) + Gu(t), \quad z(0) = 0, \\
\dot{y}(t) &= Hz(t)
\end{align*}
\]

\( F \in R, \ G \in R, \ H \in R, \ u(\cdot) \in S^{(m)}. \) (12)

In this paper, we consider that the system \( S \) has the property of \( L_2 \)-input and \( L_2 \)-output stability. Then we define the state function space as follows.

**Definition 2 (State function space)**

Let \( \gamma(t) \) be the impulse response from input \( u \) to state \( z \), i.e.

\[
\gamma(t) = \begin{cases} e^{Pt}G, & t \geq 0 \\ 0, & t < 0 \end{cases}
\]

Then the state function space \( K^{(m)} \) composed of all possible \( z \) is defined as follows:

\[
K^{(m)} \triangleq \left\{ z(\cdot) \in L_2(R) \mid z(t) = \int_{0}^{t} \gamma(t-\tau)u(\tau)d\tau, \right. \\
\left. \quad z(0) = 0, \quad u(\cdot) \in S^{(m)} \right\}.
\]

State function space is a family of functions of time onto the state space.

By Eq.(12), we can find that the signal space which include output \( y \) is also \( K^{(m)} \). Therefore, we discuss the existence of sampling basis in the state function space only.
3 Isomorphism between continuous- and discrete-time systems

In this section, we prove the isomorphism between continuous-time system and the fluency discrete-time system by showing the existence of a sampling basis which characterize the state function space defined in the previous section.

We shall define the coordinate systems which are to be used to show the existence of the sampling basis. Then the existence will be shown by the following steps. First of all, by showing the existence of basis \( \{\xi_{\ell}^{(m)}\}_{\ell=0}^{\infty} \) in the state function space \( K^{(m)} \), we show that the mapping \( \varphi_{K^{(m)}} \) from \( x \) into expansion coefficients \( \lambda \) is a biauniquely and biaubounded linear mapping. Then we derive that the existence of the sampling basis \( \{\eta_{\mu}^{(m)}\}_{\mu=-\infty}^{\infty} \) for \( K^{(m)} \) is equivalent to the biuniqueness and biauboundedness of \( \varphi_{K^{(m)}} \) which transforms \( x \) into sampled values \( \mu \). Then we introduce the operator \( M^{(m)} \) which transforms \( \mu \) into \( \lambda \), and show that the existence of the sampling basis is equivalent to the boundedness of \( M^{(m)} \). Finally, by showing the boundedness of \( M^{(m)} \), the existence of the sampling basis is proved. The mutual relations between coordinate systems are illustrated in Fig. 2.

3.1 Coordinate systems in state function space and the relation to the sampling basis

Before showing the existence of the sampling basis, we have to show that there is a basis in the state function space. In general, discrete-time signals are considered to constitute the Hilbert space

\[
\ell_2 \triangleq \left\{ \mu \mid \sum_{k=-\infty}^{\infty} |\mu[k]|^2 < +\infty \right\}
\]

with the inner product

\[
(\mu, \nu)_{\ell_2} \triangleq \sum_{k=-\infty}^{\infty} \mu[k]\nu[k].
\]

Lemma 1 The function system \( \{\xi_{\ell}^{(m)}\}_{\ell=-\infty}^{\infty} \) is expressed by

\[
\xi_{\ell}^{(m)}(t) \triangleq \int_{0}^{t} e^{(t-h)\lambda} G_{\psi_0^{(m)}}(t-l\lambda-h\tau) d\tau,
\]

\( \ell = 0, \pm 1, \pm 2, \cdots \)

is in the state function space \( K^{(m)} \).

Proposition 1 The function system \( \{\xi_{\ell}^{(m)}\}_{\ell=-\infty}^{\infty} \) is a basis for the state function space \( K^{(m)} \).

Let \( \varphi_{K^{(m)}} : K^{(m)} \to \ell_2 \) be the mapping which transforms \( x \) into \( \lambda \). Proposition 1 means that \( \varphi_{K^{(m)}} \) and its inverse are unique and bounded linear mappings. Next, we define coordinate system corresponding to the sampling basis in \( K^{(m)} \).

Definition 3 (Sampling basis in \( K^{(m)} \))

For \( K^{(m)} \), the sampling basis is defined as a system of functions \( \{\eta_{\mu}^{(m)}\}_{\mu=-\infty}^{\infty} \subset K^{(m)} \) satisfying

\[
\forall x \in K^{(m)}, \ x = \sum_{\mu=-\infty}^{\infty} \mu[\mu]\eta_{\mu}^{(m)}, \ \mu[\mu] = x(t_\mu).
\]

Then we can define the mapping which transforms \( x \) into \( \mu \).

Definition 4 (Coordinate system corresponding to the sampling basis for \( K^{(m)} \))

\( \varphi_{K^{(m)}} : K^{(m)} \to \ell_2 \) is defined as the mapping which transforms \( x \) into \( \lambda \).

Then the following proposition can be derived.

Proposition 2 \( \varphi_{K^{(m)}} \) is a unique and bounded linear mapping.

The following Lemma presents the existence of the sampling basis for the state function space \( K^{(m)} \).

Lemma 2 The sampling basis \( \{\eta_{\mu}^{(m)}\}_{\mu=-\infty}^{\infty} \) exists and is expressed by

\[
\eta_{\mu}^{(m)} = \varphi_{K^{(m)}}^{-1}(\delta_{\mu}), \ \mu = 0, \pm 1, \pm 2, \cdots
\]

where

\[
\delta_{\mu}[\ell] = \begin{cases} 1, & \ell = \mu \\ 0, & \ell \neq \mu \end{cases}
\]

if and only if \( \varphi_{K^{(m)}}^{-1} \) and its inverse are unique and bounded.

In order to prove the existence of the above sampling basis, we define the operator which transforms \( \lambda \) into \( \mu \).
Definition 5 (Coordinate transform operator)

\[ M^{(m)} = \text{the coordinate transform operator which is defined by} \]

\[ M^{(m)} \triangleq \varphi_{[K^{(m)}]}^{-1} \varphi_{[K^{(m)}]}^{-1}. \]  

(21)

Figure 2 yields \( M^{(m)} \) -1 is represented by using \( \varphi_{[K^{(m)}]} \) and \( \varphi_{[K^{(m)}]}^{-1} \).

Lemma 3 \( M^{(m)} \) -1 is unique and bounded linear operator from \( \text{Im}(\varphi_{[K^{(m)}]}) \) to \( \text{Im}(\varphi_{[K^{(m)}]}) \).

Lemma 4 Both \( \varphi_{[K^{(m)}]}^{-1} \) and \( \varphi_{[K^{(m)}]} \) are bounded, if and only if \( M^{(m)} \) is bounded.

By both Lemma 2 and Lemma 4, the following Proposition can be derived.

Proposition 3 If and only if \( M^{(m)} \) is bounded, the sampling basis \( \{\eta_{k}^{(m)}\}_{k=-\infty}^{\infty} \) exists in \( K^{(m)} \) and can be expressed by

\[ \eta_{k}^{(m)} = \sum_{t=-\infty}^{\infty} M^{(m)}(\delta_{k})(t)\xi_{t}^{(m)}, k = 0, \pm 1, \pm 2, \pm \ldots. \]  

(22)

In this subsection, we have defined the coordinate systems \( \varphi_{[K^{(m)}]} \) and \( \varphi_{[K^{(m)}]}^{-1} \), and the operator \( M^{(m)} \) regarding \( K^{(m)} \). Then, we have shown that the existence of the sampling basis is equivalent to the boundedness of \( M^{(m)} \).

3.2 Existence of sampling basis

In this subsection, we clarify the existence of the sampling basis for the state function space \( K^{(m)} \) by proving the boundedness of operator \( M^{(m)} \).

Proposition 4 The coordinate transform operator \( M^{(m)} \) is bounded.

From the above, the existence of sampling basis is obtained as follows:

Theorem 1 The sampling basis exists for the state function space \( K^{(m)} \).

By the above, we have proved that the continuous-time system model \( \Sigma \) has its isomorphic fluency discrete-time system model.

4 Conclusions

This paper aimed at guaranteeing the isomorphism of the fluency system model which was proposed by the authors. We found the existence of the sampling basis in the state function and output spaces. The result obtained in this paper provided a view that the isomorphism of the fluency system model was guaranteed.

We show a practical example of the present theory. If the polygonal signal is input to the system, conventional discrete-time system model with which a staircase signal input is assumed cannot give the desired result exactly. Using the fluency discrete-time system model, we can get the desired result exactly and this paper guaranteed this fact by deriving the existence of the sampling basis.

Problem left for the future is to analyze the characteristics of function and to verify the effectiveness of constructed systems by application to actual problems.

Acknowledgments

The authors wishes to express his sincere gratitude to Professor Rudolf E. Kalman in the Mathematical System Theory at Swiss Federal Institute of Technology (ETH) for his valuable directions. The authors also wish to thank Professor H. Inaba in the Department of Information Sciences, College of Science and Engineering at Tokyo Denki University and Dr. M. Kamada in the Department of Computer and Information Sciences at Ibaraki University for the useful discussions.

References

M-Sequence Detection by Higher Order Signal Correlation Techniques

Edward S. WARNER, Bernard MULGREW and Peter M. GRANT

Dept. of Electrical Eng., The University of Edinburgh, Edinburgh EH9 3JL,
Scotland, U.K., Tel/Fax: +44 [31] 650 5659 / 650 6554.

Abstract. This paper describes the application of triple correlation techniques to identify characteristics of the class of cyclic binary sequences known as maximal length sequences. A maximal length or m-sequence is generated by a linear feedback shift register (LFSR) whose feedback function uniquely specifies the m-sequence. The cyclic nature of m-sequences leads to the appearance of peaks in their triple correlation at certain shift pairs. By exploiting the properties of m-sequences and using polynomial theory, it is shown how these shift pairs can be used to identify the feedback function for the LFSR which generates the m-sequence. Examples of how to find the feedback function from the shift pairs at which peaks occur are given.

1 Introduction

Higher order spectral analysis techniques (HoS) [1,2,3] provide important signal processing tools for communication engineers. HoS techniques aim to characterize signals in terms of both their moments and cumulants and the associated Fourier transforms to elucidate information about the signals. In particular, signal moments derived from multiple correlations can be used to establish information about periodic signals [3]. The application of such techniques for the detection of the class of cyclic binary codes known as m-sequences (or pseudo-random noise (PN) codes) [4,7] will now be considered.

2 Triple Correlation of an M-Sequence

Let $b(t)$ be an $m$-sequence of period $L = 2^n - 1$ with bit (also known as chip) duration $\tau_c$ generated by an $n$-stage linear feedback shift register (LFSR) [4]. Figure 1 shows the generation of such a sequence with the continuous autocorrelation function (ACF) included. The LFSR is specified by a primitive polynomial $h(x)$ [4].

Let the code chips map as $(0,1) \rightarrow (1, -1)$ so that addition modulo-2 ($\oplus$) can be replaced by multiplication. The triple correlation of $b(t)$ is:

$$R_0(\tau_1, \tau_2) = \frac{1}{L\tau_c} \int_{-L\tau_c}^{L\tau_c} b(u)b(u + \tau_1)b(u + \tau_2) du$$  

(1)

The ACF and shift-and-add properties of m-sequences [4] can be used to evaluate $R_0(\tau_1, \tau_2)$, i.e., $b(t + L\tau_2)b(t + m\tau_2) = b(t + (m \neq p) \tau_2)$ where $l \neq m \neq p$; and $R_0(\tau) = 1, \tau = kL\tau_c$ and is equal to $-1/L$ otherwise. $b(t)$ is cyclic so $R_0(\tau_1, \tau_2)$ need only be evaluated for $\tau_1 \geq \tau_2$ and $0 \leq \tau_1 \leq (L - 1)$; values at all other shifts can be found by symmetry. Clearly, $R_0(\tau_1, \tau_2)$ takes only two values: $-1/L$ or 1. For certain pairs of shifts $(\tau_1, \tau_2)$ the product sequence $b(t + \tau_1)b(t + \tau_2) = b(t)$ and $b(t)$ is a multiple of $L\tau_c$; or else the triple product is some shift of $b(t)$.

For each shift $\tau_1$, except the zero shift, there are two shifts $\tau_2$ (one positive the other negative) at which peaks occur ($4(L - 1)$ in total). Figure 2 shows the delay shift pairs for the m-sequence $h(x) = 67$ (octal). Table 1 lists the first 8 delay shift pairs for this and 3 other m-sequences of different period.

The shift pairs which correspond to the peaks in $R_0(\tau_1, \tau_2)$ are related to the sequence length. The m-sequence $b(t)$ can be represented as a polynomial in the shift operator $x$ [4]:

$$b(x) = b_0 + b_1x + \ldots + b_{L-1}x^{L-1}$$  

(2)

![Figure 1: LFSRG for the $L = 2^4 - 1 = 15$-chip m-sequence, $h(t)$, generated by $h(x) = 1 \oplus x \oplus x^3$ with load $(a_0, a_1, a_2, a_3) = (1, 0, 0, 0)$. $R(\tau)$ is the normalized ACF with the same period as the m-sequence.](image-url)

![Figure 2: Delay shift pairs for the m-sequence $h(x) = 67$ (octal). Table 1 lists the first 8 delay shift pairs for this and 3 other m-sequences of different period.](image-url)
<table>
<thead>
<tr>
<th>Polynomial (octal)</th>
<th>Code period (in chips)</th>
<th>Delay shifts for peaks in triple correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>15</td>
<td>(1,4) (2,8) (3,14) (4,1) (5,10) (6,13) (7,9) (8,2)</td>
</tr>
<tr>
<td>67</td>
<td>31</td>
<td>(1,13) (2,26) (3,23) (4,21) (5,7) (6,15) (7,5) (8,11)</td>
</tr>
<tr>
<td>163</td>
<td>63</td>
<td>(1,25) (2,50) (3,55) (4,37) (5,40) (6,47) (7,53) (8,11)</td>
</tr>
<tr>
<td>361</td>
<td>127</td>
<td>(1,87) (2,47) (3,57) (4,94) (5,21) (6,114) (7,98) (8,51)</td>
</tr>
</tbody>
</table>

Table 1: Location of the first 8 peaks ($r_1, r_2$) in the triple correlation of 4 m-sequences of different period.

![Figure 2: Triple correlation display for the 31-chip m-sequence $h(x) = 67$.](image)

where $b_t \in (0, 1)$ and multiplication of $h(x)$ by $x$ yields a cyclic shift of the sequence to the right (a delay) so:

$$x^p h(x) \oplus x^q h(x) = b(x) \quad p = \pm 1, \pm 2, \ldots$$  \hspace{1cm} (3)

$$x^p [1 \ominus x^{red}] = 1$$ \hspace{1cm} (4)

$$x^{2p} \ominus x^{4p} = x^p$$ \hspace{1cm} (5)

so

$$x^{2p} \ominus x^{4p} = 1$$ \hspace{1cm} (6)

but $x^{2p} \ominus x^t = 1$ and $x^t = 1$ \hspace{1cm} (7)

thus $r = (2q)$ modulo $-L$ \hspace{1cm} (8)

For example, from Figure 2 for $r_1 = -2$ and $r_2 = -26$ then, for $r_1 = -4$ we should find $r_2 = -52$ modulo $L$. From Figure 2 $r_2 = -21$ as expected. Each m-sequence has an unique pattern $R_{0}(r_1, r_2)$ because $h(x)$ is unique. When an m-sequence is the time reverse of another, the patterns are 180° rotations of each other in the $r_1 r_2$ plane.

3 Feedback Connection Determination

The feedback connections can be determined if the number of LFSRG stages $n$ and a segment of $2^n + 1$ consecutive chips of the m-sequences are known [5]. They can also be found by triple correlation. Let an LFSRG generate an m-sequence $b(t)$ whose chips take the values $(0, 1)$. Its operation can be described by a matrix equation $X' = Tx$ [6]. Addition is modulo-2, $x$ is the $(n \times 1)$ state vector of the LFSRG and $X'$ is the updated state vector at the next clock pulse. $T$ is the $(n \times n)$ transformation matrix [6] whose diagonal elements $(i, i + 1) + i \leq n - 1$ are 1 and whose last row is given by the coefficients of the LFSRG feedback taps ($c_0, \ldots, c_{n-1}$). The determinant of $T$ is $c_0$ and $c_0 = 1$ so that $T$ is non-singular. The characteristic polynomial of the matrix $T$ is [6]:

$$\phi(\lambda) = |T - \lambda I| = c_0 + c_1 \lambda + \ldots + \lambda^n$$  \hspace{1cm} (9)

where $J$ is an $(n \times n)$ identity matrix and $\phi(\lambda)$ is found by expanding $|T - \lambda I|$ by minors of the last row of $T$. The polynomials $\phi(\lambda)$ and $h(x)$ have the same form in the notation adopted [4]. We now need to relate the shift pairs at which peaks occur in the triple correlation to $T$. The signal product $b(t + r_1) \ominus b(t + r_2) = b(t + r_3)$ of eq.(1) can be expressed as a matrix equation:

$$(I \oplus T^{(r_1 - r_2)})x = T^{(r_1 - r_2)}x$$  \hspace{1cm} (10)

If $\phi(\lambda)$, $(r_1, r_2)$ are known then the product shift $r_3$ can be found by use of the Cayley-Hamilton theorem [6], i.e.

$$\phi(T) = 0$$

so that powers of $T$ can be expanded as matrix polynomials and noting that $T^{-L} = J$. Similarly, if the shift pairs $(r_1, r_2)$ for a given $\tau_2$ are known and we are free to choose the state vector $x$ then, in principle, eq.(10) can be solved for the coefficients of $\phi(\lambda)$. Note that $\tau_2 \equiv 0$, i.e.

$$b(t + \tau_3) \equiv b(t)$$

of eq.(1) as the starting epoch is an arbitrary reference point. This method is likely to be cumbersome when the $T$ matrices are raised to large powers. Alternative methods to determine the feedback connections will now be given.

The sequence length, $L$, and hence the number of LFSRG stages, $n$, are assumed to be known. For a given value of $L$ the number of unique m-sequences, i.e. primitive polynomials $h(x)$, is known to be given by the Euler totient function $\phi(n)$. The signal $b(t)$ could be cross-correlated with each of the allowed m-sequences; the one which yields a correlation peak will be the correct m-sequence.

From polynomial theory the following Lemma is well known [6]: Let the minimal polynomial of the matrix $T$ be $f(\lambda)$. If $\phi(\lambda)$ is any polynomial such that $p(T) = 0$ then $f(\lambda)$ divides $p(\lambda)$, i.e. $p(\lambda) = f(\lambda)g(\lambda)$ for some polynomial $g(\lambda)$. We note that for an LFSRG which generates an m-sequence $f(\lambda) = \phi(\lambda)$. The matrix polynomials eq.(10) represent $p(T)$ and so:

$$p(\lambda) = \phi(\lambda)g(\lambda)$$

and $g(\lambda) = \sum_{i=0}^{\text{dim}} g_i \lambda^i$  \hspace{1cm} (11)

where $g_0 = g_{\text{dim}} = 1$, $\text{dim}$ is the degree of $g(\lambda)$ which equals $(r - n)$ and $v$ is the degree of $\phi(\lambda)$, $n$ from the m-sequence length $L = 2^n - 1$ and $\phi(\lambda)$ is eq.(9). The coefficients $c_i$ of
\( \phi(\lambda) \) can be found by equating the powers of \( \lambda \) in eq. (11). The divisor, \( g(\lambda) \), of \( p(\lambda) \) with zero remainder also identifies \( \phi(\lambda) \). There are \( N = 2^{(\text{dim} - 2)} \) possible values of \( g(\lambda) \) since, both \( p(\lambda) \) and \( \phi(\lambda) \) have an even number of terms \( \lambda \) so must \( g(\lambda) \) and \( g_0 = g_{\text{rem}} = 1 \). As an example, from Figure 2 select the delay shift pair \( (5, 7) \) and substitute in eq. (10) to give \( p(T) = T^5 \oplus T^7 \oplus I \). Since \( v = 7 \) then \( \text{dim} = 2 \) and \( g(\lambda) = 1 \oplus g_1 \lambda \oplus g_2 \lambda^2 \). By substituting \( p(\lambda) \) and \( g(\lambda) \) in eq. (11) and comparing coefficients we find \( \phi(\lambda) = 1 \oplus \lambda \oplus \lambda^2 \oplus \lambda^3 \oplus \lambda^4 \) \((57\text{ octal})\). Alternatively, the remainder of the division of \( p(\lambda) \) by \( \phi(\lambda) \) is zero when \( \phi(\lambda) = 1 \oplus \lambda \oplus \lambda^2 \) which yields the same result for \( \phi(\lambda) \). The solutions of eq. (11) may not be unique, i.e. the shift pair \((\tau_1, \tau_2)\) may satisfy more than one polynomial. The axiom eq. (8) may then be used to determine the correct \( \phi(\lambda) \) by means of pattern recognition.

\( \phi(\lambda) \) can also be determined from the matrix polynomials eq. (11) by exploiting the relations among the various shifts. As an example of one algorithm consider that the first \( n \) polynomials \( p(T) \) are tabulated in ascending order \( I \oplus T^m \equiv T^m \) for \( m = 0, 1, 2 \) etc.. The last \( n \) polynomials are known since \( I \oplus T^{\text{dim}-m} \equiv T^{\text{dim}-m} \). Look for a combination in the first \( n \) terms which, when multiplied by powers of \( T \), satisfies a relation in the last \( n \) terms. Make the appropriate substitutions, multiply by powers of \( T \) and use \( T^m \equiv I \) and compare coefficients to those of \( \phi(T) \). For example, for the 127-chip m-sequence of Table 1 it is found that \( I \oplus T^4 \equiv T^4 \) and \( I \oplus T^6 \equiv T^{126} \). Multiplying the first polynomial by \( T^3 \) and substituting for \( T^6 \) in the second yields \( T^3 \oplus T^7 \equiv I \oplus T^{126} \). Now multiply by \( T \) and use \( T^{127} \equiv I \) to give \( T(T^3) \equiv T^3 \oplus T \oplus I \). Substituting for \( T^3 \equiv \sum_{i=1}^{127} c_i T^i \) and comparing coefficients leads to \( \phi(T) \equiv T^3 \oplus T^6 \oplus T^8 \oplus T^4 \oplus I \) which is equivalent to \( \phi(\lambda) \) \((361\text{ octal})\). No doubt, other algorithms which exploit the relations eq. (11) can also be found.

4 Conclusions

It has been shown that the feedback configuration for an LFSRG which generates an m-sequence can be found through the use of triple correlation. By the shift-and-add property of m-sequences peaks in the triple correlation occur at shift pairs which are unique to the LFSRG which generates the m-sequence. Methods to determine the LFSRG feedback connections from these shift pairs have been derived and verified. The robustness of higher order correlation techniques in the presence of noise and their operation with data modulation now deserve further investigation. Such techniques could obviously be used for the detection of direct-sequence spread-spectrum signals.

We have shown [8] how this technique can be applied to determine the feedback connections for m-sequences. On 63-chip m-sequences codeword detection has been accomplished at -8dB input signal-to-noise ratio (SNR). Furthermore, we have shown how the detection statistics vary with different SNR values.

Acknowledgement

This work was supported by the UK Science and Engineering Research Council under grant no. GR/H 72748.
Blind Estimation of ARMA Systems

Dieter BOSS and Karl-Dirk KAMMEYER

Hamburg University of Technology, Dept. of Communications, Essenerdorfer Str. 40, 21071 Hamburg, Germany,
Tel.: ++(49)-40/7718-2168, Fax: ++(49)-40/7718-2281, E-mail: boss@tu-harburg.de

Abstract: In this paper, we present an approach to blind estimation of non-minimum phase ARMA models using fourth order statistics. The algorithm follows a Residual Time Series (RTS) procedure, which sequentially identifies the AR and MA parts. In step 1 of the RTS concept, the AR estimation is geared to deliver the shortest impulse response of the overall system, so that a new fast approach to MA system identification can be applied in step 2. Investigations have shown that, in order to obtain a given ARMA estimation quality, the Yule-Walker based AR identification requires much more data samples of the channel output signal than does the subsequent MA estimation based on eigenvectors. In an attempt to decrease AR estimation bias and variance at a given blocklength, the Structured Total Least Squares (STLS) scheme has proved to achieve better pole estimates of channels with allpasses. However, the STLS technique can not improve on the standard least squares performance, if the cumulant estimation errors are highly correlated, which is true for systems with allpass-free poles close to the unit circle. For these reasons, the Allpass Separation (APS) approach has been conceived, which first identifies the allpass-free poles using 2nd order statistics, thereby correlating the cumulant estimation errors at the overall system's output. This is why, in a 2nd step, the STLS can be applied successfully. Briefly, for the allpass-free poles, APS benefits from high quality 2nd order estimates while STLS yields the best identification of allpass-poles known to the authors.

1 Introduction

With digital communications, the channel input random processes usually are complex, because they are described by the equivalent baseband representation of digital modulation signals such as PSK or QAM. Although this is neglected quite frequently in literature, we will apply complex cumulants as defined in [1]. Furthermore, digital modulation signals reveal symmetric distributions and thus prohibit the application of 3rd order statistics due to their zero skewness. We will therefore use 2nd and 4th order statistics, exclusively.

The complex discrete channel input random process (r.p.) \( d(k) \) is supposed to be stationary, non-Gaussian, zero mean, 2nd and 4th order white, and to have unit power \( \sigma_d^2 = 1 \). The 2nd transform of the channel's complex impulse response \( h(k) \) is modelled as a causal and stable, possibly non-minimum phase, ARMA-(p,q) model

\[
H(z) = \frac{b_0 + b_1 z^{-1} + \ldots + b_q z^{-q}}{a_0 + a_1 z^{-1} + \ldots + a_p z^{-p}},
\]

where the \( a_i, b_i \) are the AR and MA coefficients, and \( p \) and \( q \) are called the AR and MA orders, respectively. The above model is assumed to be free of pole-zero-cancellations and to have unit power amplification. Note that the most critical assumption with a multitude of applications is the time invariance of the transmission channel. This is why we shall try to use as few samples of the channel output r.p. for the estimation of moments as ever possible.

Objective: using samples of the channel output r.p. \( x(k) \) only, the aim of blind ARMA identification is to estimate the parameters \( a_i, b_i, p \) and \( q \) so that the deviation of the estimated impulse response (IR) \( \hat{h}(k) \) from the true channel IR \( h(k) \) is minimum. Therefore, the criterion for the assessment of an estimated IR will be \( \sum_e |\hat{h}(k) - h(k)|^2 \).

2 Blind MA Identification

In this section, a fast approach to blind MA system estimation will be outlined. As it is based on the eigenvectors of particular cumulant matrices, it is termed "EVI" standing for EIGENVECTOR ALGORITHM FOR IDENTIFICATION. Compared with other algorithms, EVI is a fundamentally novel approach based on a closed-form solution to the problem of blind equalization [2]. The iterative algorithm "EVA" derived from this closed-form solution has been presented in [3]; a comprehensive analysis and a generalisation of EVA to include other types of equalisation has been submitted [4]. As far as system identification is concerned, EVI was presented in [5]. A detailed description has also been submitted [6]. The purpose of this paper is to integrate EVI into a comprehensive algorithm for blind ARMA identification. We will therefore just outline its most distinctive properties.

In contrast to other approaches, such as Giannakis and Mendel's Least Squares solution [7] (modified in [1]), the Cumulant Zero Matching approach [8] or Exhaustive Search algorithms [9], EVI achieves a given estimation variance at considerably reduced block lengths of the channel output data. Rather than estimating the cumulants on the basis of tens of thousands of data values (which is indispensable on the conditions detailed in Section 1), this iterative approach requires just some hundred data values [5, 6]. This has to be em-
phasized, because future work will be directed towards blind estimation of time variant mobile communication channels. EVI is also insensitive with respect to MA order overfits, which lead to singularities in the system of equations utilised by some approaches [e.g. (7)]. Recent investigations have shown that EVI is quite robust with respect to additive white Gaussian noise, too.

3 Blind ARMA identification

In this section, a Residual Time Series (RTS) approach to blind ARMA model identification will be given.

The following three step algorithm is proposed:

S1: Given $P$ values $p_i, i = 1 \ldots P$ of the AR order $p$, estimate the channel’s AR part $A_p(z)$ with the aid of modified Yule-Walker equations [10] generalised to 4th order statistics [7]. Realising the transfer functions of these estimations $\hat{A}_p(z)$ as FIR-filters, we obtain $P$ overall systems $H_{M A P}(z) = H(z) \hat{A}_p(z)$ (see Fig. 1). The impulse responses $h_{M A P}(k)$ of these systems are called the “residual time series” (RTS).

S2: Select the estimation $\hat{A}_p(z)$ delivering the shortest RTS $h_{M A P}(k)$. This will be accomplished by minimising a higher order cost function. Note that, thereby, the AR order is estimated implicitly.

S3: The selected RTS $h_{M A P}(k)$ ideally being finite (hence its subscript), identify the overall system $H_{M A P}(z)$ with the EVI algorithm described in section 2.

These three steps S1 to S3 will be described in the sequel.

3.1 S1: Blind AR Identification

The AR coefficients $a_1, \ldots, a_p$ of the ARMA-(p,q) model $H(z)$ according to eq. (1) can be calculated from the modified 4th order Yule-Walker equation

$$c_f^2(m, n_0, n_1) + \sum_{k=1}^{p} a_k^* \cdot c_f^2(m - k, n_0, n_1) = 0,$$

where $c_f^2$ represents the 4th order complex cumulant and $m > q + \min\{0, n_0, n_1\}$ with arbitrary integers $n_0$ and $n_1$. In a first attempt, we collect $p$ equations (2) by letting $m = q + 1 \ldots q + p$ to obtain

$$\begin{bmatrix} c(q, \cdot) & \ldots & c(q + 1 - p, \cdot) & \ldots & c(q + p, \cdot) \end{bmatrix} a = \begin{bmatrix} c(q + 1, \cdot) \ldots c(q + p, \cdot) \end{bmatrix}.$$  

where $c(\cdot, \cdot)$ has to be replaced with $c_f(\lambda, n_0, n_1)$ and $a$ is defined as $a = [a_1 \ldots a_p]^T$. Note that the asterisk with a vector or matrix stands for the conjugate complex transpose. With the evident matrix definitions, the above eq. reads

$$H_{n_0, n_1} \cdot a = -c_{n_0, n_1},$$

where $H_{n_0, n_1}$ is a Toeplitz matrix.

The straightforward way to solve for the $a_i$ is to choose $n_0 = n_1 = 0$ and solve eq. (3) by inversion. On account of the symmetry property $c_f(\lambda, 0, 0) = c_f(-\lambda, -\lambda, -\lambda)$, this approach is said to use “diagonal cumulants”, only. However, the results gained by this approach generally show a high estimation variance. In order to improve the AR estimation performance, the number of equations can be increased by taking non-diagonal cumulants ($n_0, n_1 \neq 0$) into account. If we concatenate eq. (3) for different values of $(n_0, n_1)$, say $n_0 = n_00 \ldots n_0M$ and $n_1 = 0$, we obtain the overdetermined system of equations

$$\begin{bmatrix} H_{n_01, 0} & \vdots & \vdots & \vdots \\ \vdots & \ddots & \vdots & \vdots \\ H_{n_0M, 0} & \vdots & \vdots & \vdots \\ \end{bmatrix} a = \begin{bmatrix} c_{n_01, 0} \\ \vdots \\ c_{n_0M, 0} \end{bmatrix},$$

or equivalently

$$H \cdot a = -c.$$

Eq. (4) can be solved in a least squares sense.

3.2 S2: Selection of the “best” AR estimation

In step S1, we have thus obtained $P$ overall systems

$$H_{M A P}(z) = B_q(z) \cdot \hat{A}_p(z).$$

Ideally, their impulse responses $h_{M A P}(k)$ will be finite. Realistically, however, they will be infinite due to an imperfect AR estimation. Bearing in mind that, the shorter the respective RTS is, the better the estimates of the MA coefficients (c.f. sec. 2) will be, it becomes obvious that we need to select the AR order such that the RTS $H_{M A P}(k)$ has minimum length. The cost function can therefore be defined as

$$\sum_{k=0}^{\gamma} |h_{M A P}(k)|^2 \leq \min,$$

where $\gamma$ represents a threshold index in time. Let $w(k)$ be a weighting function defined as

$$w(k) = \sum_{\nu=0}^{k_{\max}} |h_{M A P}(k - \gamma - \nu)|^2,$$

which – when multiplied with the squared magnitude of the RTS – suppresses its first $\gamma$ coefficients. $k_{\max}$ is supposed to be chosen large enough such that $[h_{M A P}(k)]^2$ can be considered negligible beyond the index $\gamma + k_{\max}$. Now, considering

$$f(p, \gamma) = \gamma^d \sum_{k=0}^{k_{\max}} |[h_{M A P}(k)|^2 \cdot w(k)|$$

$$= \sum_{\nu=0}^{k_{\max}} \gamma^d \sum_{k=0}^{\infty} |h_{M A P}(k)|^2 |h_{M A P}(k - \gamma - \nu)|^2$$

$$= \sum_{\nu=0}^{k_{\max}} \gamma^d (-\gamma - \nu, 0, -\gamma - \nu),$$

1106
we conclude that minimising \( f(p_i, \gamma) \) can be regarded as equivalent to the requirement in eq. (5). To render this measure independent from the power transfer function of the respective estimation \( \hat{A}_p(z) \), we normalise this function to obtain

\[
F(p_i, \gamma) \triangleq \frac{1}{f_p(\gamma)} \sum_{r=0}^{L_{max}} c_l(\gamma + \nu, \gamma + \nu, 0).
\]  

(6)

Minimising \( F(p_i, \gamma) \) thus delivers the estimated AR order \( \hat{p} \).

3.3 S3: ARMA Identification

Having selected the AR estimation \( \hat{A}_p(z) \) with the aid of equation (6), we need to identify the RTS

\[
h_{MA,(p)}(k) = Z^{-1} \left\{ H_{MA,(p)}(z) \right\} = Z^{-1} \left\{ H(z) \hat{A}_p(z) \right\}.
\]

As this is FIR (in the ideal case), the EVI approach described in section 2 can be used. Terming its estimation \( \hat{B}_q(z) \), we achieve the final solution to the ARMA estimation problem

\[
\hat{h}(k) = Z^{-1} \left\{ \hat{B}(z) \right\}, \quad \text{where} \quad \hat{B}(z) \triangleq \frac{\hat{B}_q(z)}{\hat{A}_p(z)}.
\]  

(7)

Notice the following advantages of this RTS scheme:

(i) Usual applications of the RTS procedure to ARMA system identification attempt to estimate the ARMA parameters \( p, q, a_i \) and \( b_i \) as exactly as possible. However, in a set-up as described in section 1, this is not really the ultimate goal. We try to approximate the true IR instead.

(ii) Commonly, RTS based algorithms separately minimise the variance of the AR and MA estimates. Here, we pursue a unified approach, where the AR estimation is geared to deliver an RTS which is optimum for the subsequent MA identification (i.e. finite). The criterion how to select the "optimum" AR estimation is NOT the quality of the approximation of \( 1/\hat{A}_p(z) \) by \( 1/\hat{A}_p(z) \), but the length of the resulting RTS.

(iii) Remember that badly estimated poles\(^4\) are not relevant in the sense of our overall concept. Such an estimation represents a zero of the z transform of the RTS and will therefore be identified by EVI. The crucial point is that poles with near unity magnitude must be estimated as exactly as possible. Otherwise, the corresponding RTS would be very long and the MA estimation would fail. The algorithm described in step S2 prevents such RTS from being retained.

Given different values of \( p_i \), there are four possible cases:

1. Critical underfit: if we underestimate the number of "critical" poles (those with near unity magnitude), we will end up with a long RTS. Such an estimation will not be retained in S2 of the approach.

2. Uncritical underfit: critical poles will be compensated for. The remaining poles will lead to a (relatively quickly) decaying RTS.

3. No misfit: ideally, the RTS will be finite; however, due to estimation errors, it will be a quickly decaying infinite IR.

4. Overfit: additional zeros, which will be identified in step S3. However, the RTS will decay quickly.

\(^4\)Which occur in case of an order misfit or when the magnitude of a true pole is small.

3.4 Simulation Results

Fig. 2 shows the results of the RTS approach presented here. The input r.p. is QPSK, and the ARMA channels are of order \((p,q) = (3,1)\) with a zero at \( z = 0.4 + 0.6j \). Fig. 2a and b display the true poles as small circles (covered by dots).

![Figure 2: Estimation results of RTS approach](image)

In S1, the AR estimations (25 Monte Carlo runs for each value of \( p_i = 1 \ldots 4 \)) were gained by solving equation (4) in the least squares sense \((n_{01} = q - p, n_{0M} = q)\). The zeros of the AR identifications \( \hat{A}_p(z) \) which were retained in S2 are shown as dots (i.e. the retained estimated channel poles). Both eq. (4) and (6) use cumulant estimates based on 30000 samples. Then, in S3, the remaining 25 RTS were estimated by EVI using cumulants and autocorrelation coefficients based on blocks of 1000 data values. For channel 1 (fig. 2a), fig. c shows the mean and std dev. (dotted) of the estimated impulse responses \( \hat{h}(k) \) according to eq. (7). For comparison, a shifted version of the true channel IR is given (dashed).

4 Advanced AR estimation techniques

The above simulation results reveal some sort of discrepancy with the RTS approach presented in sec. 3. While EVI utilises just 1000 samples of the channel output signal for a proper MA identification, the Yule-Walker based AR estimation does need as many as 30000 samples. Hence, some effort has been spent on improving AR estimation quality. So far, eq. (4) has been solved in the least squares (LS) sense. However, the standard LS-approach ignores errors in \( H \), because it modifies just the righthand side vector \( c \) to reduce the rank of eq. (4). Since both \( H \) and \( c \) contain cumulants subject to estimation errors, this is at least questionable. Approaches that modify both the righthand side vector and the lefthand side matrix to reduce the rank of a linear system of equations are termed Total Least Squares (TLS) algorithms [11]. Advanced TLS-techniques, such as the STRUCTURED TOTAL LEAST SQUARES (STLS) scheme [12] permit to account for errors in the matrix while retaining the inherent multiple Toeplitz structure of matrix \( H \).
Further investigations have proven that the performance of the LS-algorithms under consideration (LS, TLS, STLS) do profit from correlations between the cumulant estimation errors. Strong correlations exist in the case of channels with allpass-free poles close to the unit circle. However, with these channels the TLS schemes can not further improve on the standard LS-performance unless the error correlations are known in advance (which is not the case). The only way to further improve AR estimation quality is to resort to 2nd order statistics. On the other hand, with channels containing allpasses, STLS outperforms any other scheme, mainly because the cumulant estimation errors are uncorrelated.

The results mentioned above have led to the development of the two step ALLPASS SEPARATION (APS) approach:

S1: Estimate the allpass-free poles by solving the 2nd order Yule-Walker eq. obtained by replacing $r_x(\lambda)$ with $r_{xx}(\lambda)$ in eq. (3). This can be accomplished by matrix inversion or with the LS-approach, as proposed by Cadzow [13]. Here, utilizing the autocorrelation sequence assures excellent pole estimates. Similar to the RTS scheme, the channel output data are fed through the FIR filter with the estimated AR transfer function.

S2: Based on this filter's output signal, estimate the channel poles belonging to allpasses by solving eq. (4) in the STLS sense. As the FIR filter has decorrelated the cumulant estimation errors, STLS yields a remarkably good identification of the allpass-poles.

To work in an optimum fashion, APS requires both the number of allpasses and the total AR order $p$ to be known. These orders can be determined by singular value decomposition of the matrix used in the respective step. The orders are equal to the number of non-zero singular values.

Fig. 3a and b display the pole-zero-plots, where 2 out of 3 zeros belonging to allpass factors are marked with a cross ("×"), the 3rd one is in $z = -2$. True poles are represented by small circles, the pole estimates by dots. The figures 3c and d show the estimated and exact magnitude spectra.

S1: Due to long data blocks (15000 samples), solving the 2nd order Yule-Walker equation by matrix inversion provides us with almost perfect estimations of the 3 allpass-free poles. This can be seen from the figures a and b, where the true ("•") and the mean estimated ("•") pole symbols lie on top of each other and no "estimation clouds" are visible.

S2: Here, eq. (4) is solved in the LS (Figures a and c) and STLS (Fig. b and d) sense, respectively. Both the pole plots and the magnitude spectra clearly show that STLS outperforms LS in terms of bias and std deviation. This is due to its application to an ARMA channel containing allpasses.

References


Time Delay/Scale Estimation

Using Wavelant Theory - Higher Order Wavelets

Malachy Devlin and Tariq S. Durrani

NEL, East Kilbride, Glasgow, G75 0QU, Scotland, U.K., Tel/Fax: +44 [3552] 72317/ 2979, E-Mail: sydevlin@uk.net

*Signal Processing Division, Dept. of Electronic and Electrical Eng., University of Strathclyde, Glasgow G1 I XW, Scotland.
U.K., Tel/Fax: +44 [41] 552 4400 / 552 2487, E-Mail: tsi@uk.ac.strathclyde.eee.spd

Abstract: A theory based on wavelet transform and higher order statistical theory has been developed by the authors which has been termed wavelants to identify its relationship with wavelets and cumulants. Definitions and a fuller intorduction to this work has been provided in previous literature [7], whereas, this paper is concerned with introducing the reader to the ideas and definitions of the third order wavelant in particular. To show the advantages of wavelants a problem scenario is given for which the solution can be easily found using the third order wavelant but is not feasible using wavelet transforms or cumulants alone. The problem to solve is the estimation of time delay and scaling parameters of signals at two or three sensors, where the signals may contain additive correlated noise consisting of both scaled and unscaled noise. Results for the new estimation algorithm based on wavelants is given and compared to results using conventional techniques.

1 Introduction

An important application of signal processing is time delay estimation [1][2][3] as it plays a key role in many areas such as sonar and radar. Consequently this has generated much research interest and resulting literature [1] on techniques for time delay estimation. These techniques have undergone an evolutionary process starting with highly constrained models and as the necessary theory has been developed and applied to time delay estimation some of these constraints have been lifted.

Making a generalisation of time delay estimation we begin with one of the simpler models -

\[ x(n) = s(n) \]  \hspace{1cm} (EQ 1)

\[ y(n) = s(n-D) + w(n) \]  \hspace{1cm} (EQ 2)

where \( x(n) \), \( y(n) \) are two sensor measurements

where \( s(n) \) is the known signal and \( D \) is the time delay parameter to be estimated, and \( w(n) \) is white gaussian noise.

This problem was solved using matched filtering/correlation approaches and could also be tailored to incorporate narrowband assumptions, i.e. where \( s(n-D) \) had experienced small frequency shifts.

This model can be extended by removing the narrowband constraint to produce wideband estimation so that doppler effects, for example, could not be confidently approximated to frequency shifts, this therefore give as a model

\[ x(n) = s(n) \]  \hspace{1cm} (EQ 3)

\[ y(n) = s\left(\frac{(n-D)}{A}\right) + w(n) \]  \hspace{1cm} (EQ 4)

Using this model fast moving targets and/or sensors could be used and improved accuracy in estimation obtained. This estimation problem can be solved using the wideband ambiguity function or the recently introduced wavelet transform [4].

Another direction of extending the initial model is relaxing the assumptions on noise, e.g.

\[ x(n) = s(n) + w_1(n) \]  \hspace{1cm} (EQ 5)

\[ y(n) = s(n-D) + w_2(n) \]  \hspace{1cm} (EQ 6)

where \( w_1(n) \) and \( w_2(n) \) are gaussian noise sources which may be correlated

For this the match filtering/correlation methods will fail. However if the signal is non-gaussian, which is the case in several scenarios [5], then the use of higher order statistics can be applied to solve this problem [6]. However the narrowband assumption is retained in this model such that the wideband model

\[ x(n) = s(n) + w_1(n) \]  \hspace{1cm} (EQ 7)

\[ y(n) = s\left(\frac{(n-D)}{A}\right) + w_2(n) \]  \hspace{1cm} (EQ 8)

where \( w_1(n) \) and \( w_2(n) \) are possibly correlated gaussian noise sources, and \( s(n) \) is non-gaussian or deterministic.
can not be processed optimally to obtain the delay estimation and optionally the scale factor/doppler shift [7]. This has generated the motivation to develop new analysis tools for scenarios such as this latter model. It can be appreciated that elements from wavelet transform and higher order analysis would contribute to the required solution. This is one of the motivating applications for the amalgamation of these two theories which gave rise to the recently defined analysis tools [7] termed ‘wavelets’ to signify the unity of wavelet and cumulant theories. With these tools it is now possible to perform broadband time delay estimation on measurements containing additive gaussian noise which may be correlated. Furthermore it will be seen that earlier models become special cases of the proposed wavelet time delay/scale estimation algorithm.

2 Higher Order Wavelets

The motivation for merging wavelets and higher order statistics has been outlined above, in defining these new analysis tools, two branches are possible. Firstly definitions based on higher order moments, which are termed ‘wavepaks’, have been developed but it has proved to be advantageous to use the second branch of definitions. This branch is based on cumulants, termed ‘wavelets’, and offers a number of properties not available when using wavepaks. In this paper we shall concentrate on the third order wavelet for solving the current application, however higher orders of wavelets are available for use in other scenarios [8].

As defined in [7] the deterministic definition of the third order wavelet is given by

\[ W_{xx}^3 (a_1, b_1; a_2, b_2) = \]

\[ = \frac{1}{\sqrt{a_1 a_2}} \int_{-\infty}^{\infty} x(t) y \left( \frac{t - b_1}{a_1} \right) \left( \frac{t - b_2}{a_2} \right) dt \]  

(EQ 9)

whereas when stochastic signals are being used then expectations are utilised.

Again as exhibited in [7], wavelets contain useful properties from higher order statistics and wavelet theory, one of the important properties and the one that the following algorithm relies on is that the third order wavelet for a gaussian (or symmetrically distributed) random process is zero.

Also it can be noted that as one of the special cases of the third order wavelet when \( a_1 = a_2 = 1 \), \( \tau_1 = -b_1 \) and \( \tau_2 = -b_2 \) then \( W_{xy}^3 (a_1, b_1; a_2, b_2) \) is analogous to the third order cumulant.

3 Estimation Algorithm

A novel algorithm for time delay and scale estimation in unknown gaussian spatially correlated noise has been developed based on wavelet theory. In addition to providing an estimate of either the difference in the time of arrival of a signal at two measurement sensors or the time of arrival of a reflected signal, the dilatory factor can also be extracted hence providing possible estimates of speed of a target or damping factor of a transient for example.

This algorithm demonstrates in particular the application of third order wavelets and their ability of suppressing additive gaussian noise and extraction of multiscale non-gaussian signals. It is this multiscale feature of wavelets which enable previous algorithms to be improved for increased flexibility of compounded wideband signal models.

This initial algorithm operates on a correlation structure and is a non-parametric method, however to obtain higher resolution parametric methods can be employed. To obtain this higher resolution a prototype parametric algorithm based on third order wavelets has also been developed and is currently undergoing performance and estimation testing.

A two sensor case is considered where the signals are as defined in (EQ 7) and (EQ 8) for which the third order wavelet given below is used for the estimation algorithm

\[ W_{xx}^3 (a_1, b_1; a_2, b_2) = \]

\[ = \frac{1}{\sqrt{a_1 a_2}} \int_{-\infty}^{\infty} y(t) x \left( \frac{t - b_1}{a_1} \right) \left( \frac{t - b_2}{a_2} \right) dt \]  

(EQ 10)

for the estimation of the parameters, these are then located by finding

\[ \max_{a_1, b_1, a_2, b_2} \left| W_{xx}^3 \right| \]  

(EQ 11)

which implies finding the values of the parameters \( a_1, b_1, a_2, b_2, \) where the magnitude of the third order wavelet is a maximum, from which the delay, \( D \) and the dilution, \( A \) can be estimated from either \( a_1, b_1 \) or \( a_2, b_2 \). \( D \) and \( A \) are calculated using

\[ A = a_1 = a_2 \]  

(EQ 12)

\[ D = b_1 = b_2 \]  

(EQ 13)

which are obtained from the relationships given in [7] and are also given here for convenience.

If \( y(t), x(t), z(t) \) maps to \( W_{xy}^3 (a_1, b_1; a_2, b_2) \) then

\[ y \left( \frac{t-D}{A} \right), x(t), z(t) \text{ maps to } W_{xy}^3 \left( \frac{a_1}{A}, \frac{b_1-D}{A}; \frac{a_2}{A}, \frac{b_2-D}{A} \right) \]

It appears apparent that there is a redundancy when calculating \( W_{xy}^3 \) for this scenario as the estimated parameters can be derived from either \( a_1, b_1 \) or \( a_2, b_2 \), therefore only a 2D plane needs to be calculated when estimating \( D \) and \( A \) thus increasing the computational efficiency of the algorithm. However this redundancy can be utilised directly in a 3 sensor case where the third sensor has the property

\[ z(n) = s \left( \frac{n - \tau}{B} \right) + w_3(n) \]  

(EQ 14)

where \( w_3(n) \) may be correlated with \( w_1(n) \) and/or \( w_2(n) \).
therefore using the third order wavelet
\[
W_{xyz}^3 (a_1, b_1; a_2, b_2) = \frac{1}{\sqrt{a_1 a_2}} \int_{-\infty}^{\infty} x(t) \frac{1}{a_1} \left( \frac{t - b_1}{a_1} \right) \frac{1}{a_2} \left( \frac{t - b_2}{a_2} \right) dt
\]
(EQ 15)

and estimating using
\[
\max_{a_1, b_1, a_2, b_2} |W_{xyz}^3|
\]
(EQ 16)

the parameters \(D, A\) can be obtained from \(a_1, b_1\) and \(\tau, B\) from \(a_2, b_2\) therefore highlighting the advantages of using the general theory of wavelets. The calculation of \(D, A, \tau\) and \(B\) from \(a_1, b_1, a_2\) and \(b_2\) in this case is performed by
\[
D = \frac{b_1}{a_1}
\]
(EQ 17)
\[
A = \frac{1}{a_1}
\]
(EQ 18)
\[
\tau = \frac{b_2}{a_2}
\]
(EQ 19)
\[
B = \frac{1}{a_2}
\]
(EQ 20)

which are derived from the property given in [7].

If \(y(t), x(t), x(t)\) maps to \(W_{xyz}^3 (a_1, b_1; a_2, b_2)\) then
\[x(t), \frac{1}{\sqrt{A}} \frac{1}{\sqrt{B}} y(t - D), \frac{1}{\sqrt{B}} (t - \tau)\] maps to
\[W_{xyz}^3 (a_1 A, b_1 + a_1 D, a_2 B, b_2 + a_2 \tau)\]

4 Results

In the first set of results we intend to show the superior performance of the developed third order wavelet based estimation algorithm over conventional methods. The scenario consists of two sensors with the desired signal \(s(n)\) as a negative exponentially distributed signal and the first noise signal made up of a unit variance normally distributed signal. The second sensor’s noise was obtained by passing \(w_1(n)\) through another MA filter with coefficients \([1 2 3 4 3 2 1]\). The delay is 8 samples and the dilution factor is 0.5.

Figure 1 shows the result from using the wavelet transform and as can easily be seen, no definite maximum appears due to the contribution of the correlated noise. Figure 2 shows the magnitude of the third order cumulant which provides no suggestion of a possible solution. This may be attributed to the fact that the cumulant method is estimating solely for translational parameters between similar signals hence the dilation factor causes the signal to be virtually unlocatable. Finally for this experiment figure 3 shows the third order wavelet which is represented by a 2D grid of 2D parameter graphs thus representing the 4D wavelet. Each graph has the translational parameters for their axis and each graphs is for a discrete scale value. In this case it is clearly seen that a maximum appears at \((0.5, 0.5, 0.5, 0.5)\) which provides the correct estimation of the delay and scale parameters.

In figure 4 the three sensor case is considered with the same signals used in the first scenario for the first two sensors and the noise for the third sensor is obtained by passing \(w_1(n)\) through another MA filter with the coefficients \([-3 -2 -1 0 1 2 3]\) and the signal has a dilation of 0.25 and a delay of 12 samples. From figure 4 it can be seen that the maximum occurs at \((2, 4, 4, -3)\) which corresponds to the correct estimation for the true delay and scaling parameters.

5 Conclusion

This paper has introduced a unified theory based on wavelets and higher order statistics and indicated some of the synergistic benefits of this marriage. This is supported by the application of wavelets to a common problem encountered in signal processing, that of time delay / scale estimation, and how it can be used in a broader case of signal models. Comparisons with conventional methods have also been given.

6 References

Figure 1: Wavelet Transform

Figure 2: Third Order Cumulant

Figure 3: Third Order Wavelant - 2 sensor case

Figure 4: Third Order Wavelant - 3 sensor case
QUADRATIC IDENTIFICATION USING BISPECTRUM

P.RUIZ*, O.MICHEL**

*École Supérieure de Mécanique de Marseille - Techniphone Développement, Technopôle de Château Gombert, F-13651 MARSEILLE Cedex 20

** ENS Lyon, 46 rue d'Italie, F-69364 LYON Cedex 07

Abstract: We study here the accuracy of the bispectrum for identifying quadratic components in a signal. Our purpose is not only to detect those quadratic components, but to quantify their magnitude and separate the quadratic contribution from the linear one. We apply the described algorithm in the separation of chaotic signals versus linear filtered noise.

1. Introduction

System identification supposes modelisation of the system, and identification of the parameters of the model. Once the linear model has been largely explored, the natural following step consists in identifying the parameters of a both linear and quadratic model. Bispectrum estimation of a signal enables the extraction of quadratic components. We use it here for identifying our model.

Such a tool can also be used for separating deterministic non linear signals from some classes of random signals. The first part is devoted to the explanation of the model, and the expansion of the different terms of its bispectrum. Such an expansion enables the separation of the quadratic terms from the linear ones.

2. The model

We are interesting in identifying the following model:

\[(\alpha + \sum_{i=1}^{p} a_i \cos(\omega_i t + \phi_i))^2 + \beta + \sum_{i=1}^{q} b_i \cos(\xi_i t + \psi_i)\]

(1)

where the quantities \( p, q, (a_i), (b_i), (\omega_i), (\xi_i), \) the offsets \( \alpha \) and \( \beta, \) the phases \( \phi_i \) and \( \psi_i \) are the unknowns. That model corresponds to a general linear-quadratic model. It corresponds to an extension of the linear model and a restriction of the Volterra model.

(1) can be rewritten as follow:

\[(1) = \alpha^2 + \sum_{i=1}^{p} a_i^2 \cos^2(\omega_i t) + 2\alpha \sum_{i=1}^{p} a_i \cos(\omega_i t) + \frac{1}{2} \sum_{i \neq j} a_ia_j \left[ \cos((\omega_i + \omega_j)t + \phi_i + \phi_j) + \cos((\omega_i - \omega_j)t + \phi_i - \phi_j) \right] + \beta + \sum_{i=1}^{q} b_i \cos(\xi_i t + \psi_i)\]

From the trigonometric identities, one may obtain the Fourier components:

\[\cos^2(\omega t) = \frac{1}{2} \left( \cos(2\omega t) + 1 \right)\]

we extract the components of the Fourier transform of the signal:

\[X(0) = (\beta + \sum_{i=1}^{q} b_i^2)N\]

\[X(\omega_0) = \alpha a_0 e^{i\phi_0} + \sum_{i=1}^{q} a_i e^{i\phi_i} e^{i\psi_i} D_0^+ \cup D_0^-\]

where \( D_0^+ \) and \( D_0^- \) are represented above:

\[D_0^+ : \omega_0 + \omega_j = \omega_0\]

\[D_0^- : \omega_0 - \omega_j = \omega_0\]

\[D_0 \] corresponds to an interquadratic line, and represents the set of the spectral contributions to \( \omega_0 \)

We obtain an estimate of the bispectrogram:

1113
For several realisations, an estimation of B will be:

\[
B(\omega_k, \omega_l) = \frac{1}{N} \sum_{i=1}^{N} B_i(\omega_k, \omega_l) \quad (3)
\]

where

\[
B_i(\omega_k, \omega_l) = P_i(\omega) \cdot P_i^*(\omega + \omega_k - \omega_l)
\]

From the development of (2), and taking the expected value, there will only remain the terms which are such that:

\[ E(\omega) \neq 0 \]

For instance, if we look at the first term of the expected value of (2):

\[
E[\alpha^2 \sum a_k \exp(i\phi_{k+l}(t)+\phi_{k+l}(t-\omega_1))]
\]

The last term of the previous expression converges towards zero.

As an example, if we examine now the following term, resulting from (2):

\[
E[\alpha^2 \sum a_k \exp(i\phi_{k+l}(t)+\phi_{k+l}(t))]
\]

The value in the first point, which represents the energy of the signal, is omitted; we extract the magnitudes of the components \( \omega_1 = 5, \omega_2 = 7 \) (main peaks corresponding to peaks 6 and 8); these peaks exist because \( B(\omega_1, \omega_2) \) and \( B(\omega_2, \omega_1) \) are significant. The peak in channel 3 is related to \( \omega_2 - \omega_1 \). The peak in channel 11 corresponds to frequency 10, because \( B(10,10) \) exists. The peak in channel 13 exists because \( B(12,2) \) exists (\( \omega_1 + \omega_2 \) is coupled with \( \omega_1 - \omega_2 \)). If we examine now the spectrum (read in the first line of the bispectrum...)

The spectrum lines present on the frequencies \( \omega_1 = 5, \omega_2 = 7 \) (main peaks), \( \omega_1 + \omega_2, \omega_1 - \omega_2, 2\omega_1, 2\omega_2, \xi_1 \). The frequencies are the same as in the bispectrum, except the peak corresponding to \( 2\omega_2 \), and the linear component in channel 21, corresponding to \( \xi_1 \). The
separation between linear and quadratic components is achieved as follows:
two frequencies are in the spectrum, not in the bispectrum: \( f_1 = 2\omega_2 \), \( f_2 = \xi_1 \). Those are linear components, noticed as \( f_1 \) and \( f_2 \).
We study now the magnitudes of the components.
If we consider the previous model:
\[
X(2\omega) = \frac{N_{\omega}^2}{4} \\
X(\omega - \omega) = \frac{N_{\omega_1}^2}{2} \\
X(\omega + \omega) = -\frac{N_{\omega_1}^2}{2}
\]
If we estimate the biperiodogram (averaged or not):
\[
B(\omega_1, \omega_2)/B(0,0) = \frac{\alpha^2 a_1^3}{4(\alpha^2 + \sum a_1^2)}
\]
\[
B(0, \omega_2)/B(0,0) = \frac{\alpha^2 a_1^2}{(\alpha^2 + \sum a_1^2)}
\]
\[
B(0, 2\omega)/B(0,0) = \frac{4}{16(\alpha^2 + \sum a_1^2)}
\]
\[
B(\omega_1, \omega_2)/B(0,0) = \frac{\alpha^2 a_1^2}{2(\alpha^2 + \sum a_1^2)}
\]
\( \alpha \) and the \( a_j \) can be obtained from the \( B(\omega_1, \omega_2) \).
If we study \( f_1 \), we see that \( f_1/2 \) exists in the bispectrum. Moreover, \( 2^nf_1 \) does not exist: we conclude that some part of \( f_1 \) is related to \( f_1/2 \). In order to know if \( f_1 \) contains a linear part, we compare \( B(0,f_1) \) and \( B(0,f_1/2) \). If \( f_1 \) is only related to the quadratic component \( f_1/2 \), then
\[
\frac{B(0, f_1)}{B(0, f_1/2)} = \frac{1}{4}
\]
We obtain that result.
On the other way, if we examine now the frequencies \( \xi_1 \) and \( \xi_1/2 \),
\[
\frac{B(0, \xi_1)}{B(0, \xi_1/2)} = \frac{1}{4}
\]
Moreover, \( \xi_1/2 \) is related to \( \xi_1/4 \) and
\[
\frac{B(0, \xi_1/2)}{B(0, \xi_1/4)} = \frac{1}{4}
\]
Thus, \( \xi_1 \) corresponds to a linear component.
4.2 Chaos versus noise
Quadratic identification can allow separation between chaotic signals and stochastic ones. Chaotic signals can appear in the output of some non linear filter. Some
5. Conclusion
We have examined how the algorithm can separate the linear contributions from the quadratic ones, and identify the quadratic magnitudes.

Such an algorithm enables a more exhaustive interpretation of the bispectrum, and a more exhaustive interpretation of non-linear dynamics. It enables too the separation chaos versus noise.

References
Cost Functions For Motion Estimation Based On Hos

Elisa Sávor, Antoni Gasull, and Javier R. Fonollosa

Universitat Politècnica de Catalunya
Department of Signal Theory and Communications
Apa. 30.002
08071 Barcelona,
SPAIN
Tel: 34-3-4016432, Fax: 34-3-4016447 e-mail: elisa@tsd.upc.es

Abstract. We propose a class of cost functions based on Higher-Order Statistics (HOS), to estimate the displacement vector between consecutive image frames. In case the image sequence is severely corrupted by additive Gaussian noise of unknown covariance, using higher than second order statistics is appropriate since cumulants of Gaussian noise are zero. To obtain consistent cumulant estimates we need several records of the same sequence, which is not generally possible. Nevertheless, previous frames, where the registration problem has already been solved, can be used to obtain asymptotically unbiased estimates. The objective of this paper is to introduce HOS-based cost functions capable of estimate motion even for small regions or blocks. A recursive version of the algorithm is also derived. Simulation results are presented and compared to those obtained from the mean kurtosis and the mean square error of the displaced frame difference.

1. Introduction

There are some situations where motion between frames has to be estimated in the presence of noise. For example, in motion estimation or in motion compensation applications, like dealing with images from surveillance cameras (which quality is often very poor) or medical images such as echographies with speckle noise. In such cases most existing methods do not work properly and more robust techniques are necessary. On the other hand, noise can be realistically described as a colored Gaussian process. In such circumstances Higher-Order Statistics, may offer some advantages since cumulants of Gaussian processes are asymptotically zero.

HOS-based methods have already begun to be used in motion estimation problems. In [1] several algorithms are developed based on a parametric cumulant method, a cumulant-matching method and a mean kurtosis error criterion. In [2] the displacement vector is obtained by maximizing a third-order statistics criterion.

In this paper we propose an alternative criterion that also exploits higher-order statistics. However, our goal is to obtain a low variance cost function to reduce the problems associated with the estimation of HOS for small blocks of data. Our method is based on an adaptive algorithm approach that was proposed in [3] for the estimation of fourth-order cumulants. The motivation behind this approach is to use previous frames and previously estimated displacements.

The paper is organized as follows. The problem formulation is introduced in Section 2. In Section 3, the new cost function based on the adaptive estimator is derived. A recursive version of the cost function is presented in Section 4. Simulation results are provided in Section 5 and Section 6 is devoted to conclusions and final remarks.

2. Problem Formulation

The problem of motion estimation can be stated as follows: "given an image sequence, compute a representation of the motion field that best aligns pixels in one frame of the sequence with those in the next" [4]. This is formulated as:

\[ s_{k-1}(m) = f_{k-1}(m) + n_{k-1}(m) \]

\[ g_k(m) = f_k(m) + n_k(m) = f_{k-1}(m - d_k^0(m)) + n_k(m), \] (1)

where \( m = (m,n) \) denotes spatial image position of a point, \( g_k(m) \) and \( g_{k-1}(m) \) are observed image intensities at instant \( k \) and \( k-1 \), respectively, \( f_k(m) \) and \( f_{k-1}(m) \) are noise-free frames, \( n_k(m) \) and \( n_{k-1}(m) \) are assumed to be stationary, zero-mean image Gaussian noise sequences with unknown covariance, and \( d_k^0(m) \) is the displacement vector of the object during the time interval \([k-1,k]\). The noise-free signals are assumed to be zero-mean non-Gaussian signals which are statistically independent of the noise. In this formulation the basic assumption is Intensity Constancy:

\[ f_k(m) = f_{k-1}(m - d_k^0(m)) \] (2)

The problem is to estimate \( d_k^0(m) \) from the observation of \( g_k(m) \) and \( g_{k-1}(m) \). The \( DFD_k(d) \), displaced frame difference, is defined as [4]

\[ DFD_k(d) = g_k(m) - g_{k-1}(m - d) \] (3)

where we omit the space dependency of the displacement to simplify notation. In our model we consider that pixels of regions are visible over the entire frame, that is, no occlusions occur, pixels of regions do not move in or out.
Given the previous assumptions we propose a motion estimation scheme that is divided in two steps, the second of which we concentrate our efforts:

**Segmentation.** We work with motion estimation based on a segmentation approach, which is called object-oriented motion estimation [5]. Several problems that are inherent to block-oriented approaches are avoided, i.e., blocking artifacts are drastically reduced and small region sizes are not imposed. However the method becomes increasingly complex as the number of regions undergoing different displacements increases. In this case we can also apply block-oriented methods without loss of generality. Hence, depending on the complexity of the scene we can choose one or the other approach.

**Motion estimation.** For every moving region we are going to estimate motion using a cost function that is minimized for the desired displacement. We are interested in choosing a cost function based on HOS to reduce the effect of colored Gaussian noise. We define a criterion based on the minimization of a HOS function of the displaced frame difference.

3. **HOS-based cost functions**

The classical solution to obtain the displacement vector from the DFD \( \hat{d} \) is the minimum square error [4], defined as:

\[
J_{2k}(d) = E[\text{DFD}_k^2(d)]
\]  
(4)

An estimation of this cost function is given by the sample averaging

\[
\hat{J}_{2k}(d) = \frac{1}{N} \sum_{m \in \Omega_m} \text{DFD}_k^2(d)
\]

(5)

where \( \Omega_m \) denotes the spatial domain that contains the pixels from a given region or block, and \( N \) the number of such pixels.

When the signal is contaminated with correlated noise the cost function above is biased. Nevertheless, this effect is reduced in the HOS domain. Since regions might be symmetrically distributed, fourth-order cumulants are more adequate than third-order, which are zero. The approach in [1] uses the kurtosis of the DFD \( \hat{d} \) which is asymptotically unaffected by correlated Gaussian noise. It is given by:

\[
J_{41k}(d) = E[\text{DFD}_k^4(d)] - 3[E[\text{DFD}_k^2(d)]]^2
\]

(6)

This cost function needs to be minimized when the region is characterized by a positive kurtosis and maximized when the kurtosis is negative. This criterion had been proposed by Tugnait [6] to estimate the time delay between two signals as an extension to the second-order cost function. It is estimated from

\[
\hat{J}_{41k}(d) = \frac{1}{N} \sum_{m \in \Omega_m} \text{DFD}_k^4(d) - 3 \left( \frac{1}{N} \sum_{m \in \Omega_m} \text{DFD}_k^2(d) \right)^2
\]

(7)

**Modified Kurtosis of the DFD.** The kurtosis is a very sensitive measure and we need long data records to avoid low variance estimates. On the other hand, image information is repeated along the sequence as it is established in eq.(2). This redundancy may be used to obtain better estimates of the higher-order statistics of the region, reducing the effect of additive noise. Ambland et al. [3], proposed an adaptive scheme for the estimation of fourth-order cumulants for transient detection. It was proven for the case of i.i.d. random variables that the estimator is asymptotically unbiased. The kurtosis of the displaced frame difference at time \( k \) becomes:

\[
\hat{K}_k(\text{DFD}(d)) = \hat{K}_{k-1}(\text{DFD}(d)) + \gamma [\hat{K}(\text{DFD}(d)) - \hat{K}_{k-1}(\text{DFD}(d))]
\]

(8)

where \( \hat{K}(\text{DFD}(d)) \) is the "instantaneous kurtosis" given by,

\[
\hat{K}(\text{DFD}(d)) = \frac{1}{N} \sum_{m \in \Omega_m} \text{DFD}_k^4(d)
\]

\[
- 3\left( \frac{1}{N} \sum_{m \in \Omega_m} \text{DFD}_k^2(d) \right) \hat{E}_{k-1}(\text{DFD}^2(d))
\]

(9)

where

\[
\hat{E}_{k}(\text{DFD}^2(d)) = \hat{E}_{k-1}(\text{DFD}^2(d))
\]

\[
+ \mu \left( \frac{1}{N} \sum_{m \in \Omega_m} \text{DFD}_k^2(d) \hat{E}_{k-1}(\text{DFD}^2(d)) \right)
\]

(10)

\( \mu \) and \( \gamma \) are forgetting factors that adapt the estimation in slowly changing conditions. The sub index \( k \) has been suppressed in the DFDs where previous frames are involved. This adaptive scheme leads to consistent estimates of the kurtosis of the DFD. However we are interested in obtaining the displacement vector at time \( k \) and this only depends on the instantaneous kurtosis, which involves eq.(9) and eq.(10). The cost function in eq.(5) is the update term in eq.(10). For low SNR it exhibits a local minimum at \( d = d_k^0 \). In the following time iteration, a \( d_k^0 \) shift in \( \hat{E}_{k-1}(\text{DFD}^2(d)) \), causes \( \hat{K}(\text{DFD}(d)) \) to have a maximum at \( d = 0 \) and a minimum at \( d = d_k^0 \) independently of the sign of the kurtosis. Furthermore, this behavior is emphasized normalizing by the square of the current variance. This can be shown for any signal distribution and noise level. Thus, we propose the following cost function:

\[
\hat{J}_{42k}(d) = \frac{1}{\left( \frac{1}{N} \sum_{m \in \Omega_m} \text{DFD}_k^2(d) \right)^2} \cdot \left( \frac{1}{N} \sum_{m \in \Omega_m} \text{DFD}_k^4(d) - 3 \left( \frac{1}{N} \sum_{m \in \Omega_m} \text{DFD}_k^2(d) \right) \hat{E}_{k-1}(\text{DFD}^2(d)) \right)
\]

(11)

For \( \mu = 1 \) just the previous DFD intervenes. For small \( \mu \) past frames are more significant than the previous one.
An alternative estimation must be provided when only two frames are available. In this case we should use the kurtosis as defined in eq.(7). Nevertheless, we have found that using eq.(11) with:

\[
\hat{c}_{[k]}(DFD^2(d)) = \frac{1}{N} \sum_{m \in \Omega_n} [f_k(m) - f_k(m-d)]^2
\]  

(12)

the resulting cost function displays a behavior similar to the one using many images, yielding to better estimates of the displacement, as it will be seen in the examples.

4. Recursive estimation of the displacement

In the previous section the estimated displacement was taken from an exhaustive search of the displacement that provided the absolute minimum or maximum of a given cost function. Recursive estimation algorithms aim to reduce the computational load, specially for sub-pixel displacements, since they use a priori information on the location of regions. Thus, given an i-th estimate of displacement, we obtain the (i+1)-th estimate such that, the value of the cost function resulting from the (i+1)th estimate is lower (or higher) than that used in the ith. The gradient search procedure is [4]

\[
d^i = d^{i-1} - e \nabla J(d^{i-1})/||\nabla J(d^{i-1})||
\]  

(13)

The recursive estimation of \(J_{42k}(d)\) for \(\mu = 1\) becomes:

\[
d^i - d^{i-1} = \frac{1}{N} \sum_{m \in \Omega_n} DFD_k^2(d^{i-1}) \cdot 2
\]

\[
[4 \frac{1}{N} \sum_{m \in \Omega_m} DFD_k^3(d^{i-1})V_{m+\xi,\mu}(m+d_k^o,d^{i-1}]
\]

\[
+ \frac{1}{N} \sum_{m \in \Omega_m} DFD_k^2(d^{i-1}) \frac{1}{N} \sum_{m \in \Omega_m} DFD_k(d^{i-1})V_{m+\xi,\mu}(m+d_k^o,d^{i-1})
\]

\[
- \frac{1}{N} \sum_{m \in \Omega_m} DFD_k^2(d^{i-1}) \frac{1}{N} \sum_{m \in \Omega_m} DFD_k(d^{i-1})V_{m+\xi,\mu}(m+d_k^o,d^{i-1})
\]

\[
- 4 \frac{1}{N} \sum_{m \in \Omega_m} DFD_k(d^{i-1})V_{m+\xi,\mu}(m+d_k^o,d^{i-1}) \cdot
\]

\[
\frac{1}{N} \sum_{m \in \Omega_m} DFD_k^2(d^{i-1}) \left[ \frac{1}{N} \sum_{m \in \Omega_m} DFD_k^2(d^{i-1}) \right]^{-1}
\]

(14)

where the gradients of the DFDs are expressed as a function of the gradient of the intensity at time \(k\) considering the previous and current displacements of the region. This expression is rather complex, since we have obtained the gradient of a fractional function.

5. Simulation Results

Example 1. We compare \(J_{42k}(d)\) for \(\mu = 0.89\) and \(J_{2k}(d)\) for an uncorrelated 2-D uniformly distributed rectangular object given 10 realizations of a sequence of 7 synthetic noise free images. The object was previously segmented and was moving (3,1) pixels per frame. Colored Gaussian noise was generated from a first order AR model and added to the sequence. For this size and signal distribution, the cost function \(J_{42k}(d)\) did not work at all. Table 1 shows the percentage errors to reach the final position for different sizes and SNR.

Example 2. We demonstrate the performance of the cost function using the estimation in eq. (12). We reproduce here, the example that was used in [5]. Time delay between two 1-D signals has to be estimated. The signal is given by \(f_k(m) = W_k(m) \cdot \beta_k(m)\) where \(W_k(m)\) is Laplace i.i.d and \(\beta_k(m)\) is a Bernoulli process. The noise sources are spatially correlated, \(n_k(m)\) is generated first, whereas \(\frac{n_k(m)}{\sigma_n}\) is computed from

\[
n_k(m) = \sum_{i=0}^{10} b(i) n_{k-i}(m+i)
\]

where \(b(i)\) takes the values \(0.2, 0.4, 0.6, 0.8, 1, 1, 0.7, 0.5, 0.3, 0.1\). It was seen that \(J_{42k}(d)\) yields the true time delay, \(d_k^o = 16\), whereas \(J_{2k}(d)\) does not. \(J_{42k}(d)\) can also be used to obtain the true delay time for SNR = -5 dB, and signals of length 2000. Figures 1 and 2 show the mean plus/minus standard deviation of the two HOS-based cost functions for 10 realizations. Figures 3 and 4 show the two cost functions for a signal of length 64 for 10 realizations. We obtained 90% of error for \(J_{42k}(d)\) versus 20% of error for \(J_{2k}(d)\). The new cost function outperformed the unnormalized kurtosis.

Example 3. We compare the recursive schemes of \(J_{42k}(d)\) and \(J_{2k}(d)\) [4], for a situation similar to example 1. The object in this case, is a 2-D uniformly distributed rectangular object of size 16x16 which pixels are correlated given a first order AR model. We obtain 10 realizations of a sequence of 3 images for three different SNR. The object is moving (3,1) pixels per frame. For the first pair of images we estimate the displacement using eq.(12) in the recursive scheme. The initial position was \((1,0)\) and \(\varepsilon\) was taken 0.2. The percentages of error after 20 iterations are given in Table 2. The second-order cost function often leads to a local minimum at \((0,0)\). For the second pair of images the correct displacement was given and we checked if the cost functions were able to track the minimum, see Table 3. For very low SNR, \(J_{42k}(d)\) has a 50% of error whereas \(J_{2k}(d)\) always failed.

6. Conclusions and final remarks

In case image frames are corrupted with colored Gaussian noise, HOS-based cost functions are specially adequate. However, cumulant estimates exhibit high variance for short length signals, as it is the case for regions or blocks of pixels. We have derived a cost function that uses previous frames to overcome this problem. We have also found a HOS-based cost function that outperforms the mean kurtosis criterion when only two signals are available. A recursive version to estimate the displacement has also been presented. The results of this paper show the possibilities of
HOS for motion estimation even for reasonable small regions.

References


<table>
<thead>
<tr>
<th>size</th>
<th>32x32</th>
<th>16x16</th>
<th>32x32</th>
<th>16x16</th>
<th>32x32</th>
<th>16x16</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>0 dB</td>
<td>0 dB</td>
<td>-2 dB</td>
<td>-2 dB</td>
<td>-5 dB</td>
<td>-5 dB</td>
</tr>
<tr>
<td>$J_{41k}(d)$</td>
<td>0%</td>
<td>10%</td>
<td>10%</td>
<td>30%</td>
<td>80%</td>
<td>90%</td>
</tr>
<tr>
<td>$J_{2k}(d)$</td>
<td>0%</td>
<td>20%</td>
<td>40%</td>
<td>40%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 1 Percentage of error for example 1

<table>
<thead>
<tr>
<th>SNR</th>
<th>0 dB</th>
<th>-2 dB</th>
<th>-5 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_{41k}(d)$</td>
<td>10%</td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td>$J_{2k}(d)$</td>
<td>90%</td>
<td>90%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2 Percentage of error to reach minimum, example 3

<table>
<thead>
<tr>
<th>SNR</th>
<th>0 dB</th>
<th>-2 dB</th>
<th>-5 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_{41k}(d)$</td>
<td>0%</td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td>$J_{2k}(d)$</td>
<td>10%</td>
<td>30%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 3 Percentage of error in tracking, example 3

Figure 1 $J_{41k}(d)$ for a 1-D signal of length 2000

Figure 2 $J_{42k}(d)$ for a 1-D signal of length 2000

Figure 3 $J_{41k}(d)$ for a 1-D signal of length 64

Figure 4 $J_{42k}(d)$ for a 1-D signal of length 64
Bispectral Analysis of Periodic Signals in Noise: Theory, Interpretation and Condition Monitoring Applications*

Justin W.A. FACKRELL† Paul R. WHITE, Joe K. HAMMOND, Roger J. PINNINGTON
and Alan T. PARSONS †

Institute of Sound and Vibration Research, University of Southampton, SO9 5NH, UK.
Tel/Fax +44 (0)3 593221/593033.
†Defence Research Agency, Portland, Dorset, UK.

Abstract Although many important discoveries have been made recently in the field of Higher Order Statistics (HOS), the number of successful applications of the theory remains small. This is partly due to the difficulty of interpreting the Higher Order Spectra of experimental data. This paper describes recent work investigating the origins of some commonly-seen bispectral features, with particular attention paid to signals containing periodic components and noise. Theoretical results are derived, and synthetic signals are analysed, to investigate the forms that the bispectrum and the bicoherence will take for the types of signals commonly encountered in machine condition monitoring applications. These results are then used as aide to help interpret the bicoherences of real experimental condition monitoring data. In particular it is shown that the bicoherence takes a particular form for idealised periodic signals, and that deviations of this form can be due to a variety of practical factors. The results presented here are based on acoustic and vibration measurements, but generalise to many other types of signals.

1 Theory

The bispectrum can be estimated by averaging periodograms [1], in a way similar to the Welch method for power spectrum estimation. However, the variance of this bispectral estimate is not uniform across frequencies, and second order (i.e. energy) properties of the signal can dominate the estimate. In order to make the variance more uniform across frequencies a form of normalisation is often used. Most research to date has concentrated on one of two normalised bispectra, called the 'skewness function' [1] or the 'bicoherence function' [2] respectively. The bicoherence has been used in this work because it is bounded between 0 and 1 (which can be shown from the Cauchy inequality), whereas the skewness function is not. The discrete bicoherence function is computed from the signal Discrete Fourier Transform (DFT) \( X(k) \) as [2]

\[
\hat{b}^2(k, l) = \frac{1}{K} \sum_m \frac{X_m(k)X_m(l)X_m(k + l)}{K} \frac{1}{K} \sum_k \frac{X_m(k)X_m(l)}{K} \frac{1}{K} \sum_k \frac{X_m(k + l)}{K} \]

where the signal is divided into \( K \) non-overlapping segments, \( k \) and \( l \) are the discrete frequency indices, and \( \sum_k \equiv \sum_{k=1}^K \). The skewness function has a denominator that is the triple product of the power spectra \( S(k)S(l)S(k + l) \) and often shows similar features to the bicoherence, so much of what will be said below applies to the skewness function also.

Since many machines contain rotating parts, the acoustic and vibration fields they generate are often periodic (or quasi-periodic). It is well known that such a periodic signal can be thought of as the convolution of a system impulse response function with a train of impulses. Now the (continuous) power spectrum of the impulse train will itself be an impulse train (in frequency)

\[
I(f) = \frac{1}{\Delta} \sum_{q=-\infty}^{\infty} \delta(f - q/\Delta) \]

where \( f \) is continuous frequency, and it follows that the continuous (unnormalised) bispectrum \( I(f)I^*(f + g) \) of such a signal is a 2-dimensional impulse train, with the form

\[
B_i(f, g) = \frac{1}{\Delta^3} \sum_{q_1=-\infty}^{\infty} \sum_{q_2=-\infty}^{\infty} \delta(f - q_1/\Delta, g - q_2/\Delta) \]

It will now be shown that the bicoherence of this impulse train signal can, under certain restrictions, also take this bed of nails form.

The form of the bicoherence of an impulse train signal contaminated with additive non-skewed noise can be written as [4]

\[
\hat{b}^2(f, g) = \frac{|B_i(f, g)|^2}{|B_i(f, g)|^2 + E[M(f, g)]} = \frac{1}{1 + \frac{E[M(f, g)]}{|B_i(f, g)|^2}} \]

\[ (3) \]
where $B_I(f, g)$ is the bispectrum of the impulse train and $M(f, g)$ represents a combination of terms arising from the signal energy, including cross signal and noise terms and noise energy terms.

If the impulsive signal has a non-zero power spectrum (due to noise and rounding errors) in the frequency range of interest, then the bicoherence can be shown to be

$$b^2(f, g) = \sum_{q_1 = -\infty}^{\infty} \sum_{q_2 = -\infty}^{\infty} \delta(f - \frac{q_1}{\Delta}, g - \frac{q_2}{\Delta})$$  \hspace{1cm} (4)

When working with discrete time signals windowing effects also come into play. If it is assumed that there are an integer number of signal periods in each analysis frame, then smoothing windows will be unnecessary, and there will be no windowing effects and the DFT of the signal will take the exact form of eq.(2), with zeroes between the impulses (in the frequency domain). The discrete bicoherence $b^2(k, l)$ of this impulse train signal takes a value of unity at frequencies $k$, $l$ where the DFT's $I(k)$, $I(l)$ and $I(k + l)$ have peaks, and a value of 0/0 for all other $k$ and $l$. This is important to bear in mind when constructing simulation signals for bispectral analysis, since large bispectral content can arise simply due to low signal energy between harmonic peaks. However this is not usually the case in practice, as the effects of windowing and estimation will result in $I(k)$ being non-zero between the frequency impulses, so the bicoherence of a periodic signal also takes the bed of nails form (see Figure 1).

---

2 Simulations and Experiments

Some interesting properties of the bicoherence have been investigated using simulated signals. In the first simulation, a periodic impulse train with low level Gaussian noise was sampled at two slightly different sampling frequencies. The bicoherences of these signals (see Figure 2) are very similar in the Inner Triangle (IT) but differ greatly in the Outer Triangle (OT). In the first case, the sampling rate is an integer multiple of the signal fundamental ($= 1/\delta$, where $\delta$ is the pulse spacing), so that for $k + l > f_s/2$ (i.e. in the OT region), there exist $(k, l)$ pairs with peaks in the DFT at frequencies $k$, $l$ and $k + l$. If the sample rate is changed slightly, the bicoherence in the IT changes very little, because the DFT still peaks in a regular way (see Figure 3), but for $k + l > f_s/2$ there exist no $(k, l)$ such that there are peaks at $k$, $l$ and $k + l$. Hence the OT for this signal is zero. This is shown diagrammatically in Figure 3.

A simple mechanical system was made to vibrate with a shaker (see Figure 4) which was fed with white noise. The nut fixing the beam to the shaker could be adjusted to make the beam tightly secured to the shaker, or loosely attached. This second mode of operation is analogous to a typical machine fault. The input voltage to the shaker and the acceleration at the end of the beam were recorded. By making these two measurements the 2nd order coherence function could also be calculated. The power spectra of the output acceleration and the co-

---

Figure 1: The bicoherence of a periodic impulse train signal - the 'Bed of Nails'.

Figure 2: The bicoherences of a periodic impulse train signal sampled at two slightly different frequencies.
Figure 3: The DFT’s of an impulse train signal sampled at slightly different frequencies.

Figure 4: Experimental rig for simple beam condition monitoring experiment.

Figure 6: Beam experiment. Coherence between acceleration and shaker input for the two operating conditions.

Figure 7: Bicoherence of acceleration on beam with a) beam tightened, b) beam slightly loosened. (Vertical scale: bicoherence 0-1).

Figure 5: Beam experiment. Power Spectra of acceleration measurements from beam experiment for each of two operating conditions.

Coherence functions between the input and output of the system for the two operating conditions, are shown in Figures 5 and 6. There is no significantly large change in the power spectra, and it would probably be difficult to use the power spectrum to tell whether or not the nut was tightened. The coherence function, however, changes markedly as the nut is loosened. However, two measurement sensors are needed to calculate the coherence, so it may have limited use in condition monitoring. The bicoherence, which requires only a single measurement sensor, also shows a very significant difference between the operating states of the test rig (see Figure 7). When the nut is tightened the bicoherence is close to zero everywhere, but when the nut is loosened the bicoherence content rises at several bifrequencies.
Further experimental results were obtained from a second test rig, schematically shown in Figure 8. The rig consists of an electrically powered air compressor mounted on a perspex model of a ship double hull. The hull model consisted of 36 air-filled compartments. The bicoherences of the hull acceleration at several measurement positions are shown in Figure 10. Each bicoherence has the underlying bed of nails form in the IT region, with little content in the OT. The measurements at sensor 3 were found to have a power spectrum which tailed off at about 300Hz (see Figure 9), much lower than the other sensors. This explains why the bicoherence of the sensor 3 measurement has a diagonal $k+l=300$Hz cut-off. This emphasises the fact that the bicoherence provides supplementary information to second order measures, since power spectral features are often useful in explaining bicoherence features.

3 Conclusions

Features occurring in the bicoherences of machine noise signals have been investigated, and explanations can now be given for features arising from periodic signals. These features have been found in data from real machine condition monitoring signals, and show the potential value of third order measures in single-sensor signal analysis.

4 Acknowledgements

This work was supported by contract 2040/516/csm from DRA Portland under the MOD Strategic Research Programme.

References


MIMO Volterra System Input/Output Relations for Cyclic Higher-Order Statistics

Luciano Iizzo, Antonio Napolitano and Luigi Paura

Dipartimento di Ingegneria Elettronica, Università di Napoli Federico II, via Claudio 21, I-80125 Napoli, Italy, Tel/Fax: +39-81-7683156/7683149, E-Mail izzo@nadis.dis.unina.it

Abstract. Input/output relations in terms of cyclic higher-order statistics are derived with reference to multiinput multoutput Volterra almost-periodically time-variant systems that are excited by time series exhibiting (possibly joint) higher-order cyclostationarity. The relations are expressed in terms of cyclic temporal cross-moment functions, cyclic spectral cross-moment functions, cyclic temporal cross-cumulant functions and cyclic cross-polyspectra. The time-series framework is utilized. As example of application, with reference to filtered $p$th-law devices, a system identification formula based on noisy measurements is stated. Higher-order cyclostationary statistics can provide a powerful tool to single out detection and estimation algorithms that are highly tolerant to noise and interference and can operate properly with signals exhibiting higher-order cyclostationarity rather than second-order cyclostationarity.

1. Introduction

The theory of second-order cyclostationarity (SOCS) [1], which has been developed in recent years, provides a powerful tool to single out signal processing algorithms highly tolerant to noise and interference for such purposes as detection, parameter estimation, and waveform extraction. According to such a theory, a signal exhibits SOCS if and only if there exists a quadratic homogeneous time-invariant transformation that converts into spectral lines the hidden periodicities due to some periodically time-variant operations, such as modulation, sampling, coding and multiplexing.

Very recently, some efforts (see [2-5] and papers referenced therein) have been made to introduce a theory for the higher-order cyclostationary signals, i.e., the signals for which spectral lines can be generated by means of $N$th-order ($N > 2$) homogeneous nonlinear time-invariant transformations. Higher-order cyclostationarity (HOCS) properties are expected to provide a powerful tool to single out highly tolerant to noise and interference estimation and detection algorithms that can operate properly with signals exhibiting HOCS rather than SOCS. In fact, there is a wide class of signals, whose degree of SOCS is low or zero, whereas the degree of HOCS is substantial. This class includes pulse-amplitude-modulated (PAM) signals with less than 50% excess bandwidth, $M$-ary phase-shift-keyed (PSK) signals with $M \geq 4$ and severely bandlimited digital signals.

In the present paper, the Volterra representation with kernels that are almost-periodically time-variant is considered in that it is an attractive model for describing several natural and manmade systems subject to (possibly multiple) temporal periodic intervals. Specifically, after a brief introduction on the joint HOCS, the input/output relations for a multiinput multoutput (MIMO) Volterra almost-periodically time-variant (APTV) system excited by time series exhibiting (possibly joint) HOCS are derived. The relations are expressed in terms of cyclic temporal cross-moment functions, cyclic spectral cross-moment functions, cyclic temporal cross-cumulant functions and cyclic cross-polyspectra. An example of application of the derived relations is provided. Specifically, for a special class of single-input single-output (SISO) Volterra time-invariant systems, referred to as the filtered $p$th-law devices, a system identification formula based on noisy measurements is stated.

2. Joint Higher-Order Cyclostationarity

Let us consider the column vector $\mathbf{z}(t) \triangleq [z_1(t), \ldots, z_N(t)]^T$ whose components are $N$ not necessarily distinct complex-valued time-series. The $N$ time-series exhibit joint cyclostationarity of order $N$ with cycle frequency $\alpha \neq 0$ if at least one of the $N$th-order cyclic temporal cross-moment functions (CTCMFs)

$$R_N^\alpha(\tau)_N \triangleq \langle L_N(\tau, \tau)_N e^{-j2\pi \alpha t} \rangle \triangleq (\prod_{i=1}^N \langle x_i^* (t) x_i(t + \tau) e^{-j2\pi \alpha t} \rangle$$

is not identically zero. In eq. (1), $\tau \triangleq [\tau_1, \ldots, \tau_N]^T$, $(\cdot)$ denotes infinite time averaging and $(\cdot)^*$ represents optional conjugation of the $i$th factor $x_i(t)$ of the $N$th-order lag product waveform $L_N(\tau, \tau)_N$.

The magnitude and phase of the function $R_N^\alpha(\tau)_N$ are the amplitude and phase of the sine-wave component with frequency $\alpha$ contained in the lag product. In the fraction-of-time probability context for time-series that exhibit cyclostationarity [3,6],

$$\langle L_N(t, \tau)_N e^{-j2\pi \alpha t} \rangle \triangleq \sum_{\alpha} \langle L_N(u, \tau)_N e^{-j2\pi \alpha u} \rangle e^{j2\pi \alpha t}$$

1125
\[ R^\alpha_2(\tau) = \sum_\alpha R^\alpha_2(\tau) e^{j2\pi \alpha \tau} \]  

(2)

is called the \(N\)th-order temporal cross-moment function (TCMF) and will be denoted by \(R^\alpha_2(t, \tau_N)\). In eq. (2), the sum ranges over all \(N\)th-order cycle frequencies \(\alpha\) and \((1)\) stands for the almost-periodic-component extraction operator.

The \(N\)-dimensional Fourier transform of the CTCMF, which is called the \(N\)-th-order cyclic spectral cross-moment function (CSCMF), can be written as [2,3]

\[ S^\alpha_2(f) = \mathcal{F}_\alpha[S^\alpha_2(f')_N \delta(f^T 1 - \alpha)], \]

(3)

where \(f \triangleq [f_1, ..., f_N]^T\), 1 is the vector \([1, ..., 1]^T\), \(\delta(\cdot)\) is the Dirac delta function and prime denotes the operator that transforms the vector \(u \triangleq [u_1, ..., u_N]^T\) into \(u' \triangleq [u_1, ..., u_{N-1}]^T\). The function \(S^\alpha_2(f')_N\), referred to as the reduced-dimension CSCMF (RD-CSCMF), can be expressed as the \((N-1)\)-dimensional Fourier transform of

\[ R^\alpha_2(\tau')_N \triangleq \mathcal{F}_\alpha[R^\alpha_2(\tau')_N |_{\tau_N = 0}], \]

(4)

which is the reduced-dimension CTCMF (RD-CTCMF).

Let us note that in general both \(S^\alpha_2(f)_N\) and \(S^\alpha_2(f')_N\) contain products of impulses and, then, are not well-behaved functions. This stems from the fact that in general the function \(R^\alpha_2(\tau')_N\) is not absolutely integrable [2,3]. However, a well-behaved function in the spectral-frequency domain can be introduced starting from the \(N\)-th-order temporal cross-cumulant function (TCCF)

\[ C_2(t, \tau)_N \triangleq \text{cum}\left\{ x_k^{(\ast)}(t + \tau_k), \quad k = 1, ..., N \right\} \triangleq \]

\[ (-1)^N \frac{\partial^N}{\partial \omega_1 \cdots \partial \omega_N} \log_e \left( \exp\left\{ \sum_{k=1}^{N} \omega_k x_k^{(\ast)}(t + \tau_k) \right\} \right) |_{\omega = 0} = \sum \left[ (-1)^p (p - 1)! \prod_{i=1}^{p} R_{\omega_i}(t, \tau_{\mu_i})_{|_{\mu_i = 1}}, \right], \]

(5)

where \(\omega \triangleq [\omega_1, ..., \omega_N]^T\), \(p\) is the set of distinct partitions of \([1, ..., N]\), each constituted by the subsets \(\{\mu_i : i = 1, ..., p\}\), \(|\mu_i|\) is the number of elements in \(\mu_i\), \(x_{\mu_i}\) is the \(|\mu_i|\)-dimensional vector whose indices components are those (possibly conjugate) of \(x\) having indices in \(\mu_i\) and \(\{\gamma\}\) is the set of frequencies of the \(2N\)-variate fraction-of-time probability density function of the real and imaginary parts of the \(N\) complex time-series \(x_i(t)\) [6]. In fact, taking the \(N\)-dimensional Fourier transform of the coefficient

\[ C^\alpha_2(\tau) = \langle C_2(t, \tau)_N e^{-j2\pi \beta t} \rangle \]

(6)

of the Fourier series expansion of the almost-periodic function \(C_2(t, \tau)_N\), which is referred to as the \(N\)-th-order cyclic spectral cross-cumulant function (CSCCF) \(P^\beta_2(f)_N\). It can be written as [2,3]

\[ P^\beta_2(f)_N = \mathcal{F}_\alpha[P^\beta_2(f')_N \delta(f^T 1 - \beta)], \]

(7)

where the \(N\)th-order cyclic cross-polyspectrum (CCP) \(P^\beta_2(f')_N\) is the \((N-1)\)-dimensional Fourier transform of

\[ C^\alpha_2(\tau)_N \triangleq C^\alpha_2(\tau)_{|_{\tau_N = 0}}, \]

(8)

which is the reduced-dimension CTCMF (RD-CTCMF). The cyclic cross-polyspectrum is a well-behaved function under the mild conditions that the time-series \(x_N(t)\) and \(z_i(t + \tau_i)\) \((i = 1, ..., N - 1)\) are asymptotically \((\tau_i \rightarrow \infty)\) independent \((\text{in the fraction-of-time probability sense})\) so that \(C^\alpha_2(\tau)_N \rightarrow 0\) as \(|\tau_i| \rightarrow \infty\) and, moreover, there exists an \(\varepsilon > 0\) such that \(C^\alpha_2(\tau')_N = o(|\tau'|^{-N+1-\varepsilon})\) as \(|\tau'| \rightarrow \infty\).

Finally, we note that in Section 3, for analytical simplicity, the input/output relations for MIMO Volterra APTV systems are first derived in terms of cyclic moments and, then, in terms of cyclic cumulants, by exploiting the relationship between moments and cumulants.

3. MIMO Volterra APTV Input/Output Relations

Let us consider a MIMO Volterra APTV system with \(N\) inputs \(z(t) \triangleq [z_1(t), ..., z_N(t)]^T\) and \(M\) outputs

\[ y_m(t) = \sum_{n=1}^{\infty} \int_{I^{(m)}} \sum_{i^{(m)}} h_{i^{(m)}}(t, 1t + \tau^{(m)}) \]

\[ L \tilde{z}_{i^{(m)}}(t, \tau^{(m)})_{n^{(m)}}, \]

(9)

where \(\tau^{(m)} \triangleq [\tau_1^{(m)}, ..., \tau_m^{(m)}]^T\), \(i^{(m)} \triangleq [i_1^{(m)}, ..., i_m^{(m)}]^T\) is the index vector of the \(n\)th-order terms, \(z_{i^{(m)}}(t) \triangleq [z_{i_1^{(m)}}(t), ..., z_{i_m^{(m)}}(t)]^T\) and the inner summation is over all vectors \(i^{(m)}\) belonging to the set \(I^{(m)} \triangleq \{1, ..., N\}^m\).

The Volterra kernels can be expanded into the Fourier series

\[ h_{i^{(m)}}(t, 1t - \tau^{(m)}) \]

\[ = \sum_{\sigma_{i^{(m)}}} h_{i^{(m)}}(\tau^{(m)}) \exp\{j2\pi \sigma_{i^{(m)}} t\}, \]

(10)

where the sum ranges over all the harmonics of the almost periodic function \(h_{i^{(m)}}(t, 1t - \tau^{(m)})\).

The \(M\)th-order CTCMF of the \(M\) outputs is given by

\[ \mathcal{F}[y_m(t + \tau_m)e^{-j2\pi \lambda t}] = \]

\[ \sum_{n_1, ..., n_M = 1}^{+\infty} \sum_{i^{(1)}} \cdots \sum_{i^{(M)}} \sigma_{i^{(1)}} \cdots \sigma_{i^{(M)}} \exp\left\{ j2\pi \left( \sigma_{i^{(1)}} \cdots \sigma_{i^{(M)}} \right) \tau \right\} \]

1126
where the inner summation is over all vectors \( \sigma_{1}^{(n_{1})} \cdots \sigma_{M}^{(n_{M})} \triangleq [\sigma_{1}^{(n_{1})}, \ldots, \sigma_{M}^{(n_{M})}]^{T} \), the symbol \( \otimes \) denotes \( \mathbb{R}^{m} \)-dimensional convolution with respect to \( \nu^{(m)} \) and the equality has been obtained interchanging the order of the integration and summation operations. Note that optional conjugations in the expression of the lag product waveform have been omitted here without loss of generality. The reduced-dimension version \( \tilde{R}_{\gamma}^{(T)}(\tau_{M}) \) of \( R_{\gamma}^{(T)}(\tau_{M}) \) is obtained setting \( \tau_{M} = 0 \).

The input/output relation in terms of RD-CSMF's can be obtained by Fourier transforming both sides of eq. (11) and accounting for eq. (3). It results that

\[
\mathbb{S}_{\gamma}^{\sigma}(f)^{M} \triangleq \mathbb{S}_{y_{1}, \ldots, y_{M}}(f^{M}) = \sum_{n_{1}, \ldots, n_{M}=1}^{+\infty} \int_{\mathbb{R}^{n_{1}+\cdots+n_{M}}}^{+\infty} \mathbb{S}_{x_{1}, \ldots, x_{M}}(f_{x_{1}, \ldots, x_{M}}^{M}) \sigma_{1}^{(n_{1})} \cdots \sigma_{M}^{(n_{M})} \mathbb{S}_{\lambda}^{\sigma}(f_{x_{1}, \ldots, x_{M}^{M}}^{M}) \mathbb{S}_{\lambda}^{\sigma}(f_{x_{1}, \ldots, x_{M}^{M}}^{M})
\]

where \( H^{(n)(m)^{M}} \) is the \( n_{m} \)-fold Fourier transform of \( \mathbb{S}_{x_{m}}(f) \), and \( x_{m}^{(n)} \) is the \( n_{m} \)-dimensional vector \( [x_{m}(1), \ldots, x_{m}(n_{m})]^{T} \).

Equation (14) leads to a useful system identification formula for the filtered pth-law device (p integer number), which is characterized by the input/output relationship

\[
y(t) = \int_{-\infty}^{+\infty} h(u) x^{p}(t-u) du.
\]

In fact, denoted with \( H(f_{1}) \) the Fourier transform of \( h(u) \), by letting in eq. (14) \( H^{(n)(m)}(f^{(n)}) = \delta_{n_{m},1} \) for \( m = 2, \ldots, N \) and \( H^{(n)(m)}(f^{(n)}) = H(f^{(n)} T_{1}) \delta_{n_{m},1} \) assuming as input the \( N \)-dimensional vector \( x(t) \triangleq [x(t), \ldots, x(t)]^{T} \triangleq \begin{bmatrix} x(t), x_{0}(t) \end{bmatrix}^{T} \), one has

\[
H(f_{1}) = \frac{\mathbb{S}^{\sigma}}{\mathbb{S}^{\sigma}}(f N^{M+1}) \cdot \prod_{m=1}^{M-1} H_{\sigma_{m}^{(n_{m})}} \left( \begin{bmatrix} \lambda^{(n_{m})} T_{1}, \lambda^{(n_{m})} T_{1} \end{bmatrix} \cdot \begin{bmatrix} f_{1} - \sigma_{1}^{(n_{1})} T_{1}, \ldots, \lambda^{(n_{m})} T_{1} \end{bmatrix} \right)
\]

Let us note that, for \( p = 1 \), this identification formula reduces to the linear system identification formula stated in [4]. Moreover, it is worthwhile to underline that no method exploiting only second-order cyclostationarity can be derived to identify nonlinear \((p > 1)\) filtered pth-law devices.

Let us consider now the case of MIMO memoryless PTV systems, which are characterized by the input/output relations

\[
y_{m}(t) = G_{m}[t, x(t)] = \sum_{n_{m}=1}^{+\infty} \frac{1}{\delta_{n_{m},1}} \sum_{i_{1}, \ldots, i_{m}=0}^{+\infty} \frac{\delta_{n_{m}}}{\partial x_{i_{1}}(t) \cdots \partial x_{i_{m}}(t)} L_{x_{i_{1}}(t) \cdots x_{i_{m}}(t)} x_{i_{m}}(t, o)_{n_{m}}
\]

In such a case, the Fourier coefficients of the Volterra kernels are given by

\[
h_{\sigma_{m}^{(n_{m})}}(t) = G_{\sigma_{m}^{(n_{m})}}[t, x(t)] = \mathbb{S}_{\sigma_{m}^{(n_{m})}}(f^{(n)}) \delta_{n_{m},1},
\]

where

\[
G_{\sigma_{m}^{(n_{m})}} \triangleq \frac{1}{\delta_{n_{m},1}} \left\langle \mathbb{S}_{\sigma_{m}^{(n_{m})}}(f^{(n)}) \exp(-j 2 \pi \sigma_{m}(t)) \right\rangle
\]

Therefore, eq. (12) becomes

\[
\mathbb{S}_{\gamma}^{\sigma}(f^{M}) = \sum_{n_{1}, \ldots, n_{M}=1}^{+\infty} \int_{\mathbb{R}^{n_{1}+\cdots+n_{M}}}^{+\infty} \mathbb{S}_{x_{1}, \ldots, x_{M}}\left( f_{x_{1}, \ldots, x_{M}^{M}}^{M} \right) \sigma_{1}^{(n_{1})} \cdots \sigma_{M}^{(n_{M})} \mathbb{S}_{\lambda}^{\sigma}(f_{x_{1}, \ldots, x_{M}^{M}}^{M}) \mathbb{S}_{\lambda}^{\sigma}(f_{x_{1}, \ldots, x_{M}^{M}}^{M})
\]

1127
\[
\sum_{\mathbf{i}^{(m_1)}\ldots\mathbf{i}^{(m_M)}} \left( \prod_{m=1}^{M} G_{\mathbf{i}^{(m_1)}}^{\mathbf{z}^{(m_1)}} \right) \int_{R_{n_1}+\ldots+n_M} \left( \lambda^{(n_1)} (f_1 - \sigma^{(n_1)}) - \lambda^{(n_1)} \right) \cdot \ldots \cdot \lambda^{(n_M)} d\lambda^{(n_1)} \ldots d\lambda^{(n_M)}. \tag{20}
\]

As regards the CTCGF of the M outputs of a MIMO Volterra APVT system, from eq. (5) and eq. (6), accounting for eq. (11) and observing that the subsets \( \mu_i \) are disjoint, it follows that

\[
C_{\mathbf{y}_M}^\beta (\tau)_M \triangleq C_{\mathbf{y}_M}^\beta (\tau)_M = \sum_{n_1,\ldots,n_M=1}^{+\infty} \sum_{i^{(n_1)}\in I^{(n_1)}} \ldots \sum_{i^{(n_M)}\in I^{(n_M)}} \exp \left( j2\pi \left( \sigma^{(n_1)} \ldots \sigma^{(n_M)} \right)^T \right) 
\times C_{\mathbf{z}_i^{(n_1)}}^\beta (\tau) \cdot C_{\mathbf{z}_i^{(n_2)}}^\beta (\tau) \ldots C_{\mathbf{z}_i^{(n_M)}}^\beta (\tau) 
\times h_{\mathbf{z}_i^{(n_1)}}(n_1) \ldots h_{\mathbf{z}_i^{(n_M)}}(n_M) \nonumber |_{\mathbf{u}(n_1) = \mathbf{i}_1} \ldots |_{\mathbf{u}(n_M) = \mathbf{i}_M}, \tag{21}
\]

where

\[
C_{\mathbf{z}_i^{(n_1)}}^\beta (\tau) = \frac{\sum \left\{ L_{\mathbf{z}_i^{(n_1)}}(t_1, n_1), \ldots, L_{\mathbf{z}_i^{(n_M)}}(t_M, n_M) e^{-j2\pi \beta t} \right\}}{\mathbf{c}(\mathbf{z}_i^{(n_1)}, \ldots, \mathbf{z}_i^{(n_M)}; \mathbf{z}_i^{(n_1)}, \ldots, \mathbf{z}_i^{(n_M)})}. \tag{22}
\]

Let us note that, since in eq. (11)

\[
R_{\mathbf{z}_i^{(n_1)}}^\beta (\tau) \ldots R_{\mathbf{z}_i^{(n_M)}}^\beta (\tau) = R_{\mathbf{z}_i^{(n_1)}}^\beta (\tau) \ldots R_{\mathbf{z}_i^{(n_M)}}^\beta (\tau) \text{,} \tag{23}
\]

the input/output relations in terms of CTCMF's and CTCGF's are the same (compare eq. (11) with eq. (21)). Moreover, the relationship in terms of CSCMF's can be immediately stated from that in terms of CSMCF by replacing \( \mathbf{z}_i^{(n_1)} \ldots \mathbf{z}_i^{(n_M)} (\lambda^{(n_1)}, \ldots, \lambda^{(n_M)}) \) with the \((n_1 + \ldots + n_M)\)-fold Fourier transform of the function defined in eq. (22).

4. Filtered \( p \)-th-Law Device Identification

In this section, a system identification formula, based on noisy measurements of the input and output signals, is stated for the filtered \( p \)-th-law devices. Specifically, let us consider the model

\[
\begin{align*}
\mathbf{w}(t) &= \mathbf{x}(t) + \mathbf{n}(t), \quad \mathbf{z}(t) = \mathbf{y}(t) + \mathbf{m}(t),
\end{align*}
\]

where both \( \mathbf{n}(t) \) and \( \mathbf{m}(t) \), which are possibly correlated with each other, model noise and interference at the input and output, respectively, and are assumed to be independent (in the fraction-of-time probability sense) of \( \mathbf{x}(t) \).

The system identification problem can be solved by the formula

\[
H(f_1) = \frac{\mathbf{S}_{\mathbf{wn}}(f')N}{\int_{R^{\tau-1}} \mathbf{S}_{\mathbf{wn}}((\lambda^T, f_1 - \lambda^T 1, f_2, \ldots, (f_{N-1})^T)N_{+p-1} d\lambda^T}
\]

where \( \mathbf{N} \)-dimensional vector \( \mathbf{w}(t) \triangleq [w(t), \ldots, w(t)]^T \triangleq [w(t), w_0(t)]^T \) has been introduced. This formula follows from eq. (16) on the assumptions that \( \mathbf{z}(t) \) exhibits \((N+P-1)\)th-order cyclostationarity with cycle frequency \( \alpha \) and, moreover,

\[
\mathbf{S}_{\mathbf{n}}(f_k) \equiv 0, \quad \gamma \neq 0; \quad k = 1, \ldots, N + P - 2, \tag{27}
\]

\[
\mathbf{S}_{\mathbf{n}}(f_k) \equiv 0, \quad k = 1, \ldots, N, \tag{28}
\]

\[
\mathbf{S}_{\mathbf{n}}(f_k) \equiv 0, \quad k = 1, \ldots, N + P - 1, \tag{29}
\]

where \( \mathbf{n}(t) \triangleq [n_1(t), \ldots, n_N(t)]^T \triangleq [n_1(t), n_0(t)]^T \).

Finally, it is worthwhile to underline that no problem arises from assumption (27) since the cyclostationarity of lowest order exhibited by \( \mathbf{z}(t) \) is usually exploited to avoid unnecessary computational complexity.

References


Ventricular Late Potential Detection from Bispectral Analysis of ST-segments

Cameron A. SPEIRS, John J. SORAGHAN, Robert W. STEWART and Shaun BRYNE

University of Strathclyde, Signal Processing Division, Dept. of Electronic and Electrical Eng., Glasgow, G1 1XW, Scotland, UK.
Tel/Fax: +44 [41] 552 4400 / 552 2487, Email: cameron@spd.eee.strathclyde.ac.uk
1 Oxford Medical Limited, Cardiovascular Group, I Kinbar Road, Abingdon, Oxon, OX14 1BZ, England, UK.
Tel/Fax: +44 [235] 533433 / 534465

Abstract. Post Myocardial Infarction subjects are deemed at risk of developing sustained Ventricular Tachycardia if small delayed fragmented signals—or Ventricular Late Potentials (VLPs)—are detected continuous with the end of the QRS complex during sinus rhythm. In this paper bispectral analysis is demonstrated to provide a non-invasive risk indicator by detecting these late micropotentials. Synthesised VLP sequences were corrupted with real ECG noise from the MIT-BIH Arrhythmia database and bispectral analysis showed a discriminatory difference between cases with VLPs and without. The corruptive noise was analysed using Hinich’s statistic [1] and it was hypothesised to be Gaussian in content. The VLP like the entire cardiac complex was considered non-linear hence the problem reduced to one of detecting a non-linearity within a Gaussian environment. The bispectrum suppressed the Gaussian component and unlike 2nd order statistical techniques, was able to detect the underlying non-linearity.

1 Introduction

Studies have correlated arrhythmogenic ventricular activity occurring late in the QRS and extending to within the ST-segment to future episodes of sustained ventricular tachycardia in post myocardial infarction subjects [2]. This delayed activity in the form of fragmented deflections, seen terminal with the ventricular depolarisation wavefront, is termed ventricular late potentials (VLPs). Ventricular late potentials are usually found in border zones surrounding the scar tissue of previous myocardial infarction (MI). The border zone that exists is composed of conducting and non-conducting tissue which slows and fragments the wave of electrical depolarisation as it sweeps through the ventricular myocardium.

Ventricular late potentials are considered to be microvolt-level (1-20uV), high frequency (40-200Hz), waveforms which are continuous with the QRS complex and last into the ST-segment [4].

Typical non-linear behaviour such as phase locking, period doubling bifurcations and chaotic activity have all been observed from the heart in response to electrical stimulation [5, 6] of the Purkinje conduction network. Investigation of such systems resulting from non-linear behaviour using only second order statistical techniques such as power spectral density will yield results based on minimum phase (and hence a linear assumption). This stems from the amount of information present in the autocovariance sequence which accordingly only provides a complete statistical description for processes having a Gaussian distribution. Analysis of ECG signals using second order techniques will provide no information with respect to any underlying non-linear phenomena resulting from interaction between late potential activity and the host waveform (QRS→ST).

Higher order spectral/statistical techniques on the other hand exhibit phase coupled information arising from non-linear mechanisms which would otherwise be lost through the power spectrum [7]. The higher order spectral technique considered here is the bispectrum which is based on the spectral representation of the ECG data's 3rd order statistics (cumulant sequence). The bispectrum is capable of relaying non-linear information as produced by a quadratic operation through spectral peaks at phase coupled components as seen on a two-dimensional frequency plane. Another important property of the bispectrum is that since a Gaussian distributed process may be fully characterised by its mean and variance, the bispectrum of such a process based on 3rd order statistics will yield a flat zero bispectrum over all frequencies in the region.

2 Bispectral Estimation

A real discrete zero mean third order stationary process, X(k), has its third order moment matrix defined as:-

\[ R(m, n) \triangleq E[X(k).X(k + m).X(k + n)] \]  (1)

For a zero mean process the 3rd order moment \( \equiv 3rd \) order cumulant sequence.

The resulting 2-D Fourier Transform of the 3rd order cumulant sequence yields its bispectrum:-

\[ B(\omega_1, \omega_2) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} R(m, n) \exp(-j(\omega_1 m + \omega_2 n)) \]  (2)

Two of the important properties of the bispectrum are:

1. for Gaussian sources the bispectrum is zero over all frequencies in the range.
2. non-linearities maybe detected through phase coupled components in the bispectrum.

Hinich [1] demonstrates a test for Gaussianity and linearity of a stationary time series which we adopt here to determine such characteristics from our ECG sequences. Essentially the bispectrum of a Gaussian process is zero over all frequencies (zero cumulant matrix). The Gaussianity test hypothesises whether there is sufficient statistical deviation from zero to imply a non-Gaussian source. Strictly this test determines the factor of zero-skewness whereby a Gaussian distribution is symmetrical. The linearity test is performed on
the bicoherence function (normalised bispectrum with respect to its power spectrum) and measures the degree to which the bicoherence remains constant statistically over the frequency plane.

3 3rd Order Spectral Estimation of ECG Data Sequences

Synthesised normal ECG complexes were obtained from an ECG generator written by the authors’ in MATLAB. The sampling rate of the data could be arbitrarily changed by the user. The synthetic complexes were compared in the time and frequency domain to real data taken from the MIT-BIH Arrhythmia database (Lead II, sampled at 360Hz) and shown to match favourably. Synthesised ventricular late potentials could also be added to the data records, where their characteristics were modelled as decaying transitory broadband signals [4]. They are applied to the R-wave and allowed to decay into the ST-segment, hence the terminology “late”.

The ST-segments from many complexes of normal sinus rhythm data (Record 103, Lead II) of the MIT-BIH Arrhythmia database were concatenated together to form a single time series of ST-segments. The ST-segment for normal subjects is simply an isoelectric region where no myocardial electric deflection is present and hence allows access to the corrupting noise exclusively. This corrupting noise results from many sources. Electromyographic (EMG) noise due to surrounding muscle stimulus may be considered to give the most wideband characteristics. Prior to segmentation baseline wander due to respiration and electrode movement was removed.

Friesen [8] has modelled this electromyographic noise as having zero mean Gaussian bandlimited characteristics. In performing Hinich’s test for Gaussianity [1] on our new concatenated time series we can hypothesise that the data (noise) does appear to support this Gaussian model.

From our synthetic database we generate sequences which contain late potential activity and sequences which are normal. To these we add the concatenated isoelectric regions (interpolated to match the chosen sampling rate of the synthetic data) to yield synthetic ECGs corrupted by real electrocardiographic noise. The synthetic data is sampled at 2kHz where each time series contains 64 complexes of at least 80 samples per ST-segment.

Hinich’s statistical tests were conducted on our data sets along with the estimation of the bispectrum. The bispectrum was computed using the direct class of conventional estimators [7] with a FFT window length of 128 samples, a Rao–Gabr window and a 50% overlap between segments.

The signal-to-noise ratio between the late potentials and the corrupting noise in our trials was approximately -5dB. The QRS complex, as defined from the unprocessed QRS onset — QRS offset, was not included in the forming the bispectrum since it was found to bias the result and hence mask late potential activity. Figure (1) shows the overall black diagram of the aforementioned procedure culminating in a discriminatory decision as to the existence of ventricular late potentials in a given patient. Figures (2) and (3) show typical ST-segments for data without VLPs and data with respectively. Also shown are their corresponding magnitude bispectra.

4 Discussion of Results

The two main problems in detecting ventricular late potentials from surface ECGs is that the inherent additive noise and QRS morphology effectively masks the time and frequency domain late potential characteristics. Therefore from the surface ECG the morphology of a complex containing late potential activity is similar to the morphology of one without. Time domain signal averaging many complexes is one existing technique which suppresses the random noise component of periodic data. Simson [4] adopts this method of noise reduction before high pass filtering XYZ lead data and forming the vector magnitude to determine the existence of late potentials. However, typically 300 complexes (or more) are required to reduce the noise to a level below that of the late potential activity.

On consideration, our task is the common signal processing problem of estimating a signal from only its noisy observation. As we determined previously from many isoelectric regions of real ECG data, the corrupting noise appears Gaussian in characteristic—as defined using Hinich’s statistics. Similar analysis of the entire ECG cycle—again using Hinich’s tests—shows that the underlying complex characteristics to be non-linear which agrees with [6].

The two properties of the bispectrum noted in Section 2 are our motivation for transforming to the bispectral domain. The Gaussian additive noise is “automatically” suppressed and the bispectrum detects any system non-linearities. Using the power spectrum would simply characterise the Gaussian noise and the non-linear VLP would be rejected through the assumption of a minimum phase system.

Figures (2) & (3) show the resultant bispectra for ECG records with and without late potential activity respectively.

Although ventricular late potentials are documented to arise due to non-conducting regions in the ventricles, their delaying of the excitation wavefront results in ventricular depolarisation components—albeit very small compared to QRS—extending to within the ST-segment. It is within this region that we are interested in detecting phase coupling (non-linearities) and hence inferring activity outwith that of a “normal” isoelectric region.

Looking at the magnitude bispectrum over the principle domain, the discriminatory difference between the two cases is the obvious peak at low interaction frequencies when VLPs are present. In comparison the bispectrum for no VLPs shows no obvious peaks over the entire principle domain, thus implying a Gaussian source or alternatively no non-linear activity. In both cases there appears a small peak at normalised location ~0.2Hz, 0.2Hz). This is we are attributing to the small discontinuities between alignment of ST-segment noise sequences, thus leading to higher frequency components such as those observed. In Figure (2) most of the interaction occurs around the 0 — 0.15 (normalised), or 0 — 300Hz given 2kHz sampling, which is indicative of VLPs [4].

Given that late potential activity can not usually be observed directly from the raw ST-segment time series, the bispectrum is shown to detect phase coupled features within the segment indicative of non-linearities, Figure (2).

Incorporating the QRS only biased the bispectrum to show lower frequency, higher magnitude phase coupling (5 — 40Hz) due to the QRS-to-late potential ratio being so large. In these experiments unprocessed QRS offset was know apriori and the data was segmented accordingly.

Detection of such small non-linearities from within real data ST-segments would certainly be a non-trivial exercise. Initially to yield enough samples within the ST-segment to compute reliable bispectral estimates, the data must be high frequency sampled (2kHz and above). This maybe a worthwhile expenditure given very few complexes are required (we used 34) compared to time averaging detection techniques (approximately 300 complexes) [4].

Automatic ST-segment recognition is possible as shown by Sko-
rdalakis [9]. Any deviation from isoelectric within this region which is the result of phase coupling will be detected. This means that baseline wander due to respiration etc. may also exhibit in the bispectrum and bias the estimates away from the late potential activity. Care should be taken that if removal of baseline wander is desired, suitable high pass filtering is undertaken which will not distort the late potential's phase characteristics. This means filtering in a bidirectional sense such that the resulting phase distortion is zero.

5 Conclusions

In this paper we have used the inherent properties of the bispectral transformation (Gaussian component suppression and non-linearity detection) as a discriminatory tool to detect the presence or absence of terminal ventricular late potentials extending to within the ST-segment. We have been able to utilise these properties since we initially tested the corrupting ECG noise and found it to be Gaussian (or at least originated from a symmetrically distributed PDF), whereas the underlying ECG component was found to result from a non-linear operation. Preprocessing the ST-segments was found to be particularly important given the extremely low SNRs involved, since any low frequencies (baseline wander) effectively biased the resulting bispectra and hence its discriminatory outcome.

Acknowledgements

This work is supported by Oxford Medical Ltd. Cardiovascular Group and the Scientific and Engineering Research Council, UK. CASE Studentship.

References


ST-segments with VLP activity: concatenation of 64 complexes

ST-segments with no VLP activity: concatenation of 64 complexes

Magnitude Bispectrum: over region $0 \rightarrow \Phi_{th}$; $0 \rightarrow (\phi/2)$

Figure 2: Time series and bispectrum for ST-segments with underlying ventricular late potentials.


1132
Identification of Nonlinearities in Vowel Generation

Stephen McLaughlin, Simon Hovell* and Andrew Lowry*

Dept. of Electrical Eng., The University of Edinburgh, Edinburgh EH9 3JL, Scotland, U.K., Tel/Fax: +44 [31] 650 5578 / 650 6554. E-Mail: sml@ee.ed.ac.uk
*BT Laboratories, Signal Processing Division, Edinburgh IP7 5RE, England, U.K., Tel/Fax: +44 [473] 643567 /

Abstract. In this paper the generation of vowel sounds is studied to determine if there are nonlinearities present. This is an important issue since all synthesis models at present are based on linear models. Consequently, if nonlinearities are present in the generation of speech it is clear that linear approaches will not offer optimum performance. This paper looks at the bispectra of vowel sounds to determine the presence of quadratic nonlinearities indicated by the presence of phase coupling between formant frequencies.

1 Introduction and Motivation

This paper is primarily concerned with the role that signal processing techniques based on higher order statistics may play in speech analysis. The work is only a preliminary study and concentrates on determining if there are any significant nonlinearities in vowel generation. To date virtually all practicable digital signal processing techniques have been based on second order statistics. Consequently, the assessment of signals is often based on an examination of the signal spectrum. The conclusion is drawn that if a signal has a flat or near-flat spectrum then the quality of any prediction will be poor. This line of reasoning while useful for linear predictive systems which only exploit the first and second order statistics of the signal, is not true in general since it ignores the higher order statistics of the signal. The last two decades have seen a growing interest, within the signal processing community, in the use of higher order statistics in a variety of applications. The reader is referred to the tutorial papers of [1] and [2].

However, the use of higher order statistics is not new. The work of Pearson around 1900 with the method of moments being the most obvious example. Around, 1920 this approach fell out of favour with the development of Maximum Likelihood by Fischer. This was due to the ML approach providing the optimal estimate although in practice estimating the log-likelihood functions is often non-trivial. Around the early 1960s a group of statisticians at the University of California began to explore the use of these techniques again [3] [4] [5] [6]. However it was not until Mendel in 1980 and his researchers at the University of Southern California began to develop system identification techniques based on HOS methods and apply them to seismic deconvolution problems that the interest of the signal processing community was aroused. At that point in time it was however impractical to implement any of these techniques in real time. In addition, there were questions marks over the robustness of the numerous methods in real world environments particularly with short data records. It is only recently that the advances in DSP technology coupled with research on the robustness issue [7, 8, 9] and more recently [10, 11] that many HOS methods are now potentially feasible. This is one of the motivations for considering these methods in a speech analysis/synthesis application.

The second element to the motivation for this work is the ability of HOS based methods to identify the underlying phase of systems and also nonlinearities present in the signal generation process. This point is important with regard to speech since techniques such as LPC which is based on second order statistics reproduce a spectrally equivalent minimum phase solution and are also linear. It has been conjectured that speech can be considered as being generated by a mechanical system with inherently nonlinear dynamics. The justification for this comment is apparent when one considers the speech generation process. In voiced speech, (such as vowel sounds), airflow through the glottis is modulated by periodic opening and closing of the vocal folds. This airflow is further modified by the vocal tract before emerging as an acoustic pressure wave from the mouth. This highly complex process is generally modelled as a linear system with a time-varying filter (representing the vocal tract) excited by an impulsive periodic signal (representing closure of the vocal folds). Two major assumptions are implicit in this model:

1. The excitation and filter are mutually independent
2. Airflow through the vocal tract is laminar.

It has been known for some time that the former is not true. More controversially, Teager, [12] and [13]
has presented physical measurements that suggest vowel sounds are characterised by highly complex flows involving jets and vortices in the vocal tract rather than well-behaved laminar flow.

2 Experimental Results

In this paper the vowel sounds were analysed by using the bispectrum. The bispectrum of a signal $x(t)$ can be defined in terms of its Fourier transform $X(k)$ [1, 2]:

$$B(k, l) = E[X(k)X(l)X(k + l)],$$

where $E[]$ denotes the expectation operator. The bispectrum will permit us to identify any quadratic phase coupling between the underlying signals in the vowel sound. It was necessary to average over several data records to obtain the results. This is in fact a pre-requisite as was shown in [14]. The phase coupling is revealed in the following manner, consider the signal described below:

$$n_x(k) = \cos(\lambda_1 k + \theta_1) + \cos(\lambda_2 k + \theta_2) + \cos(\lambda_3 k + \theta_3)$$

where $\lambda_1 + \lambda_2 = \lambda_3$. It is clear that if $\theta_3$ is an independent random variable, then the three cosines are independent. On the other hand if $\theta_3$ is dependent of $\theta_1$ and $\theta_2$ then the third cosine is a result of phase coupling between the other two. The power spectra for both these situations would be identical. The bispectrum for the uncoupled case is zero but for the coupled case will have a peak present at $(\lambda_1, \lambda_2)$.

One of the most serious problems when working with higher-order statistics is that the variance of the estimation of the higher-order cumulants is high and in general very long data sequences are needed to reduce the error of estimation. In [10] alternative cumulants estimates are proposed that result in lower variance of estimation at least for symmetrically distributed data. A different approach which was adopted here. This is based on segmenting and averaging the data sequence and usually results in smooth higher-order cumulant estimates:

1. Segment the data into $K$ records of $M$ samples each ($N=KM$)
2. Subtract the average value of each record from the data
3. Assuming that $\{z^{(i)}(k), k = 0, 1, ..., M - 1\}$ is the data set per record $i=1, 2, ..., K$, obtain estimates of the higher-order moments

$$m^{(i)}_{n}(\tau_1, ..., \tau_{n-1}) = \frac{1}{M} \sum_{k=5}^{S_n} z^{(i)}(k+\tau_1) ... z^{(i)}(k+\tau_{n-1})$$

where $n=2, 3, ..., N$, $i=1, 2, ..., K$ and

$$S_1 = \max(0, -\tau_1, ..., \tau_{n-1})$$
$$S_2 = \min(M - 1, M - 1 - \tau_1, ..., M - 1 - \tau_{n-1})$$

4. Average over all segments

$$\hat{m}_{n,2}(\tau_1, ..., \tau_{n-1}) = \frac{1}{K} \sum_{i=1}^{K} m^{(i)}_{n}(\tau_1, ..., \tau_{n-1})$$

5. Finally generate the $n$th-order cumulant sequence $\hat{C}_{n,2}(\tau_1, ..., \tau_{n-1})$, which is known to be a function of moments from second to $n$th order.

If the data sequence is not long enough we can use overlapping records for the calculation of the estimates. In [14] Raghuveer and Nikias provide an example of a process for which segmentation into records is necessary in order to form consistent estimates of third order moments. In this example the segmentation is necessary in order to detect quadratic phase coupling between pairs of sinusoids.

The bispectral analysis of signals is dependant on the signals being studied having some skewness. Figures 1 and 2 clearly demonstrate this to be the case for the vowel sounds studied here.

![Figure 1: Histogram of 4th segment of vowel sound](image)

The basic technique utilised here was to take the vowel sounds and segment them so that an ensemble average of bispectra could be produced as described above) to eliminate any statistical variation which would lead to erroneous assessment of the results. The data was segmented into segments of 128 data points each. This resulted in 64 x 64 bispectra plots which were then averaged to produce the figures. The first vowel sound which was analysed was the vowel sound [a]. The bispectral plot is shown in d in this investigation was from a speaker with a healthy larynx, and care was taken to
avoid aperiodic voicing modes. Any or nonlinear behaviour could then be attributed to normal speech production. The data was sampled at 12kHz with 16 bit linear ADC. The vowel sounds studied are listed below in the speech assessment methodologies - phonetic alphabet (SAM-PA) notation:

- (a) the vowel [], as in cat;
- (b) the vowel [i], as in heed;

Figure 3 shows the bispectrum for the vowel sound [i] and it is clear that significant phase coupling is present. In order to determine if this is due to the presence of a quadratic nonlinearity it is necessary to look at the bicoherence. If significant bicoherence exists at the same point as the peaks exist in the bispectrum then this is strong evidence for the presence of a quadratic nonlinearity. The bicoherence (or to be more precise the skewness function) is defined as follows,

$$
s^{2}(k, l) = \frac{\frac{1}{E[|X(k)X(l)X^{*}(k + l)|^2]} \frac{E[|X(k)|^2]E[|X(l)|^2]E[|X(k + l)|^2]}{E[|X(k)|^2]}.
$$

Figure 4 illustrates the bicoherence for the bispectral data shown in Figure 3.

It is clear that there is significant bicoherence present at the peaks in the bispectrum. Figures 5 and 6 represent the results of the same analysis for the vowel sound [i].

### 3 Conclusions

In this paper we have demonstrated that for simple vowel sounds there is significant phase coupling present between the three formant frequencies in two vowel sounds, and that this is different from sound to sound. The presence of this quadratic nonlinearity in the generation of such sounds has significant implications for speech processing based on linear second order statistics based signal processing.
References


A New Adaptive Algorithm For AR Parameters Estimation Using Cumulants

A. Meziane, A. Adib & D. Aboutajdine
LEESB B.P. 1014 Faculté des Sciences de Rabat-Morocco

Abstract The aim of the present paper is to derive a new adaptive algorithm based on Higher Order Statistics (HOS) with very low complexity and very ease of implementation for Non-Gaussian signals. The most widely used in conventional adaptive methods (based on the autocorrelation) is the least mean squares (LMS) algorithm [11]. But where Non-Gaussian processes or non-linearities are present, the least squares solution fails to provide proper estimations of the system to identify. An iterative solution similar to the LMS algorithm based on a non-conventional orthogonality condition is developed in this paper. A comparative performance analysis of the proposed algorithm and other previous algorithms developed in the literature [10,11] is performed on simulated examples. Some convergence results are also proved.

1. Introduction
Non-Gaussian processes are quite common in signal processing problems. Current adaptive processing techniques based on Higher Order Statistics (HOS) have lately received extensive attention to solve the linear time invariant identification problem. As HOS are phase sensitive, and vanish for coloured Gaussian noise, system identification schemes are very robust if the process to be characterized is corrupted by this kind of noise. Because of these properties, cumulants have found direct application in blind equalization of communication channels, echo cancellation, image and speech processing. Those applications require adaptive algorithms, suitable for real-time processing and non-stationary environment.

An important problem in system identification is the estimation of the system parameters. Friedlander and Porat [4] have developed an Overdetermined Recursive Instrumental Variable algorithm. A gradient type algorithm is also proposed by Nikias and Chiang [2]. They are derived in order to identify non-minimum phase systems. The last algorithm is reformulated by Bakrim and al [1] for time-varying AR parameter estimation. Those methods cannot be used in real-time applications because they need a lot of computations. Concerning the adaptive methods, On one hand, for the non gaussian case, a recursive lattice algorithm for the estimation of the parameters of an AR process, using a 1-D slice of the kth order cumulant General Lattice (GL) is developed in [10]. The complexity is reduced to o(N) operations per iteration. Nevertheless, it has been shown in [7] that it may not exist any 1-D slice that gives a full rank set of normal equations. The W-Slice (WS) algorithm (using a linear combination of 1-D slices) presented in [7] overcomes this problem. In order to reduce more the complexity a Fast Transversal Recursive Instrumental Variable (FTRIV) is proposed in [6]; it is equivalent to two fast transversal filters, one excited by the observed process y(t), and the other by the instrumental process z(t).

On the other hand, for the Gaussian case (autocorrelation), various recursive and iterative techniques were developed to reduce the amount of computation. The most known of them is a stochastic gradient search technique for solving underlying least squares optimization problem, called the Least Mean Squares (LMS) algorithm [11]. It requires only (2N) operations per time-step (iteration). The LMS algorithm may still be appealing because of its simple structure which makes it easy to implement. A novel version of LMS algorithm, based on a modified Mean Square Error (MSE) criterion, is developed in [3].

The objective of this study is to find a new adaptive algorithm, insensitive to additive Gaussian noise (AGN), based on HOS for estimating AR parameters with a computational load comparable to that of the LMS. Instead of searching to minimize the mean square error criterion, we will realize iteratively a non-conventional orthogonality condition between the output prediction error ε(t) and the instrumental process z(t), where appropriate choices for z(t) will be shown to lead to normal equations based on cumulants. Development and results are shown for 3rd and 4th order cumulants.

2. Adaptive AR parameters estimation
Consider a causal real autoregressive (AR) process x(t) described by:

\[ x(t) = \sum_{i=1}^{N} a(i) x(t-i) + w(t) \]  

(1)

\[ y(t) = x(t) + v(t) \]  

(2)

Where the input w(t) is considered to be zero-mean, stationary, independently identically distributed (i.i.d.) process, and hence, has non-zero m-th order cumulant γ_m for some m ≥ 2. y(t) is the noisy output, v(t) is an additive gaussian noise independent of x(t).

The parameters of an AR process can be obtained from the following set of normal equations based on cumulants [10]:

\[ \sum_{i=0}^{N} \gamma(t) C_y(i-1) = \gamma_k \delta(t) \quad \forall \ t \geq 0 \]

(3)

Where C_y(·) stands for a fixed 1-D slice of a cumulant of an arbitrary order. Equations (3) may be compactly written as:

\[ C_{k,y} A_{N+1} = \gamma_k \epsilon_0 \]

(4)

Where A_{N+1} = [a(1) ... a(N)]^T is the AR coefficient vector.

*γ_k* is the k-th order cumulant of the input process w(t).

*ε_0* is an (N+1) vector with unity as its first element and zeros elsewhere.

* C_{k,y} is an (N+1)x(N+1) matrix constructed from the output cumulant and has its C_{k,y}(i,j) elements given by:

\[ C_{k,y}(i,j) = C_{k,y}(i-j,0 \ldots 0) \]

(5)

The instrumental variable (IV) principle has been widely used for constructing identification methods [8,16]. The purpose of this paper is to apply this technique to the identification of a non-Gaussian AR process. The following condition (6) which is an orthogonality condition between the error sequence and the instrumental variable z(t) is assumed to be satisfied for an IV method [8].

\[ E[z(t-i) \epsilon_k(t)] = E[z(t-i) \epsilon_k(t)] = 0 \quad i = 1 \ldots N \]

(6)
Where \( e_n(t) \) is the so-called forward prediction error defined by:
\[
\varepsilon_n(t) = y(t) + A_N^T(t-1) Y_N(t-1)
\]
and appropriate choices for \( Z(t) \) leads to normal equations based on cumulants.

The set of equations in (6) can be written as:
\[
\left( \varepsilon_n(t), Z_n(t-1) \right) = 0
\]
with \( Z_n(t) = [z(t_1), ..., z(t_{N+1})] \)

Hence by approaching the optimal vector \( A_N \) by a set of vectors \( \theta_N(t) \) in equation (7), the result is a formula involving only the vector error \( \delta_N \) given by:
\[
\left( \varepsilon_n(t), Z_n(t-1) \right) = \delta_N
\]

Then, \( \delta_n(t) \) converges exactly to \( A_N \) when \( \delta_N \) tends to the zero vector \( 0 \).

In the autocorrelation case, (i.e.) when \( z(t) = y(t) \) the equation (9) denotes the gradient of the mean square error \( E \left[ \varepsilon_n^2(t) \right] \). The LMS algorithm introduced by Widrow [10] use the steepest descent minimization method to update \( \theta_n(t) \). But, in the presence of low signal to noise ratio (SNR), the LMS algorithm becomes biased and the choice of optimum value \( A_n \) will be erroneous. A new MSE objective function, insensitive to AGN, expressed in terms of the third or fourth-order cumulant of the process \( y(t) \) is introduced by Giannakis in [3] (the key idea is that we can express the output autocorrelation and the MSE objective function as a projections of third or fourth-order statistics). Based on this objective function, a novel consistent version of the LMS algorithm is developed.

3. Proposition of a deterministic algorithm

The algorithm proposed here is not based on the minimization of any objective function, but it is certainly justified as an equation error method to solve a set of linear equation based on cumulants.

Our objective here is to describe a method that starts with any initial parameter guess \( \theta_0(t) \) and seeks \( A_N \) using the update formula:
\[
\theta_n(t) = \theta_n(t-1) - \mu \left( \varepsilon_n(t), Z_n(t-1) \right)
\]
\[
\varepsilon_n(t) = y(t) + \theta_n(t-1) Y_n(t-1)
\]

Where the index \( t \) is the iteration number and \( \mu \) is the constant gain that regulates the speed and stability of adaptation.

Equations (10) and (11) lead to:
\[
\theta_n(t) = \theta_n(t-1) - \mu E \left[ y(t) + \theta_n(t-1) Y_n(t-1) Z_n(t-1) \right]
\]

We can however, solve (12) by transforming to the principal coordinate system. First we translate as in [5], using \( V_n(t) = \theta_n(t) - A_N \), so that it becomes:
\[
V_n(t+1) = (1 - \mu \theta_n) V_n(t)
\]

* If \( z(t) = y(t) \), then the \( \phi \) matrix is the autocorrelation matrix and hence is symmetric and positive definite, implying that it has non negative real eigenvalues. Thus it can be diagonalized by using an orthogonal matrix \( K \) such that:
\[
\phi = K^T Q K, \quad \text{where} \quad Q = \text{diagonal} \left[ \lambda_1, \lambda_2, ..., \lambda_r \right]
\]
and \( \lambda_k \) is the \( r \)th eigenvalue of \( \phi \). In this form, we see that the convergence condition is satisfied by choosing \( \mu \) such that:
\[
0 < \mu \left( \frac{1}{\lambda_{\text{max}}} \right)
\]
where \( \lambda_{\text{max}} \) is the largest eigenvalue of \( \phi \). The condition (14) is a necessary and sufficient condition for the convergence of the algorithm.

* If \( z(t) = y(t) - 3 \epsilon_y(t) y(t) \), then \( \phi \) is the matrix obtained from the diagonal slice of the third-order cumulant.

* If \( z(t) = y(t) - 3 \epsilon_y(t) y(t) \), then \( \phi \) is the matrix obtained from the diagonal slice of the fourth-order cumulants, and so on. Note that in the last two cases, the matrix \( \phi_N \) is not Hermitian. Then, it can be written in Jordan form with \( k \) square blocks on the diagonal:
\[
\phi_N = B^{-1} J B, \quad \text{where} \quad B \text{ is a non-singular matrix and } J \text{ given by:}
\]
\[
J = \begin{bmatrix}
J_1 & & & \\
& \ddots & & \\
& & \ddots & \\
& & & 1
\end{bmatrix}
\]
where \( J_k = \begin{bmatrix}
\lambda_k & & & \\
& \ddots & & \\
& & \ddots & \\
& & & 1
\end{bmatrix}
\]

The matrix \( J_k \) is the Jordan canonical form, whose diagonal elements are the eigenvalues of \( \phi_N \) and 1's just above the diagonal.

Using the expression of \( \phi_N = B^{-1} J B \) and setting \( W_n(t) = B V_n(t) \), equation (13) becomes:
\[
W_n(t) = (1 - \mu \lambda \theta_n) W_n(t-1)
\]

Theorem [9]: The iteration: \( X_{N+1} = M X_N + G \) converges if:
\[
\rho(M) < 1
\]
where \( \rho(M) = \max |\lambda_i| \), \( i = 1...k \) is the spectral radius of the matrix \( M \).

The system in equation (15) converges if:
\[
\max |1 - \mu \lambda| < 1
\]

Let \( \lambda_m = a_m + j b_m \), where \( \lambda_m \) is the eigenvalue of a largest modulus, then:
\[
|1 - \mu a_m| - j \mu b_m^2 < 1
\]
\[\mu \left( \mu (a_m^2 + b_m^2) - 2 a_m \right) < 0
\]

We assume \( \mu > 0 \), then:
\[
0 < \mu \left( 2 \frac{a_m}{|\lambda_m|^2} \right)
\]

Then we get:
\[
0 < \mu \left( 2 \text{ Real} \left( \frac{1}{\lambda_m} \right) \right)
\]
If $\text{Real} \left( \frac{1}{\alpha_m} \right) < 0$, a steepest ascent algorithm must be employed with $2 \text{Real} \left( \frac{1}{\alpha_m} \right) \lessgtr \mu \lessgtr 0$.

4. Proposition of a stochastic algorithm (GLMS)

As was mentioned before, the predictor is determined through the statistics of the observed process $y(t)$. In the most practical applications these statistics are not known or are partially known. Then, the statistic mean $\left\langle \bar{e}_N(t), Z_N(t-1) \right\rangle$ is replaced by its instantaneous value. This leads to the stochastic algorithm given by:

$$\quad \bar{e}_N(t) = y(t) + \Theta_N(t-1) Y_N(t-1)$$

$$\quad \Theta_N(t) = \Theta_N(t-1) - \mu \bar{e}_N(t) Z_N(t-1)$$

First, we show that the stochastic algorithm is equivalent to the deterministic one. We have:

$$\quad \Theta_N(t+p) = \Theta_N(t+p-1) - \mu \bar{e}_N(t+p) Z_N(t+p-1)$$

$$\quad \Theta_N(t+p) = \Theta_N(t) - \mu \left[ \frac{1}{p} \sum_{j=1}^{p} \bar{e}_N(t+j) Z_N(t+j-1) \right]$$

Where:

$$\mu = p \mu$$

For high values of $p$ and with suitable ergodisme conditions, we have:

$$\frac{1}{p} \sum_{j=1}^{p} \bar{e}_N(t+j) Z_N(t+j-1) = \left\langle \bar{e}_N(t+1), Z_N(t) \right\rangle$$

Hence, the stochastic algorithm can be approximated by:

$$\quad \Theta_N(t+p) = \Theta_N(t) - p \mu \left( \bar{e}_N(t+1), Z_N(t) \right)$$

(21)

That is just the deterministic algorithm with $p \mu$ the new gain constant.

The prediction error sequence calculated previously is called "a priori error sequence", because it uses the predictor coefficients vector before its update. The a posteriori error sequence is given by:

$$\quad \bar{e}_N(t) = y(t) + \Theta_N(t) Y_N(t-1)$$

(22)

calculated after updating the predictor coefficients vector. It may be reformulated as:

$$\quad \bar{e}_N(t) = y(t) + \left( \Theta_N(t-1) + \mu \bar{e}_N(t) Z_N(t-1) \right) Y_N(t-1)$$

$$\quad \bar{e}_N(t) = \bar{e}_N(t) \left( 1 + \mu Z_N(t-1) Y_N(t-1) \right)$$

(23)

As the a posteriori error sequence exploits more informations \cite{5}, the system is considered stable if:

$$\quad \text{E} \left( \bar{e}_N(0) \right) \lessgtr \text{E} \left( \bar{e}_N(t) \right)$$

(24)

If the a priori error sequence is assumed to be uncorrelated with the output data ( confirmed by the linear prediction ). The stability condition is given by:

$$\quad \text{E} \left( 1 + \mu Z_N(t-1) Y_N(t-1) \right) \lessgtr 1$$

Then:

$$\quad 0 \lessgtr \mu < \frac{2}{\text{E} \left( Z_N(t-1) Y_N(t-1) \right)}$$

Hence, the step size $\mu$ of the algorithm has to verify the following condition to ensure the stability of the algorithm:

$$\quad 0 \lessgtr \mu < \frac{2}{\text{Trace} \Phi_N}$$

(25)

The structure of $\Phi_N$ leads to:

$$\quad \text{Trace} \Phi_N = N \varphi_{00}$$

(26)

We know that the cumulant matrix obtained from a single 1-D slice is not, in general, of full rank. The diagonal term $\varphi_{00}$ may be equal to zero. In order to overcome this problem we can use a linear combination of $N+1$ slices \cite{7} ( that gives a full rank matrix ). In this case, (26) becomes:

$$\quad \text{Trace} \Phi_N = N W_{00}$$

(27)

Where $W_{ij} = Y_N(i-j)$, $h(t)$ is the impulse response of the system. Then:

$$\quad 0 \lessgtr \mu < \frac{2}{N W_{00}}$$

(28)

In practice, when the sign of $W_{00}$ is unknown, one may start assuming that $\mu > 0$, and switch to $\mu < 0$ if divergence is observed.

5. Simulations

To illustrate the behavior of the adaptive algorithm described above, we performed a number of simple simulation experiments using two kinds of second and third order all pole-filter. The statistics ( mean $\pm$ standard deviation ) of the resulting parameter estimators were approximated by performing 100 Monte Carlo runs.

Case 1: Model leading to a full rank matrix with the following parameters:

$$\quad A(z) = 1 - 0.5z^{-1} + 0.125z^{-2}$$

Case 2: The parameters of the AR process have been chosen as to give a system that does not yield any full rank matrix or when the diagonal term $C_3(0,0) = 0$ of the AR process has been chosen as to give a system that does not yield any full rank matrix or when the diagonal term $C_3(0,0)$ is equal to zero:

$$\quad A(z) = 1 + 1.125z^{-1} + 0.5z^{-2} \quad ( C_3(0,0) = 0 )$$

$$\quad A(z) = 1 + 1.125z^{-1} + 0.5z^{-2} \quad ( C_3(0,0) = 0 )$$

A(z) = 1 + 1.125z^{-1} + 0.5z^{-2} ( Singular matrix )

The following constatations can be drawn from these cases:

* The convergence of the GLMS algorithm is clearly affected by the values of the gain constant $\mu$ ( setting the step size $\mu$ to a smaller value reduces the variance but makes the tracking very slow ). See Fig. 1.

* For a singular matrix, the results show that consistent estimates are given by the GLMS algorithm using a linear combination of $N+1$ slices Fig. 5. When we use a single slice, the algorithm becomes biased and a dramatical increase in the value of deviations is found Fig. 4.

* Note the superior performance of the GLMS and the Giannakis algorithms ( lower bias ) in comparison to the conventional LMS one when the output is corrupted by an AGN. As expected though, they have higher variance. See Fig. 1, 2, 3.

6. Conclusion:

In this paper, a new adaptive algorithm for system identification using cumulants has been proposed. It can be used for AR or ARMA non-Gaussian processes, providing accurate estimations of the AR parameters even in the
The presence of additive coloured Gaussian noise. Development of a normalized version of the GLMS algorithm, as in the case of conventional LMS algorithm, should be fairly straightforward. Note that the described algorithm allows a simple implementation for real time applications than the Giannakis one [3]. Theoretical and practical aspects of the GLMS algorithm have been also investigated.

7. References
Time Delay Estimation Using a Single Cumulant

Chih-Chun FENG and Chong-Yung CHI

Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
Tel: 886-35-731156, Fax: 886-35-715971, E-Mail: cychi@ee.nthu.edu.tw

Abstract. This paper proposes a new cumulant based method for time delay estimation using a single $M$th-order cumulant where $M$ is an even integer greater than three. It is applicable when two sensor measurements are non-Gaussian and contaminated by Gaussian noise sources with unknown statistics. The time delay estimate is obtained using Broyden-Fletcher optimization algorithm. Some simulation results are provided to support that the proposed method works well and is more robust than Tugnait's CUM-TDPE method.

1. Introduction

Assume that $x(n)$ and $y(n)$ are two spatially separated sensor measurements that satisfy

\[ x(n) = s(n) + w_1(n) \]  \hspace{1cm} (1)  
\[ y(n) = s(n - D) + w_2(n) \]  \hspace{1cm} (2)

where $s(n)$ is an unknown signal and $w_1(n)$ and $w_2(n)$ are unknown noise sources. Time delay estimation is a problem to estimate the parameter $D$ from finite measurements of $x(n)$ and $y(n)$ which can be found in such application areas as sonar, radar, biomedicine and geophysics. A number of cross-correlation (second-order statistics) based methods for estimating $D$ have been reported [1,2]. However, correlation based methods are sensitive to additive noise, especially for the case of colored and coherent noise.

Recently, higher-order ($\geq 3$) statistics (HOS) [3], known as cumulants, have been considered in various signal processing areas where desired signals are non-Gaussian and contaminated by Gaussian noise sources, partly because cumulants of a non-Gaussian signal can be used to extract the amplitude information but also phase information of the signal and partly because cumulants of Gaussian noise are totally zero. Some cumulant based methods for time delay estimation [4-6] have also been reported.

Tugnait [5] proposed a discrete searching method which estimates the time delay $D$ by finding the integer $\tau$ such that the following absolute fourth-order cumulant function

\[ S_4(\tau) = |C_{4,x}| = |E[z^4(n)] - 3E^2[z^2(n)]| \] \hspace{1cm} (3)

is minimum where $C_{M,x}$ denotes the $M$th-order cumulant of the error signal $z(n)$ defined as

\[ z(n) = x(n - \tau) - y(n). \] \hspace{1cm} (4)

It can suppress spatially correlated Gaussian noise sources at the expense of higher variance than the method associated with the cross-correlation based mean square error $S_2 = E[z^2(n)]$.

In this paper, we propose a new time delay estimation method based on an objective function given by eq.(7) below which is a generalization of eq.(3) with the error signal $z(n)$ replaced by the one given by eq.(6) below. The proposed method is presented in Section 2. Section 3 provides some simulation results for performance comparison with Chan, Riley and Plant's correlation based method [2] as well as Tugnait's CUM-TDPE method [6]. Finally, we draw some conclusions.

2. A New Cumulant Based Time Delay Estimation Method

Assume that $x(n)$ and $y(n)$, $n = 0, 1, \ldots, N - 1$ are given non-Gaussian noisy measurements generated from the model given by eq.(1) and eq.(2), respectively, while $s(n)$ is generated from a convolutional model

\[ s(n) = u(n) * h(n) \] \hspace{1cm} (5)

where $h(n)$ is a linear time-invariant (LTI) system. Let us make the following statistical assumptions for $u(n)$, $w_1(n)$ and $w_2(n)$:
(A1) \( u(n) \) is a real, zero-mean, stationary, independent identically distributed (i.i.d.), non-Gaussian driving input sequence with \( M \)-th order cumulant \( C_{M,u} \neq 0 \) where \( M \) is even and \( M \geq 4 \).

(A2) \( w_1(n) \) and \( w_2(n) \) are zero-mean Gaussian with unknown statistics.

(A3) The input \( u(n) \) is statistically independent of \( w_1(n) \) and \( w_2(n) \).

The new cumulant based time delay estimation method uses a single cumulant of the error signal \( e(n) \) defined as

\[
e(n) = x(n) - y(n) = \sum_{i=-p}^{p} v(i)x(n-i) - y(n)
\]

(6)

where \( v(n) \) is an FIR filter and \( p \) is the largest possible delay one can expect. The new method is based on the following theorem:

**Theorem 1.** Assume that \( x(n) \) and \( y(n) \) are the noisy signals associated with eq.(1), eq.(2) and eq.(5) under the assumptions (A1) through (A3) and the time delay \( D \) is an integer. Let \( v(n) \) be the FIR filter associated with the error signal \( e(n) \) given by eq.(6) where \( p \geq |D| \). Let \( \tilde{v}(n) \) be the optimum FIR filter such that the following objective function

\[
J_M(v(n)) = |C_{M,e}|
\]

(7)

is minimum where \( M \) is an even integer and \( M \geq 4 \). Then \( \tilde{v}(n) = \delta(n - D) \) and \( J_M(\tilde{v}(n)) = 0 \).

**Proof.** Substituting eq.(1), eq.(2) and eq.(5) into eq.(6) yields

\[
e(n) = u(n) * v(n) + w(n)
\]

(8)

where \( w(n) = w_1(n) * v(n) - w_2(n) \) is also a Gaussian noise sequence by (A2) and

\[
g(n) = h(n) \cdot [v(n) - \delta(n - D)].
\]

(9)

The objective function \( J_M \) given by eq.(7) can be expressed as [3]:

\[
J_M = \left| C_{M,u} \cdot \sum_{n=-\infty}^{\infty} g^M(n) \right|
\]

(10)

which equals to zero only when \( g(n) = 0 \) for all \( n \) since \( M \) is even. One can easily see, from eq.(9), that \( g(n) = 0 \) leads to \( \tilde{v}(n) = \delta(n - D) \).

Based on Theorem 1, the proposed method estimates \( D \) by minimizing \( J_M \) given by eq.(7) with \( C_{M,e} \) replaced by \( M \)-th order sample cumulant \( \tilde{C}_{M,e} \). Since \( J_M(v) \) is a highly nonlinear function of \( v \), where \( v = [v(-p), v(-p+1), \ldots, v(p)]^T \), Broyden-Fletcher optimization algorithm [7] is used to find the optimum estimate \( \tilde{v} \). At the \( (i+1) \)th iteration, \( \tilde{v} \) is updated by

\[
v_{i+1} = v_i - \alpha_i H_i g_i
\]

(11)

where \( \alpha_i \) is a positive scaling factor for adjusting the step size, \( g_i \) is the gradient of \( J_M(v) \) with respect to \( v \) for \( v = v_i \) and

\[
H_i = H_i + \frac{1}{r_i^T s_i} \left[ (1 + \frac{s_i^T H_i s_i}{r_i^T s_i}) r_i r_i^T - r_i s_i^T H_i s_i r_i^T \right]
\]

(12)

in which \( r_i = v_i - \tilde{v} \) and \( s_i = g_i - g_i - g_i \). Any positive definite matrix (e.g., identity matrix) can be used for the initial matrix \( H_0 \) and always leads to a positive definite \( H_i \) for \( i \geq 0 \) provided that \( \alpha_i \) is chosen such that \( J_M(v_{i+1}) < J_M(v_i) \). By our experience, \( v_0 = 0 \) is a good choice to initialize the above optimization algorithm. The optimum time delay estimate \( \tilde{D} \) is determined to be the index associated with \( \tilde{v}(D) = max(\tilde{v}(n), -p \leq n \leq p) \). When time delay \( D \) is not an integer, one can estimate \( \tilde{D} \) by interpolation [2].

Several remarks regarding the proposed time delay estimation method are described as follows:

(R1) The proposed criterion given by eq.(7) can also be thought of as an extension of Wiener filter based criterion minimizing \( E[e^2(n)] \) proposed by Chan, Riley and Plant [2].

(R2) If \( v(n) \) is constrained to take the form of \( \delta(n - \tau) \), then \( e(n) = x(n) \) and \( J_4 = S_4(\tau) \) (see eq.(3) and eq.(4)).

(R3) Tugnait [8] proposed an input-output ARMA system identification method using fourth-order cumulants. For the case of MA system \( v(n) \), this method estimates \( v(n) \) by minimizing \( \sqrt{C_{4,v}} \) where \( e(n) \) is also given by eq.(6) in which \( x(n) \) and \( y(n) \) are noisy input and output measurements of the system, respectively. The proposed time delay estimation method for \( M = 4 \) is equivalent to a special case of the Tugnait’s system identification method when the MA system \( v(n) = \delta(n - D) \).

3. Simulation Results

In the simulation, the driving input \( u(n) \) used was assumed to be a zero-mean, Exponentially distributed i.i.d. random sequence with variance \( \sigma_u^2 = 1 \), Kurtosis \( C_4,u = 6 \) and sixth-order cumulant \( C_{6,u} = 120 \). The noise source \( w_1(n) \) was a colored Gaussian sequence generated from a first-order moving average system with coefficients \( \{1,0.8\} \) driven by a white Gaussian.
noise sequence. The true time delay was assumed to be $D = 8$ and the noise source $w_2(n) = w_1(n - 3)$ (i.e., sensor noise sources were spatially coherent and colored Gaussian). For comparison, the Wiener filter based method proposed by Chan, Riley and Plant [2] and the cumulant based time delay parameter estimation method (CUM-TDPE) proposed by Tugnait [6] which uses fourth-order cumulants and cross cumulants were also employed to estimate $D$ with the same data.

**Example 1. (White Signal)**

The unknown signal $s(n) = u(n)$ (i.e., $h(n) = \delta(n)$) was used to generate measurements $x(n)$ and $y(n)$ for $N = 16000$ and $SNR = 0 dB$. The optimum $\hat{\theta}(n)$ was obtained by the previous optimization algorithm with $p = 15$. Thirty independent runs were performed and the obtained thirty estimates $\hat{\theta}(n)$ are shown in Figures 1(a) through 1(d), associated with Chan, Riley and Plant’s Wiener filter based method [2], Tugnait’s CUM-TDPE method [6], and the proposed method with $M = 4$ as well as $M = 6$, respectively. From Figure 1(a), one can see that all the estimates $\hat{\theta}(n)$ approximate $0.43\delta(n - 3) + 0.57\delta(n - 8)$ because of $w_2(n) = w_1(n - 3)$ although the variance is small. From Figures 1(b) through 1(d), one can see that all the estimates $\hat{\theta}(n)$ approximate $\delta(n - 8)$ except for a scale factor. These simulation results demonstrate that both Tugnait’s CUM-TDPE method and the proposed method are effective for suppressing coherent colored Gaussian noises while the former has a smaller variance than the latter for this case.

**Example 2. (Colored Signal)**

The unknown signal $s(n)$ was generated from a first-order lowpass FIR filter $h(n)$ (see eq. (5)) with coefficients $[1, 1]$ driven by the input sequence $u(n)$. Table 1 shows MEAN and RMS (root-mean-square) error calculated from the obtained thirty estimates $\hat{D}$ for $N = 8000$ and $N = 16000$, $SNR = 0 dB$ and $SNR = -5 dB$, and $p = 15$ and $p = 30$. One can observe, from Table 1, that the proposed method for $M = 4$ provides more reliable estimates $\hat{D}$ than the other methods. However, the proposed method for $M = 6$ has quite large variance because sixth-order sample cumulants have much larger variance than fourth-order cumulants used by Tugnait’s method and the proposed method for $M = 4$.

The simulation results shown in the above examples also indicate that the proposed method for $M = 4$ is more robust than Tugnait’s method for finite data. The reason for this could be that only a single sample cumulant, which itself is random, is used by the former, but the number of sample cumulants used by the latter is proportional to $(2p + 1)^2$ which accounts for large bias and variance for large $p$.

**4. Conclusions**

We have presented a new cumulant based method for time delay estimation using a single $M$th-order cumulant where $M$ is an even integer greater than three. It is also a parameter estimation method implemented by Brodyen-Fletcher algorithm. Simulation results support that it works well and is more robust than Tugnait’s CUM-TDPE method.

**Acknowledgements**

This research was supported by the National Science Council under Grant NSC-83-0404-E-007-037.

**References**


Figure 1. Simulation results of Example 1 ($N = 16000$ and $SNR = 0 \text{ dB}$). The true time delay is $D = 8$, signal $s(n)$ is white, noise sources are spatially coherent and colored Gaussian. Thirty estimates $\hat{s}(n)$ for $p = 15$ obtained using (a) Chan, Riley and Plant’s Wiener filter based method, (b) Tugnait’s CUM-TDPE method and the proposed method with (c) $M = 4$ and (d) $M = 6$, respectively.

Table 1. Simulation results of Example 2. The true time delay is $D = 8$, signal $s(n)$ is colored, and noise sources are spatially coherent and colored Gaussian.

<table>
<thead>
<tr>
<th>Approach</th>
<th>$SNR = 0 \text{ dB}$</th>
<th>$SNR = -5 \text{ dB}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N = 8000$</td>
<td>$N = 16000$</td>
</tr>
<tr>
<td></td>
<td>MEAN</td>
<td>RMS</td>
</tr>
<tr>
<td>Wiener filter based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>method [2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 15$</td>
<td>3.00</td>
<td>5.00</td>
</tr>
<tr>
<td>$p = 30$</td>
<td>3.00</td>
<td>5.00</td>
</tr>
<tr>
<td>Tugnait’s CUM-TDPE method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>method [6]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 15$</td>
<td>7.70</td>
<td>1.64</td>
</tr>
<tr>
<td>$p = 30$</td>
<td>8.00</td>
<td>0.00</td>
</tr>
<tr>
<td>The proposed method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>with $M = 4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 15$</td>
<td>8.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$p = 30$</td>
<td>8.00</td>
<td>0.00</td>
</tr>
<tr>
<td>The proposed method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>with $M = 6$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 15$</td>
<td>7.30</td>
<td>3.48</td>
</tr>
<tr>
<td>$p = 30$</td>
<td>6.67</td>
<td>7.97</td>
</tr>
</tbody>
</table>
Horizontal-Slice Cumulant Matrices for Parameter Estimation of Exponentially Damped Sinusoids.

Diego P. RUIZ, María C. CARRION, Antolino GALLEGO and Abdellatif MEDOURI.

Dpto. Física Aplicada, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain, Tel/Fax: 958-243229/34-58-243214, E-mail: ondas@ugr.es

Abstract. A very common problem in signal processing is that of parameter estimation of exponentially damped sinusoids embedded in additive, white or colored, Gaussian noise. Recently a new approach based on higher-order statistic was proposed in [2], which is basically a reformulated Kumaresan-Tufts (KT) method using a diagonal slice of estimated third- or fourth-order cumulants. However, when there are few data available and the SNR is low, the diagonal slice sequence fails to provide better estimations than the original KT method. In this paper the parameter estimation problem is reformulated and it is demonstrated that only certain horizontal slices of cumulants allow the exponentially damped structure to be maintained, which is necessary for the KT method to subsequently be applied to the cumulant sequence. Simulation results show that the HOS-based approach using horizontal slices performs better than the original KT method even when the data record is small.

1. Introduction

Many measurement problems can, in practice, be reduced to the of estimation of the parameters of a damped sinusoidal signal, which is also related to fitting pole-zero models to observed data [1]. The problem, however, becomes very difficult when the number of available data is small and the SNR is low. A crucial advance to solving this problem was recently proposed by Kumaresan and Tufts, who developed the Principal Eigenvector method (PE) [1], which preprocesses the Linear Prediction (LP) data matrix or autocorrelation matrix using a low-rank approximation. The PE method performs very well when the additive noise is Gaussian white. However, when the additive noise is coloured, the Kumaresan-Tufts method fails at low SNR, essentially because it views the coloured noise spectrum as extra damped sinusoids, so it tends to overestimate the model order. To overcome this problem, Papadopoulos and Nikias [2] used third- and fourth-order statistics or cumulants, which are blind to additive Gaussian noise, white or coloured. They proposed the construction of a third- or fourth-order correlation matrix along the diagonal slice instead of using the autocorrelation matrix. As shown in [2], these matrices have the correct rank and thus the prediction-error filter polynomial can be rooted to estimate the frequencies.

Nevertheless, when the number of available data is finite, as occurs in practice, the matrices constructed using the diagonal line of estimated cumulants do not maintain the correct rank due to the estimated cumulant sequence no longer being modeled as a damped exponential model with the same frequencies as the data. This problem becomes more serious when the number of data is lower. Our interest in this paper lies in finding which slices of third- or fourth-order cumulants allow us to obtain the complex frequencies of exponentially damped sinusoidal models, not only asymptotically, when the number of data is infinite, but also with finite data length. Simulation examples demonstrate that matrices constructed with the estimated horizontal-slice cumulants perform better than diagonal-slice ones when the number of data is small and the SNR low. Recently, different matrices based on an alternative [3] method have been proposed in [4] to handle this problem, which show better behaviour in frequency resolution that the diagonal-slice matrices at moderate SNR.

2. Cumulant-Based Approach

Consider the noiseless signal defined in eq. (1), for any time instant \( n \geq 0 \) as:

\[ x(n) = \sum_{m=1}^{M} a_m e^{j\omega_m n} \]  

(1)

where the signal contains \( M \) exponentially damped sinusoids with complex amplitudes \( a_m \) and complex frequencies \( \omega_m = \sigma_m + j\gamma_m \), \( m = 1,2,...,M \). We are concerned with the estimation of the frequencies and damping factors from a one-dimensional slice (full-rank slice) of the higher-order correlation sequence of the energy signal \( x(n) \) using a finite number of observed data \( x(n) \) \( n = 0,1,2,3,\ldots \). In the third-order case, the third-order correlation sequence is defined as:

\[ R_x(r_1,r_2) = \sum_{n=-\infty}^{\infty} x(n)x^*(n+r_1)x^*(n+r_2) \quad r_1, r_2 = 0,1,2,3,\ldots \]  

(2)

One dimensional slices of third-order correlations \( r_x(t) \) can be defined taking \( r_x = at+b \) in eq. (2) where \( a \) and \( b \) are the slope and intercept respectively of the line along the moments plane.

The signal parameters could be recovered from a one
dimensional slice (1-D slice) in the moment plane provided that the 1-D slice can be modeled as the sum of M decaying complex exponentials oscillating with the same frequencies and damping factors as the original signal. Since the 1-D slice behaves like a damped exponential signal, the KT method can be applied, and so the frequencies and damping coefficients of the signal can be identified.

The question that arises here is when the 1-D slice can be modeled as the sum of M damped exponentials. In the case of a deterministic signal, we can distinguish two cases:

2.1 Infinite Data Length
If we know the energy signal elsewhere, we can construct the true third-order correlation sequence without "deterministic" errors. Taking into account the energy signal defined in eq.(1) and performing the geometric sum, the 1-D slice in the plane $\tau_1, \tau_2$ becomes:

$$r_{(\tau)} = \sum_{0}^{\infty} \alpha \alpha^{*} e^{i \phi} e^{i \phi(\alpha - b)} e^{i \phi(\alpha + b)} 1 - e^{-2 \alpha \tau_1 \tau_2}$$

$$n = \max(0, -\tau_1, -\tau_2)$$

since $x(n) = 0$ for $n < 0$. To build a damped exponential model for $r_{(\tau)}$, we can vary $\tau$ from 0 to $\infty$ or from 0 to $-\infty$. Several cases then appear depending on the quadrant in the moment plane the 1-D slice lies on. These cases are collected in the following proposition:

**Proposition 1.** Only horizontal slices in the first and fourth quadrants and diagonal slices in the third quadrant allows the original damped exponential structure of data for the moment sequence to be preserved.

Proof. See appendix A.

2.2 Finite Data Length and Noisy Data.
In the previous subsection we dealt with true higher-order correlations of the energy signal $x(n)$. In practice we only have a finite subset of data samples, namely $n = 0, ..., N-1$ and we want to estimate nm moments from this data. In this case there will be "deterministic" errors associated with imperfect estimation of correlations. These errors can eliminate the fundamental property of higher-order correlations which allows the signal parameters to be recovered, i.e. the moment sequence should be modeled as a sum of exponentially damped sinusoids even for the finite data case. The higher-order correlation of the deterministic signal can be estimated using the biased estimator [3]:

$$\hat{R}_{(\tau_1, \tau_2)} = \frac{1}{N} \sum_{n=0}^{N-1} x(n)x(n+\tau_1)x(n+\tau_2)$$

$$S_1 = \max(0, -\tau_1, -\tau_2)$$

$$S_2 = \min(N-1, N-1-\tau_1, N-1-\tau_2)$$

Using this estimated moment sequence, not all the slices allowed by proposition 1 maintain the exponential structure with the same frequencies as the data. This situation can invalidate the damped exponential model for $r_{(\tau)}$. The different cases that arises are summarized in proposition 2:

**Proposition 2.** If a biased (or unbiased) estimator is used to estimate the moment sequence from N data, only certain horizontal slices (with high lags) in the first quadrant can be used to model $r_{(\tau)}$ as a sum of M decaying complex exponentials. In this case the diagonal line in the third quadrant is not valid.

The proof of this proposition follows the same reasoning as proposition 1 and is omitted for brevity. From proposition 2 it is clear that the horizontal slice which allows a higher number of correlations corresponds to that defined by $b \neq 0$.

In addition, in many practical cases the signal is contaminated with additive noise. Therefore, let us suppose that the observed data sequence consists of N samples from M exponentially damped signals in additive complex valued Gaussian noise:

$$y(n) = x(n) + w(n) \quad n = 0, 1, ..., N-1$$

where $x(n)$ is defined in eq.(1) and the additive noise $w(n)$ is a zero-mean, colored or white complex stationary Gaussian process, with independent real and imaginary parts identically distributed, and also independent of the signal. Assuming a large number of realizations available and defining the signal $y'(n)$ as $x'(n)+w(n)$, $x'(n)$ being the signal in eq.(1) with the mean removed, the third-order cumulant sequence of $y'(n)$ can be expressed [3]:

$$\hat{c}_{(\tau_1, \tau_2, \tau_3)} = E \left[ \frac{1}{N} \sum_{n=0}^{N-1} y'(n)y'(n+\tau_1)y'(n+\tau_2) \right] = 0$$

$$= \frac{1}{N} \sum_{n=0}^{N-1} x'(n)x'(n+\tau_1)x'(n+\tau_2) = \hat{R}_{(\tau_1, \tau_2, \tau_3)}$$

where the summation have been included since the process is not stationary, and it has been taken into account that the signal is independent from the noise and that the third-order moments of a Gaussian process is zero. From eq.(6) it can be seen that the third-order cumulant sequence is the third-order correlation sequence of $x'(n)$, to which can be applied proposition 2 to model it as a damped exponentials model with $M+1$ sinusoids (M frequencies of the original signal plus an additional frequency at zero).

When only a unique realization is available (which occurs in many practical situations), the expected value is not taken, and undesirable terms which behave as noise appear, disturbing the moment sequence. In this case the influence of the noise on the third-order correlations of the signal cannot be avoided, although the combination with robust techniques such as SVD to solve for signal parameters allows better results to be obtained than the use of this technique with raw
data such as is shown in our simulations.

Likewise, the same results can easily be extended to fourth-order cumulants, with both proposition 1 and 2 remaining valid in particular.

3. Simulation Results

In this section we provide computer simulation results using fourth-order cumulants (FOC), since they are clearly superior to third-order moments. In order to compare the different methods in our simulations, we use the diagonal slice sequence of FOC (DS-FOC) used in [2], a horizontal slice of FOC allowed by proposition 2 (HS-FOC), and the KT method. In the case of HS-FOC the slice allowing a higher number of correlations is chosen. The experiment carried out in this simulation corresponds to that used in [1]-[2], described by the formula:

\[ y(n) = a_1e^{j2\pi n} + a_2e^{j2\pi n} + w(nT) \quad n=0,1,2,\ldots,29 \]  

where \( s_1=-0.2+j2\pi(0.42), s_2=-0.1+j2\pi(0.52), a_1=1, a_2=1, T=0.5 \). The sequence \( w(n) \) is an i.i.d. complex Gaussian process with a variance \( 2\sigma^2 \). The signal-to-noise ratio is defined as \( SNR = 10 \log(1/\sigma^2) \). For brevity’s sake, the results shown in this section correspond only to the more interesting case of coloured noise, generated by passing a complex white Gaussian process generated by the IMSL library through an FIR filter with an impulse response given by \( h=[0.5, 0.6, 0.7, 0.8, 0.7, 0.6, 0.5, 0, 0.5, 0.6, 0.7, 0.8, 0.7, 0.6, 0.5] \).

\[ 10\log_{10}(1/MSE) \]

![Figure 1](image1.png)

**Figure 1.** MSE of estimated damping factor (-0.2) as a function of the SNR in additive coloured Gaussian noise.

A statistical measurement of the performance of these methods was obtained via Monte Carlo simulations, doing 500 independent runs with the transient signal kept the same and generating independent noise realizations using different seeds. From 30 data points, correlations were estimated along the diagonal slice and the horizontal slice with \( b=25 \), and a Hankel matrix was constructed with a filter length of \( K=13 \). For the purpose of comparison it is also shown the KT method for \( K=18 \) and \( N=25 \) data. The results of these runs were used to calculate the mean square error (MSE) of the estimated frequencies (0.42 and 0.52) and damping factors (-0.2 and -0.1) as functions of the SNR over the range of 5 to 30 dB, (Figures 1-4). Threshold points occur when the bias is too large (100% of the true value) or more than M zeros appear outside the unit circle. These threshold points happen at low SNR and constitute the practical limitation in SNR of these methods. It is clear in these figures that the HS-FOC has an advantage over the DS-FOC based method, and also performs better with respect to the KT method, even with a small number of data. This is not surprising since this method is influenced by coloured noise whose components are closer to the frequency 0.42 and this frequency is thus poorly estimated at low SNR.

\[ 10\log_{10}(1/MSE) \]

![Figure 2](image2.png)

**Figure 2.** MSE of estimated damping factor (-0.1) as a function of the SNR in additive coloured Gaussian noise.

\[ 10\log_{10}(1/MSE) \]

![Figure 3](image3.png)

**Figure 3.** MSE of estimated frequency (0.42) as a function of the SNR in additive coloured Gaussian noise.

\[ 10\log_{10}(1/MSE) \]

![Figure 4](image4.png)

**Figure 4.** MSE of estimated frequency (0.52) as a function of the SNR in additive coloured Gaussian noise.
The better performance of the HS-FOC method compared to
the DS-FOC version shown in Figures 1-4 is explained by the
effect of the "deterministic errors" associated with the
cumulant sequence, which cause the cumulant sequence to
contain several oscillating components. This is clarified in
Figure 5, which shows the normalized singular values of the
Hankel matrix constructed by using the DS and the HS with
no noise.

![Figure 5. Normalized singular values for DS-FOC matrix (dashed line) and HS-FOC matrix (solid line).](image)

A strong break at order 3 in the HS-FOC matrix can be
observed here but, on the other hand, the DS-FOC matrix
exhibits a slow decrease in the singular values, indicating that
the cumulant sequence in not modeled as a damped
exponential model with the same frequencies as the data. This
modelling error is the cause of the worse estimation, specially
in the complex frequency ($-0.1 + j2\pi 0.52$).

When the number of data N increases, the DS-FOC method
shows better performance in noise, since the modelling errors
decrease quickly with N. At high N and low SNR, the DS
method performs better than the HS one, due mainly to the
high lags used in the computation of higher-order correlations,
which implies using data strongly corrupted by noise making
the estimation more sensitive to noise.

4. Conclusions

In summary, only specific slices of third- or fourth-order
cumulants obey an exponential model with the same
frequencies as the original data. Therefore, the PE method can
be used to compute the frequencies of exponentially damped
sinusoids with cumulants instead of data, with the added
advantage that the Gaussian additive noise cumulants of an
order greater than two are identical to zero. When data length is
finite, similar conditions are established for estimates of
cumulants. Simulation examples show that horizontal-slice
matrices perform better than diagonal-slice ones when the
number of available data is small.

Acknowledgements

This work was partially supported by the "Comisión
Interministerial de Ciencia y Tecnología" (CICYT) of Spain
under project TIC92-0663.

References

parameters of exponentially damped sinusoids and pole-zero
Estimation of Exponentially Damped Sinusoids Using
1990.
(Spectra) in Signal Processing and System Theory:
Theoretical Results and Some Applications, Proceedings
(1994), A "Constrained Third-Order Mean"-type
estimator to calculate the parameters of exponentially damped
sinusoids using third-order statistics, to appear in Signal
Processing North-Holland.

Appendix A

Proof of Proposition 1.

Depending on the quadrant the 1-D slice lies on in the $\tau_1$-$\tau_2$
plane, four cases can be distinguished:

Case 1.- The 1-D slice lies on the first quadrant.- In this case,
$\tau$ ranges from 0 to $\infty$, the Slope $a$ is greater than 0, and
the Intercept $b$ can be taken as positive since $b<0$ would
lead to a change in the order of $\tau$. Taking these
values into account, $n_0$ defined in eq.(9) is equal to 0, so the
moment sequence becomes:

$$r(\tau) = \frac{M}{\prod_{k} \beta_k} \sum_{\beta_k} a_k^\tau_a^\tau e^{i(\alpha_k+\theta_k)} \quad (A.1)$$

To model $r(\tau)$ as a damped exponential sinusoids with the
same frequencies as the original signal, $a$ must be chosen
equal to 0 which is equivalent to taking horizontal slices.

Case 2.- Second quadrant.- In this case, $\tau=0$, $\tau_1<\infty$, $\tau_2<0$ and
$\tau_2\tau_1>0$. Now $n_0=0$ and exponentials terms appear in the
moment sequence, impeding any slice from being modelled as
an exponential sum in this quadrant.

Case 3.- Third quadrant.- $\tau=0$, $\tau_1<\infty$, $\tau_2<0$ and
$\tau_2\tau_1>0$. In this case, the following results can be obtained:

If $a>1$ or $c<1$, there is no slice which can be modelled as the
same exponential model as data.

If $a=1$ (diagonal slice), the moment sequence becomes:

$$r(\tau) = \frac{M}{\prod_{k} \beta_k} \sum_{\beta_k} A(0) e^{\alpha_0} \sum_{\beta_k} \frac{A(0) \beta_k}{\prod_{k} \beta_k} e^{-\beta_k \theta_k + \alpha_k \tau} \quad (A.2)$$

As may be seen, the diagonal slice allows the signal
parameters to be obtained. This diagonal slice was used in [2]
with $b=0$ to recover signal parameters.

Case 4.- Fourth quadrant.- In this case, $\tau_1>\tau_2$, $a>0$, $b<0$ and
$n_0=0$. Since $a<0$, $a$ must be equal to 0 (horizontal slice).
Identification of a Quadratic System using only Output Cumulants

Pascal Bondon and Michel Krob

Laboratoire des Signaux et Systèmes, CNRS (UMR 14)
École Supérieure d'Électricité, Plateau de Moulon
91129 Gif-sur-Yvette Cedex, France.
Tel/Fax: 33 1 69 41 80 40 / 30 60, E-mail: bondon@lss.supelec.fr

Abstract This paper deals with the identification of the coefficients of any discrete and finite extent quadratic system driven by an unobservable independent identically distributed random sequence. Based on output cumulants up to the third order, a sufficient condition to identify the parameters is established, and some examples are given.

1 Introduction

The input-output relation of a discrete and time-invariant quadratic system writes

\[ y(n) = \sum_{k,l} h(k,l) x(n-k) x(n-l) \]  

(1)

where, without loss of generality, the kernel \( h(\cdot) \) is assumed symmetric in its arguments. A recent review of the use of quadratic systems in signal processing may be found in [9]. For example, they appear when testing two zero mean Gaussian distributions with different covariance matrices [5, chap. 11], for estimating the power of a random signal [10, Part I chap. 2] and in image processing [8]. The quadratic systems can have particular forms in practical situations. For instance, \( h(\cdot) \) can be diagonal, i.e., \( h(k,l) = 0 \) for \( k \neq l \), which corresponds to a special case of the Hammerstein model, \( h(\cdot) \) can be dyadic, i.e., \( h(k,l) = h(k)h(l) \), which is a particular case of the Wiener model, \( h(\cdot) \) can be positive definite as happens if the quadratic system is used to estimate the power of a random signal and \( h(\cdot) \) can be Toeplitz, i.e., \( h(k,l) = h(k-l) \), as arises in some applications of adaptive detection [7]. Finally, efficient implementations of quadratic digital filters have been considered in [4].

In this work, we just assume that \( h(\cdot) \) is non zero only for a finite number of its parameters, and we study the identification of \( h(\cdot) \) when the input is unobservable but is independent and identically distributed (i.i.d.). For this purpose, we use only output cumulants up to the third order. The previous contributions to this blind identification problem are the following. In [1], \( x(n) \) was assumed to be Gaussian and the parameter set was chosen to be the one that minimizes the sum of the squared differences between the second-and third-order moments of the output of the system and the moments of the output of the model. A nonlinear optimization routine was used and the authors observed in practice that the solution was generally not unique. In [6], \( y(n) \) was an arbitrary input but only specific linear-quadratic systems were considered. For these systems, the authors showed that the third-order moments of \( y(n) \) are very simply related to the coefficients of the systems. Of course this property does not hold for a general linear-quadratic system. Finally in [2], a general quadratic system driven by an arbitrary unobservable sequence was considered, and necessary and sufficient conditions to determine the order of the kernel were established.

In this paper, we show that under some conditions the coefficients \( h(k,l) \) can be expressed in terms of the coefficients \( h(k,k) \) and of some third-order output cumulants. The parameters \( h(k,k) \) are then a solution of nonlinear equations. It is shown that the solutions to these equations are not unique up to a time shift and a scaling factor, and a sufficient condition of uniqueness is given. To illustrate the contents of the paper some particular models of quadratic systems are then studied.

The organization of the paper is as follows. Some preliminary results are given in section 2. Section 3 is devoted to the determination of the parameters of the system. Some examples are discussed in section 4 and finally, the conclusion is given in section 5.

2 Preliminary results

In all the following, the sequence \( x(n) \) is assumed i.i.d. and zero mean. Regrouping in (1) the terms which belong to the same parallel to the principal diagonal,
we obtain
\[ y(n) = y_0(n) + 2 \sum_{k>0} y_k(n) \]
where
\[ y_k(n) = \sum_{l} h_k(l)v_k(n-l) \]
\[ h_k(l) = h(k+l, l) \quad v_k(l) = \zeta(l-k) \zeta(l) \]
As \( x(n) \) is stationary and (1) is time-invariant and absolutely summable, \( y(n) \) is also stationary and the pth-order cumulant function of \( y(n) \), \( C_{yp}(.) \), exists. \( C_{yp}(.) \) is complicated to calculate because it depends on \( h(.) \) and on the input cumulants \( \gamma_x \) for \( 2 \leq q \leq 2p \). The expressions of \( C_{yp}(.) \) for \( p = 1, 2, 3 \) given hereunder are demonstrated in [3].

\[ C_{1y} = \gamma_x \sum_n h_0(n) \]
\[ C_{2y}(i) = (\gamma_x + 2\gamma_x^2) \sum_n h_0(n)h_0(i + n) + 4\gamma_x^2 \sum_{k>0} h_k(n)h_0(i + n) \]
\[ C_{3y}(i_1, i_2) = \gamma_x \psi_0(i_1, i_2) + \gamma_x \gamma_x \omega_1^{[3]}(i_1, i_2) + \gamma_x^2 \phi_0(i_1, i_2) + \omega_1^{[3]}(i_1, i_2) + \omega_2^{[3]}(i_1, i_2) + \omega_3^{[3]}(i_1, i_2) \]
(4)
where the functions \( \phi_0(.) \) and \( \omega_1^{[3]}(.) \) are defined by
\[ \psi_0(i_1, i_2) = \sum_n h_0(i_1 + n)h_0(i_2 + n)h_0(n) \]
\[ \omega_1(i_1, i_2) = 4 \sum_n h_0(i_1 + n)h_k(i_2 + n)h_0(n) \]
\[ \phi_0(i_1, i_2) = 4 \sum_n h_0(i_1 + n)h_0(i_2 + n)h_0(n) \]
\[ \omega_2(i_1, i_2) = 2 \sum_n h_0(i_1 + k)h_0(i_2 + n)h_0(n) \]
\[ \phi_0(i_1, i_2) = 8 \sum_{n} h(n,k)h(i_1+k,i_1+l)h(i_2+l,i_2+n) \]

**Remark 1** Taking the Fourier transform of (3) we obtain the expression of the output spectrum
\[ S_{yp}(\nu) = (\gamma_x + 2\gamma_x^2)[H_0(\nu)]^2 + 4\gamma_x^2 \sum_{k>0} |H_k(\nu)|^2 \]
and we see that \( S_{yp}(\nu) \) does not depend on the phases of the frequency responses \( H_k(.) \) for \( k \geq 0 \). Then we find as in the linear case that the phase cannot be recovered from the second-order characteristics of the output. Notice also that \( S_{yp}(\nu) \) gives only a sum of the modulus of the \( H_k(.) \)'s. Therefore, even if the \( H_k(.) \)'s are minimum phase, it is necessary to analyse the third-order properties of \( y(n) \).

**Remark 2** As \( y(n) \) is stationary, its statistical properties are independent of a time shift. Then any identification method based on cumulants of \( y(n) \) cannot distinguish between two kernels which only differ from a time shift. Therefore, since we consider the case where the kernel \( h(.) \) is non zero only for a finite number of its parameters, we can assume without loss of generality in all the following of this paper that \( h(i,j) \neq 0 \) only if \( (i,j) \in \{0, \ldots, n\}^2 \) where \( n \) is the order of \( h(.) \). In other words, \( h(i,j) \) can be considered as the \( (i,j) \)th element of a matrix whose \( l \)th diagonal is \( h_l(\cdot) \). Furthermore, we denote \( m_l \) the order of the subsystem \( h_l(.) \).

3 Blind identification of the parameters

In [2], we have established necessary and sufficient conditions to determine the order \( n \). We assume in this paper that \( n \) is known and we address the problem of recovering the parameters of the quadratic system using (2), (3) and (4). We give in property 1 a sufficient condition under which the non diagonal parameters can be expressed in terms of the diagonal parameters and of some output third-order cumulants. We show in property 2 that if this condition is not satisfied, then some systems are not identifiable. Now if it is satisfied, then the diagonal parameters are solution of a system of nonlinear equations. In property 3, we show that this solution is not unique. Lastly, a sufficient condition of uniqueness is proposed in property 4. Due to the space limitation the proofs of properties 2 and 3 are not reported here but may be found in [3].

**Property 1** If \( \gamma_x \neq 0 \) and \( m_0 = n \), then the coefficients \( h(k, l) \neq l \) can be expressed in terms of the coefficients \( h_0(k) \), \( \gamma_x \) and of some output third-order cumulants.

**Proof** We deduce from (4) that for all \( 0 \leq j < i \leq n \),
\[ C_{3y}(n + i - j, i) = \gamma_x \omega_2(n - j, -i) - 2\gamma_x \sum_{k=0}^j \sum_{i=1}^n h_0(n - j + k) h_0(l - i) h(l, k). \]
(7)
Let \( C_{ij} \) (resp. \( h_{ij} \)), \( 0 \leq j \leq n - 1 \), be the \( n - j \) dimensional vector whose \( p^{th} \) order component, \( 0 \leq p \leq n - j \), is \( C_{ij}(2n - j - p + 1, n - p + 1) \) (resp. \( h_{ij}(n - p + 1, j) \)). We deduce from (7) that
\[ C_{ij} = 2\gamma_x^2 \sum_{k=0}^j h_0(n - j + k) h_0(n - k) h_k \]
where \( H(n, n) \) is the \((n \times n)\) lower triangular Toeplitz matrix defined by
\[ H(n, n) = \begin{pmatrix} h_0(0) & \cdots & 0 \\
0 & \ddots & \cdots \\
h_0(n-1) & \cdots & h_0(0) \end{pmatrix} \]
and \( H(k, l) \), \( 1 \leq k, l \leq n \), is the matrix constituted with the \( k \) upper rows and the \( l \) left columns of \( H(n, n) \). Relation (8) can be written as
\[ C = 2\gamma_x^2 Hh \]
(9)
where
\[ C^T = (C_0^T, \ldots, C_{n-1}^T) \quad h^T = (h_0^T, \ldots, h_{n-1}^T) \]
and \( H \) is obviously defined. The matrix \( H \) is lower triangular and all its diagonal elements are equal to \( h_0(0) h_0(n) \). Then \( H \) is invertible if \( h_0(0) h_0(n) \neq 0 \).
which is equivalent to $m_0 = n$. Therefore, if $\gamma_{2r} \neq 0$ and $m_0 = n$, the vector $h$ which components are the coefficients $h(k, i)$ for $k \neq i$ can be expressed in terms of the coefficients $h_0(k)$ appearing in $H$, $\gamma_{2r}$ and of the output third-order cumulants element of $C$ by inverting the linear relation (9).

**Property 2** If $\gamma_{2r} = 0$ or $m_0 < n$, then it is not always possible to recover the non-diagonal coefficients even if the diagonal coefficients are known.

In the following section, we assume that $\gamma_{2r} \neq 0$ and $m_0 = n$. Then the vector $h$ of the non-diagonal coefficients can be expressed in terms of the diagonal parameters $h_0(k)$ up to a scaling factor by inverting (9). Therefore, it remains to determine the parameters $h_0(k)$. Of course, the important question is the one of the uniqueness of the solution up to a scaling factor. The existence of this factor can be seen by multiplying $z(n)$ by $\alpha$ and dividing $h(k_1, k_2)$ by $\alpha^2$ in (1) which let $\gamma(n)$ unchanged. Unfortunately, the solution is generally not unique as illustrated by property 3 hereunder. A sufficient condition of uniqueness which assumes that two output sequences are available is proposed in property 4. This situation arises when the input is under the control of the experimentalist but is not observable at the receiving end.

**Property 3** Different quadratic systems of the same order $n$ for which $m_0 = n$, driven by the same input satisfying $\gamma_{2r} \neq 0$, can have the same output cumulants up to the third order.

**Property 4** Assume $\sum_{j=0}^{n} h_0(j) \neq 0$, $h_0(.)$ is minimum phase, and two output sequences are available. One, denoted $y(n)$, corresponds to a non Gaussian input $x(n)$ satisfying $\gamma_{2r} \neq 0$, whereas the other, $\gamma(n)$, corresponds to a Gaussian input $\gamma(n)$. Then the kernel $h(.)$ can be recovered from the cumulants of $y(n)$ and $\gamma(n)$ up to the third order.

**Proof** As $\sum_{j=0}^{n} h_0(j) \neq 0$, it results from (2) that $C_{1y} \neq 0$, $C_{1y} \neq 0$ and that $\gamma_{2r} C_{1y} = \gamma_{2r} C_{1y}$. Then we deduce from (3) that

$$C_{2y}(i) - \left( \frac{C_{1y}}{C_{1y}} \right)^2 C_{2y}(i) = \gamma_{2r} \sum_{j=0}^{n} h_0(j) h_0(i + j) \quad (10)$$

and (10) is solved using the $z$-transform. As $h_0(.)$ is minimum phase, (10) has only two solutions which only differ by their sign. We denote $k_0(.)$ any one of these solutions and we have

$$k_0(.) = \epsilon \sqrt{\gamma_{2r}} h_0(.) \quad (11)$$

where $\epsilon = \pm 1$. According to (2) and (11),

$$\alpha = \frac{1}{C_{2y}} \sum_{j=0}^{n} k_0(j) = \epsilon \frac{1}{\gamma_{2r}} \frac{1}{C_{2y}}.$$

Now we calculate the matrix $K$ defined as $H$ but where $h_0(.)$ is replaced by $k_0(.)$ and we solve the linear equation $C = 2Kk$. Then according to (9) and (11),

$$k = \frac{\gamma_{2r}}{1/|\gamma_{2r}|} h. \quad (12)$$

According to (3) and (11)-(12), we have

$$C_{2y}(i) - \frac{2}{\gamma_{2r}} \sum_{j=0}^{n} k_0(j) h_0(i + j) =$$

$$4 \gamma_{2r} \gamma_{2r} \sum_{p \geq 0, q \geq 0} k_0(p) k_0(q) \gamma_{2r} (i + j). \quad (13)$$

If the right hand side of (13) is zero for $i = 0$, then the quadratic system is diagonal and the identification problem is solved very easily using for instance (3) and (4) whether $h_0(.)$ is minimum phase or not. Otherwise, (13) allows to determine

$$\beta = \frac{\gamma_{2r} \gamma_{2r} + 1}{\gamma_{2r}}.$$

Then multiplying (12) by $\alpha \beta$, we obtain

$$\alpha \beta \epsilon = \epsilon \sqrt{\gamma_{2r}} h. \quad (14)$$

The relations (11) and (14) constitute the identification up to a scaling factor of the parameters of the quadratic system.

If the function $h_0(.)$ is not minimum phase, the previous approach fails. Indeed, (10) admits at most $z^n$ solutions (up to the sign) obtained by reflecting one or more minimum phase zeros to reciprocal locations outside the unit circle and then (11) does not hold. Of course in this case, if several output sequences are available and that the cumulants of the corresponding inputs are known in such a way that the function $\phi_0(.)$ can be determined with (4), then according to (5) it is well known that $h_0(.)$ can be found knowing $\phi_0(.)$. Finally, the non diagonal parameters are recovered by the linear equation (9).

**4 Examples**

In this section, we consider two particular models of quadratic systems. We show in the first case that the parameters can be identified using the cumulants up to the third order of only one output sequence, and in the second that the coefficients $h_i(.)$ can only be recovered up to some time shifts which depend on $i$.

**4.1 Toeplitz kernel**

A Toeplitz kernel is defined by $h_k(i) = \alpha_k$ for all $0 \leq k \leq n$ and $0 \leq i \leq n - k$. The system is therefore characterized by $n + 1$ parameters and its order is equal to $n / \alpha_0 \neq 0$. The method presented in section 3 to determine the coefficients can be simplified as follows. We deduce respectively from (2) and (3) that

$$C_{1y} = (n + 1) \gamma_{2r} \alpha_0.$$
\(C_{2y}(n-i) = (i+1)(\gamma_{2e} + 2\gamma_{2e}^2 \alpha_i^2 + 4\gamma_{2e}^2 \sum_{k=1}^{i}(i-k+1)\alpha_k^2)\) for all \(0 \leq i \leq n\) and then 
\[C_{2y}(n-i) - 2C_{2y}(n-i+1) + \]
\[C_{2y}(n-i + 2) = 4\gamma_{2e}^2 \alpha_i^2. \tag{15}\]
for all \(1 \leq i \leq n\). Now taking \(j = 0\) in (7), we obtain that 
\[C_{3y}(n+i, i) = 2\gamma_{2e}^2 \alpha_i^2 \sum_{t=1}^{n} \alpha_t\]
for all \(1 \leq i \leq n\) from which we immediately deduce that
\[\xi_i = C_{3y}(n+i, i) - C_{3y}(n+i+1, i+1) = 2\gamma_{2e}^2 \alpha_i^2 \alpha_{i+1}. \tag{16}\]
for all \(1 \leq i \leq n\). If \(\gamma_{2e} \alpha_0 \neq 0\), the sign of \(\alpha_i\) is deduced from (16) and then using (15) we obtain \(\gamma_{2e} \alpha_i\) for all \(1 \leq i \leq n\). Finally, \(\gamma_{2e} \alpha_0\) is given by \(C_{1y}\). Then the identification of the parameters of the kernel is achieved up to the factor \(\gamma_{2e}\).

4.2 A class of kernels without a principal diagonal

We now consider an example where \(m_0 = 0\) and then the results presented in property 1 for determining the class of systems defined by
\[y(i) = 2 \sum_{k \in I} y_{2k+1}(i) \tag{17}\]
where \(I \subseteq \{0, \ldots, n\}\). As \(h_{0y}(.) = 0, C_{1y} = 0\) and the functions \(\phi_y(.)\) and \(\omega_y(.)\) are all zero. Furthermore we deduce from (17) that \(\phi_y(.) = 0\). Then \(C_{3y}(.) = \gamma_{2e}^2 \psi_y(.)\). We deduce from (3) and (6) that \(C_{2y}(.)\) and \(C_{3y}(.)\) are unchanged when any linear subsystem \(y_{2k+1}(.)\) is shifted by an arbitrary integer. Hence many quadratic systems in the class defined by (17) have the same order and the same output cumulants up to the third order. We now show that if the kernel satisfies some constraints which remove the previous degrees of freedom, then the system can be identified. Assume \(h_{2p+1}(0) \neq 0\) and \(m_{2p+1} \neq m_{2p+1} \) for all \((k, i) \in I^2, k \neq l\). Let \(m_{2p+1} = \max_{x \in \mathbb{E}}(m_{2k+1})\). We respectively deduce from (3) and (6) that \(C_{2y}(.)\) and \(C_{3y}(.)\) (if \(\gamma_{2e} \neq 0\)) have the same order \(2q\) where \(q = m_{2q+1}\). Then \(m_{2p+1}\) can be found from the order of \(C_{2y}(.)\) or if \(\gamma_{2e} \neq 0\) from the order of \(C_{3y}(.)\). Now
\[C_{3y}(q, i) = 8\gamma_{2e}^2 h_{2p+1}(i) h_{2p+1}(i) h_{2p+1}(0) \]
for all \(0 \leq i \leq q\). Then if \(\gamma_{2e} \neq 0\), we have
\[g(i) = \frac{\xi_i}{\sqrt{C_{3y}(q, i) C_{3y}(q, 0)}} = 2\gamma_{2e}^2 \alpha_i \alpha_{i+1} \]
which gives the parameters \(h_{2p+1}(.)\) up to the factor \(2\gamma_{2e}^2\). Replacing \(C_{3y}(t_1, t_2)\) by \(C_{3y}(t_1, t_2) - g(t_1 + t_2)\), we deduce from (6) that the previous approach for identifying \(h_{2p+1}(i)\) can be applied if \(I\) is replaced by \(I - \{p\}\). Then the kernel can be recovered up to a scaling factor by an iterative procedure.

5 Conclusion

The problem of blind identification up to a time shift and a scaling factor of any discrete time-invariant and finite extent quadratic system driven by an unobservable sequence was addressed. The input \(x(i)\) was assumed i.i.d., no hypothesis on the \(p\)th-order cumulants \(\gamma_{p\text{e}}\) of \(x(i)\) was made, and our approach was based on the use of the output cumulants of \(y(i)\) up to the third order. The following results were established. If \(\gamma_{2e} = 0\) or \(m_0 = 0\), it is not always possible to recover the coefficients \(h(k, l)\) for \(k \neq l\) even if the coefficients \(h(k, l)\) are known. If \(\gamma_{2e} \neq 0\) and \(m_0 = 0\), the coefficients \(h(k, l)\) for \(k \neq l\) can be expressed in terms of the coefficients \(h(k, l)\), \(\gamma_{2e}\) and of some third-order output cumulants. The parameters \(h(k, l)\) then satisfy a system of nonlinear equations but unfortunately the solution of this system is generally not unique. Then a sufficient condition of uniqueness was proposed. This condition assumes that the system \(y(0)\) is minimum phase, satisfies \(\sum_{k=0}^{\infty} h(k, l) \neq 0\) and that two output sequences are available. One output must correspond to a non Gaussian input, whereas the other must correspond to a Gaussian input.

References


Blind Separation of Sources: A Comparative Study of a 2-nd and a 4-th Order Solution

S. VAN GERVEN, D. VAN COMPERNOLLE, H.L. NGUYEN-THI\textsuperscript{†} and C. JUTTEN\textsuperscript{†}

K.U.Leuven - ESAT, K. Mercierlaan 94, 3001 Heverlee, Belgium, E-Mail: vgerven@esat.kuleuven.ac.be
\textsuperscript{†}L.N.P.G. - TIRF, 46 Av. Felix Viallet, 38031 Grenoble Cedex, France, E-Mail: inguyen@tirf.ingp.fr

Abstract. This paper compares optimal performance and parameter settings in signal separation algorithms. It presents the different parameters involved in building a fully adaptive signal separator and studies the influence of these parameters on the performance of different algorithms. Two algorithms, built around different separation criteria, are compared in the case of convolutive mixtures modeled by the strictly causal FIR filters. Parameters were varied to find optimal settings. It was found that with equal optimisation the decorrelation based algorithm and the fourth order moment based algorithm perform quite identical. Because parameter optimisation relies heavily on feedback from the original sources, which are not available in a real-life situation, test simulations were performed in which the optimum settings, 'learned' from some examples, were frozen. The test simulations show a similar sensitivity to changes in mixture characteristics for both algorithms.

1 Introduction

The common philosophy in recent signal separation solutions is to use a symmetric adaptive filter in a feedback structure as shown in figure 1. The key differences are found in the type of separation criterion used, which is directly reflected in the adaptation rule for the adaptive filter coefficients.

Furthermore parameter settings play a crucial role in algorithm performance. Unfortunately parameter settings are not always described extensively. A correct comparison of separation criterias is possible only if algorithms are equally developed up to this separation criterion. According to the work done in [1, 2, 3, 4], this paper will focus on the comparison of 2 different criterias: output decorrelation and cancellation of fourth order moments. Performance as well as optimal parameter settings will be compared.

![Figure 1. General Structure for Symmetric Adaptive Signal Separation](image)

2 Parameters

From the detailed description of our adaptive signal separation algorithm, 7 parameters appear that will influence the performance. Two parameters are related to the adaptation speed. Three parameters are used in the estimation of energies and (cross-)moments. Finally some parameters are introduced to stop adaptation under unfavourable signal conditions.

All moments and cross-moments are estimated consistently with a first order autoregressive processing each involving a time constant $\tau$. With general notations this gives:

$$\alpha = 1 - \frac{1}{\tau}$$  (1.a)

$$M_{pq}^{(k+1)}(x,y) = \alpha M_{pq}^{(k)}(x,y) + (1-\alpha)e^p(k)\delta^q(k)$$  (1.b)

The exact definitions of the different moments and time constants involved are given below.

2.1 Adaptation Speed

The adaptation speed is tuned by two adaptation constants, $\mu_1$ and $\mu_2$, for filter $W_1(k)$ and $W_2(k)$ respectively. These adaptation constants are internally normalized. For the decorrelation approach they are divided by an estimate of the energy in the signal. For the method based fourth order moments the adaptation constants are divided by a fourth order moment estimate.

2.2 Centered Signals

Calculations of the correlations and fourth order moments is done with the zero mean signals. The time constant for the estimation of mean of the signals is called $\tau_m$. We will denote the signal itself with an upper case letter and the centered signals with lower case letters.

$$M_i(U_i) = E[U_i]$$  (2.a)

$$u_i(k) = U_i(k) - M_i(U_i)$$  (2.b)
2.3 Energy Estimate

The energy estimate is also obtained with a first order autoregressive process involving the time constant $\tau_e$. The energy estimate is used for the intermittent adaptation decision discussed below.

$$M_3(U_i) = E[U_i^2]$$ (3)

The same time constant $\tau_e$ is also used to calculate the second and fourth order moments of the centered signals. These quantities are needed for the normalisation of the adaptation constants.

$$M_2(u_i) = E[u_i^2]$$ (4.a)
$$M_4(u_i) = E[u_i^4]$$ (4.b)

2.4 Cross-moment Estimate

The second and fourth order cross-moments of the centered signals are used as separation criteria. They are estimated involving a third time constant $\tau_{m}$.

$$M_{11}(u_i, u_j, l) = E[u_i(l)u_j(l)]$$ (5.a)
$$M_{31}(u_i, u_j, l) = E[u_i^3(l)u_j(l)]$$ (5.b)

2.5 Intermittent Adaptation

Intermittent adaptation is known from classical adaptive filtering. For symmetric signal separation, the problem is even more critical as explained in [2, pp. 57-59]. The idea is not to adapt the filter if the energy of the output signal fed into the filter is much lower than the energy of the other output signal. This is because at such instants there is not enough information available on the signal in the filter compared to the other signal, which would result in large undesired variances of the coefficients.

To implement intermittent adaptation adaptively 2 more parameters, $\theta_1$ and $\theta_2$, called the adaptation thresholds, appear in the system. The intermittent adaptation rule can be stated as:

if $M_2(U_3) < \theta_1 M_2(U_1)$ do not adapt $W_i(k)$ (6.a)
if $M_2(U_1) < \theta_2 M_2(U_2)$ do not adapt $W_2(k)$ (6.b)

In the case of a mixture of two speech signals, where two silence periods might coincide, an absolute silence threshold can be added. This threshold must then be chosen as a function of the dynamic range of the input signals.

2.6 Discussion

It must be clear from the above that optimal parameter settings can depend both on the separation criterion and on the signals. As analytical relations are very hard to derive in this double feedback structure, only experimental results can reveal the different influences. We will describe in section 4 how we proceeded to obtain optimal parameter settings.

3 Algorithm Implementation

In section 2 some important aspects of the algorithm were already clarified. Here the algorithm is stated more explicitly.

The input signals $Y_1(k)$ and $Y_2(k)$ are convolutive mixtures of the independent sources $S_1(k)$ and $S_2(k)$:

$$Y_i(k) = S_i(k) + H_i * S_j(k) \quad i, j \in \{1, 2\} \quad i \neq j$$ (7)

The output signals are obtained via a double feedback adaptive filter:

$$U_i(k) = Y_i(k) - W_i(k) * U_j(k) \quad i, j \in \{1, 2\} \quad i \neq j$$ (8)

The filters $H_1$ and $H_2$ are assumed to be strictly causal FIR filters. The strict causality constraint is needed for the decorrelation based criterion to work. It is not within the scope of this conference paper to discuss the validity and limitations of this simplified signal model for convolutive mixtures.

The filter coefficients are updated at each sample instant except if the intermittent adaptation rule applies:

$$W_i(k + 1) = W_i(k) + \Delta W_i(k) \quad i \in \{1, 2\}$$ (9)

with the elements of $\Delta W_i(k)$ given by:

for 2-nd order:

$$\Delta w_i(l) = \frac{\mu_i}{M_2(u_i)} M_{11}(u_i, u_j, l)$$ (10.a)

for 4-th order:

$$\Delta w_i(l) = \frac{\mu_i}{M_4(u_i)} M_{31}(u_i, u_j, l)$$ (10.b)

4 Test Procedure

To compare the performance of the different separation criteria and the optimal parameter settings in both cases, we developed an original test procedure. This test procedure consisted of a ‘training-phase’ and a ‘test-phase’.

4.1 Classes

First of all we divided our experiments in 5 different classes representing mixtures of different types of signals (table 1). The noises which were always stationary and artificially generated. The speech signals were French sentences recorded at 8kHz, the length of the sentences varied from 1.5s to 2.5s.

<table>
<thead>
<tr>
<th>class</th>
<th>source 1</th>
<th>source 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>white noise</td>
<td>white noise</td>
</tr>
<tr>
<td>2</td>
<td>colored noise</td>
<td>colored noise</td>
</tr>
<tr>
<td>3</td>
<td>speech</td>
<td>white noise</td>
</tr>
<tr>
<td>4</td>
<td>speech</td>
<td>colored noise</td>
</tr>
<tr>
<td>5</td>
<td>speech</td>
<td>speech</td>
</tr>
</tbody>
</table>

Table 1. Division of mixtures in 5 different classes

4.2 Training

During the ‘training-phase’ we created 1 example in each class and optimised parameter settings for it. For this example we choose broad FIR low-pass mixture filters. The norm of the filters $|H_i|$ varied from 0.7...0.85 and the order of the mixing filters was 10...15.

To keep the optimisation task within acceptable time limits, we agreed after some preliminary trials on fixing the parameters $\tau_m$ and $\tau_e$ identical for both criteria. The algorithms are fairly robust to changes in these parameters within a range of about 20...200. We found optimal settings for $\theta_2$ to be very close to each other for both criteria and we also agreed on taking identical $\theta_2$. Finally we optimised on the parameters $\mu_i$ and $\tau_e$ independently. As a consequence we might not have found the optimum optimum, but the parameter search was sufficiently exhaustive to ensure that we are close to that optimum. Results of the training phase are given in section 5.
4.3 Tests

During the ‘test-phase’ we created 8 different mixtures (table 2) in each class and tried to separate ‘blindly’ with the optimal parameter settings obtained from the training. Each test corresponds to a different modification in the mixture and reveals the sensitivity of parameters and optimal performance to changes in the sources, the filters etc... Such changes can and will occur in real-life situations as well.

<table>
<thead>
<tr>
<th>test</th>
<th>modification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>scaling both sources by 0.1</td>
</tr>
<tr>
<td>2</td>
<td>scaling source 2 by 0.7</td>
</tr>
<tr>
<td>3</td>
<td>scaling (</td>
</tr>
<tr>
<td>4</td>
<td>scaling (</td>
</tr>
<tr>
<td>5</td>
<td>using single delay FIR filters</td>
</tr>
<tr>
<td>6</td>
<td>using resonant FIR mixing filters</td>
</tr>
<tr>
<td>7,8</td>
<td>different sources</td>
</tr>
</tbody>
</table>

Table 2: Modifications applied in the test mixtures

The goal of test 1 was to check the absolute level sensitivity. Test 2 checks the relative level sensitivity and tests 3 and 4 the sensitivity to mixture strength and input SNR levels. Tests 5 and 6 investigate the filter type sensitivity and tests 7 and 8 the sensitivity to source characteristics.

The only a priori knowledge during the tests was that filterlengths nowhere exceeded 15 taps and that the speech signal was always on channel 1. Such a priori knowledge might be fair in practical situations although a filterlength of 16 taps is still far off from realistic lengths of acoustic transfer functions.

5 Results

5.1 Performance measure

The performance of the separation algorithm was measured using the separation error. This is the residual error at the output, relative to the source signal and is expressed in dB:

\[
Err_r = 10 \log_{10} \left( \frac{E[(S_i - U_i)^2]}{E[S_i^2]} \right)
\]  

(11)

For all experiments a mean value was calculated starting from 600 ms, up to the end of the signal. This represents a compromise between convergence speed and residual error. Remark that for classes 3,4 and 5, where the intermittent adaptation rule applies, this does not represent an identical number of coefficient updates. On the other hand speech pauses in \(S_i\) were not taken into account for the calculation of mean of the related separation error \(Err_r\).

5.2 Parameter settings

Optimal parameter settings were found after an extensive search in the parameter space. As already mentioned the \(\tau_s\) and \(\tau_m\) were fixed a priori. For class 1,2 we choose \(\tau_s = \tau_m = 100\) for classes 3,4 and 5 \(\tau_s = \tau_m = 50\). The \(\theta_i\) were optimised but identical for both criteria. For classes 1,2 the adaptation was permanent. For classes 3,4 and 5 the optimal values for \(\theta_1, \theta_2\) were respectively \((0.6, 0.6)\), \((0.4, 0.6)\) and \((0.6, 0.5)\). The optimal settings for the other parameters are given in table 3.

<table>
<thead>
<tr>
<th>class</th>
<th>parameter settings</th>
<th>4-th order mom.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\tau_s)</td>
<td>(\mu_1)</td>
</tr>
<tr>
<td>1</td>
<td>200</td>
<td>5e-3</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>10e-3</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>20e-3</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>20e-3</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>5e-3</td>
</tr>
</tbody>
</table>
classes 4 and 5 (highly colored signals). We discuss the results per test rather than per class since the performance deviations were not strongly class dependent. For the modifications applied see table 2. Instead of giving a quantitative overview of all 80 separation experiments, we will try to describe qualitatively the most important results of the tests.

- For test 1 the performances were almost identical to the training results, confirming the linearity of the algorithm and the correct use of internal normalisation.

- For test 2 the performance deteriorates for all classes in channel 1 (Err1 \( \gtrsim \)) and improves for all classes in channel 2 (Err2 \( \lesssim \)). Intuitively we expected the inverse to happen: if the input SNR in channel 1 increases we expected that source 1 should be easier to recover and hence that the separation error in channel one should decrease. The inverse behaviour is not easy to understand and theoretical calculations learned that the result will depend on the existing ratio of the source energies and on the norms of the mixing filters too. Test 2 proves that it is not necessarily the strongest source that will be recovered best.

- Test 3 and 4 are similar tests, changing the norm of filter \( \mathcal{H}_1 \) only (cfr. table 2). For test 3 the performance deteriorates with exactly the same amount in both channels with about 4dB on average since \( \|\mathcal{H}_1\|/\|\mathcal{H}_2\| \sim \) and hence the mixture is more difficult to separate. Test 4 represents an easier task and the performance improves in both channels with about 5dB on average. It is remarkable that the performance changes were identical in both channels (for all classes and for both algorithms) although we only changed the properties of filter \( \mathcal{H}_1 \). For classes 3,4 and 5 the performance changes were less severe probably due to the intermittent adaptation thresholds. Globally tests 3 and 4 learned that performance is highly sensitive to changes in the mixture strength, independent of signals and separation criteria used. We therefore suggest that all authors should specify the norms of their mixing filters when presenting signal separation results.

- In test 5 both mixing filters were of the form \( \mathcal{H}_i \delta(t-\tau) \) with \( \tau \) of the order 5 . . . 8. We kept on adapting the complete filter because no a priori knowledge on the filter was assumed. The performance increases for both criteria and for all classes, except for class 5, but the differences were not that important. We concluded that a single delay filter represents a slightly easier separation task.

- In test 6 we found performance improvements for all classes except class 1. This can be understood since class 1 contains white noise signals and resonant filters stress only a small part of the spectrum so that some information is lost. The improvements in the other classes can be explained since in most of the cases the resonant filters match more or less the spectra of the sources, so that they are introducing extra discrimination which makes the separation task easier.

- For tests 7 and 8 we saw different results for the different classes and separation criteria did not always showed the same behaviour. Results in classes 3,4 and 5 showed a greater variability compared to classes 1 and 2. This can be explained by the time varying nature of the signals and the signal dependent settings of the intermittent adaptation thresholds.

5.5 Discussion

As a conclusion on the test results we can state that changes in the mixture affect the performance in both channels equally. This emphasizes the strongly coupled nature of the algorithm. Changes in the sources lead to less predictable results although performance remained acceptable in all tests. In all cases both separation criteria behave comparably. Searching for the optimal parameter settings for each test example will surely give additional and interesting information on the role of the parameters. This however represents a tremendous work and is far beyond the scope of this paper.

6 Conclusion

In this paper we compared performances and parameter settings for different signal separation criteria. From the optimal results we conclude that both criteria lead to nearly identical performances. Since complexity in order 2 is less compared to order 4, we may state that in the case of strictly causal convolutive mixture, there is no need to use higher order statistics to achieve good separation. The key to high separation performance lies in a good choice of the parameters. While the time constants involved are fairly robust, the choice of the adaptation constants requires more attention. In the case of non-stationary signals intermittent adaptation thresholds should be chosen carefully too. Some knowledge about the dynamic range and the relative levels of the sources is quite useful for a good parameter choice and hence for successful signal separation.

Acknowledgements

This work was supported by ATHOS, Esprit Basic Research Working Group 6620, and by the CNRS, France. Dirk Van Campenolle is Senior Research Associate of the National Fund for Scientific Research of Belgium (N.F.W.O.). Stefaan Van Gerven is Research Assistant of the I.W.O.N.L.

References


COMPlex SELF–ADAPTIVE ALGORITHMS FOR SOURCE SEPARATION BASED ON HIGH ORDER CONTRASTS

Eric MOREAU and Odile MACCHI

Laboratoire des Signaux et Systèmes, CNRS-Supelec, Plateau de Moulon, 91192 Gif-sur-Yvette, FRANCE,
and Groupement de Recherche TdSI du CNRS.
Tel/Fax: 33 1 69 41 80 40 / 60 19 35 46, E-Mail: moreau@lss.supelec.fr, macchi@lss.supelec.fr

Abstract. This paper presents some algorithms for source separation in the case of complex signals. A generalized
complex version of the Herault-Jutten (HJ) adaptation law is presented for source separation. New criteria based
on the maximization of high-order contrast functions are proposed. They contain fourth order cumulants. Two
stochastic gradient based adaptive algorithms are derived in order to optimize the proposed contrast functions.
They don't require a prewhitening stage prior to the proper separation. Achievement of these three adaptive
algorithms is demonstrated by computer simulations.

1 Introduction

The source separation problem is currently receiving an
increased interest because of practical applications in
diverse fields of engineering and applied sciences like
communications and array processing. It can be simply
formulated as follows: several unknown linear mixtures
of certain independent random signals called sources are
observed. Using an adaptive system, the objective is to
recover the unknown original sources without knowing
the mixture parameters.

It is well known that methods based on second order
statistics (i.e. correlation or power spectrum) are not
sufficient to solve the problem. The first designed adap-
tive algorithms (available in the real case) include non
linearity functions in the updating increments [1],[2].
However these adaptation laws may converge to a non
separating state, depending on the density of the sources
and on the initial state.

Recently algorithms based on maximization of a con-
tраст criterion were proposed. Some of them require a
prewhitening stage prior to the proper separation [3]-
[6]. In the approaches of [7],[8] this preprocessing is not
necessary, but signals cannot be complex.

In this paper, we generalize the one-stage separation
system of [7],[8] to complex signals. Indeed, we are able
to build up contrast criteria that impose restriction nei-
ther on the signals (white or not, circular or not) nor on
the signs of the kurtosis of sources. Two self-adaptive
algorithms are derived in order to maximize the pro-
posed contrasts. Computer simulations corroborate the
presented theory and show ability to adaptively realize
source separation, at least in the presented situations.

2 Problem formulation

The classical linear memoryless noise free model is con-
sidered. It reads \( x = G a \) where

\[ \bullet x \text{ is the } (n,1) \text{ vector of observations; } \]

\[ \bullet a \text{ is the vector formed by the } n \text{ sources } a_i \text{ which are as- } \]

sumed to be zero-mean, statistically independent, with
unit-variance and non-zero kurtosis;

\[ \bullet G \text{ is the } (n,n) \text{ square mixture matrix. } \]

The source separation problem consists in estimating
a linear system \( H \) operating on the observations \( x \), such
that its \( n \) outputs \( y_i, i=1,\ldots,n \) restores the \( n \) sources.
This transformation reads

\[ y = H x \]

(1)

where \( y \) is the \((n,1)\) vector of outputs. It has been
shown [9] that this problem is equivalent to generating
\( n \) outputs \( y_i \) statistically independent. If \( y \) has inde-
dependent components, and if a matrix \( Q \) has one and only
one non-zero element in each row and each column, then
the vector \( Q y \) also has independent components. Such
a matrix \( Q \) can be factorised in \( Q = A P \) where \( A \) is an
invertible diagonal matrix and \( P \) a permutation matrix.
Thus independence of the outputs is satisfied if and only
if (iff) the estimated matrix \( H \) satisfies the relationship

\[ S \overset{\Delta}{=} H G = A P , \]

(2)

where \( A \) and \( P \) are arbitrary and correspond to undeter-
derd parameters. The specific value of \( A \) corresponds
to arbitrary attenuations for restored sources, while the
\( P \) value corresponds to an arbitrary order. \( H \) is called
a "separating matrix".

3 Contrast functions

In the context of source separation, contrast functions
were first defined by Comon [9]. They constitute sepa-
ration criteria in the sense that their global maxima

\[ \text{correspond to separating states. Thus the problem re- } \]

duces to an optimization one. For complex signals a
contrast is defined as follows:
**Definition:** A contrast function is a multivariate mapping \( J(\cdot) \) from a space of random variables in \( \mathbb{C}^n \) onto \( \mathbb{R} \) satisfying the four requirements

R-1. \( J(y) \) depends only on the probability density of \( y \) and is symmetrical w.r.t. the components \( y_i; \)

\[ \forall P \text{ permutation matrix, } J(Py) = J(y); \]

R-2. \( J(y) \) is invariant under a scale change:

\[ \forall A \text{ diagonal invertible matrix, } J(Ay) = J(y); \]

R-3. A linear mixture of independent components can only decrease the contrast:

\[ \forall \alpha \text{ independent vector, } \forall S, J(Sa) \leq J(\alpha); \]

R-4. Only the permutations and scale changes keep unchanged the contrast of independent sources:

\[ \forall \alpha \text{ independent vector, } \forall S, J(Sa) = J(\alpha) \iff S = AP. \]

The fourth requirement specifies that those matrices \( S \) which maximize the contrast satisfy the relation (2). Thus the corresponding \( H \) is a separating matrix and the problem is solved. Before presenting contrast functions, it is useful to define fourth-order cumulants. For zero-mean complex variables \( y_1, y_2, y_3 \) and \( y_4 \), they can be defined as:

\[ \text{cum}[y_1, y_2, y_3, y_4] := E[y_1 y_2 y_3 y_4^*] - E[y_1 y_2 y_4^*] E[y_3 y_4^*] - E[y_1 y_3 y_4^*] E[y_2 y_4^*] - E[y_1 y_2 y_3 y_4^*] \]

(3)

where the * denotes complex conjugation. Let

\[ \gamma_{y_i} \triangleq \text{cum}[y_1, y_2, y_3, y_4] = E[|y_1|^4] - 2E[|y_1|^2 |y_2|^2] - E[|y_1|^2] E[|y_2|^2] \]

(4)

\[ \gamma_{y_{ij}} \triangleq \text{cum}[y_i, y_j, y_i, y_j] = E[|y_i|^4] - 2E[|y_i|^2 |y_j|^2] - E[|y_i|^2] E[|y_j|^2] \]

(5)

\[ \chi_{y_i} \triangleq E[|y_i|^2], \quad \lambda_{y_i} > 0. \]

Notice that \( \gamma_{y_i} \) and \( \gamma_{y_{ij}} \) are real numbers, so their signs are well defined.

In his early definition [9], Comon chooses

\[ J_B(y) = \sum_{i=1}^{n} \gamma_{y_i}^2. \]

(7)

But his definition is restricted to spatially "white" vectors \( y \), i.e. \( E[y y^H] = I \) where the superscript \( H \) means "transpose and conjugate". It means that source separation requires two successive stages, first the whitening stage, then the stage maximizing (7). In the following we generalize the contrast (7) to non white complex random variables. Therefore source separation will be performed in a single stage.

**Statement 1.** If the fourth-order cumulants (4) of all the sources \( a_i \) have the same sign, then the function

\[ l(y) = \sum_{i=1}^{n} \gamma_{y_i} - \alpha \sum_{i<j}^{n} \gamma_{y_{ij}} \]

(8)

where \( \alpha \geq 1 \), is a contrast over the set of normalized random vectors \( y \) (i.e. \( E[|y|^2] = 1 \), \( \forall i \)).

**Statement 2.** If the fourth-order cumulants (4) of all the sources \( a_i \) have the same sign, then the function

\[ l(y) = \sum_{i=1}^{n} \gamma_{y_i} - \alpha \sum_{i<j}^{n} \gamma_{y_{ij}} \]

(9)

where \( \alpha \geq \cdot \), is a contrast for all random vectors \( y \).

**Statement 3.** The function

\[ J_N(y) = \sum_{i=1}^{n} \gamma_{y_i}^2 - \alpha \sum_{i<j}^{n} \gamma_{y_{ij}}^2 \]

(10)

where \( \alpha \geq 1 \), is a contrast over the set of normalized random vectors \( y \) for \( n = 2 \).

**Statement 4.** The function

\[ J(y) = \sum_{i=1}^{n} \left( \frac{\gamma_{y_i}}{\lambda_{y_i}} \right)^2 - \alpha \sum_{i<j}^{n} \left( \frac{\gamma_{y_{ij}}}{\lambda_{y_i} \lambda_{y_j}} \right)^2 \]

(11)

where \( \alpha \geq \cdot \), is a contrast for all random vectors \( y \) for \( n = 2 \).

These statements can be proved in a way similar to the proof given in [8] for the real case, by simply introducing squared moduli when suitable.

## 4 Self-adaptive algorithms

In order to find a matrix \( H \) which separates the sources, stochastic adaptive algorithms are now proposed to maximize the above contrasts. Because the \( n \) diagonal elements of \( A \) are arbitrary, among the \( n^2 \) complex unknowns of \( H \), \( n \) elements can be fixed in order to get a well-posed problem. One way is to set at 1 the \( n \) diagonal entries of \( H \)[10].

### 4.1 Complex HJ adaptation law

As a basis for comparisons, a generalized complex version of the well-known Herault-Jutten adaptation law is presented. The original (HJ) structure for \( H \) is a recursive one [1], but a direct structure can be used instead (see [10]) while retaining the same adaptation law. In the real case, with a positive constant step-size \( \mu \), the increment after each iteration is

\[ \Delta h_{tm} = -\mu f_1(y_t) f_2(y_{tm}), \quad t \neq m \]

(12)

where \( f_1(\cdot) \) and \( f_2(\cdot) \) are two different odd functions. The simplest choice is \( f_1(u) = u^3, f_2(u) = u \). Here we propose the generalized complex form below

\[ \Delta h_{tm} = -\mu |y_t|^2 y_t y_{tm}^*, \quad t \neq m. \]

(13)
Computer simulations:
The algorithm performance is illustrated for \( n = 2 \) sources. Two types of sources are considered:

**type 1.** \( a_i \) admits the four equiprobable values: \( \pm 1 \pm j \);  
**type 2.** \( a_i = \exp(j\phi) \) where \( \phi \) is a random variable with a uniform probability density over \([0, 2\pi]\).

The performance index associated to the matrix \( S \) is defined as [7]:

\[
p(S) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{|s_{ij}|^2}{\max_i |s_{ij}|^2} - 1  
+ \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{|s_{ij}|^2}{\max_i |s_{ij}|^2} - 1
\]  

(14)

In Fig.1 (resp. Fig.2), it is plotted versus iterations for two sources of type 1 (resp. type 2). The performance index converges to zero which means that separation is realized. Indeed there is one and only one non zero element per row and column in \( S \).

4.2 Gradient based adaptive algorithms

Now consider maximization of the contrasts (3)-(11). Although it is indeed possible to write the gradient algorithms associated to the contrasts (10) and (11), the squared values make it cumbersome. We restrict the present contribution to maximization of the contrasts (8) and (9). It means that all the \( \gamma_{ni} \)'s have the same sign called \( \epsilon \). Moreover, for simplicity, sources are assumed to be circular and thus

\[
E[|y_i y_j|] = 0, \quad \forall i, j
\]  

(15)

Maximization of \( I_N(y) \)

The computation of (8) requires that all the \( y_i \)'s have unit power. This is easily obtained thanks to \( n \) automatic gain controls (AGC) placed on the \( n \) output channels as indicated in Fig.3 with \( n = 2 \). The positive constants \( f_i \) are worth \( E[|y_i|^2]^{1/2} \).

The non-diagonal entries of \( H \) are optimized thanks to the gradient algorithm:

\[
h_{tm}(k+1) = h_{tm}(k) + \mu \frac{\partial I_N}{\partial h_{tn}} H = H(k)
\]  

(16)

where \( \ell \neq m \) and \( \mu \) is a positive step-size. From (1)

\[
\frac{\partial y_i}{\partial h_{tm}} = \delta_{it} e_m, \quad m \neq i
\]  

(17)

where \( \delta_{it} \) is the Kronecker symbol. From (4), (5) and (8), a stochastic version of (16) can be written

\[
h_{tm}(k+1) = h_{tm}(k) + \mu \epsilon e_m[2|y_i|^2 y_i
\]

\[
- \alpha \sum_{i \neq t=1}^{n} \frac{(|y_i|^2 y_t - y_i E[y_i y_t])}{E[|y_i|^2]} H = H(k)
\]  

(18)

The unknown moments in (18) have to be estimated thanks to the available data. A classical adaptive procedure is

\[
\hat{E}[y_i y_t](k) = (1-\mu) \hat{E}[y_i y_t](k-1) + \mu y_i y_t(k) \quad (19)
\]

where \( \mu \) is a positive step-size. The AGC procedure is realized in the same time thanks to

\[
d_i(k) = d_i(k-1) + \mu_a [1-|y_i(k)|^2]  
\]

(20)

\[
f_i(k) = \sqrt{d_i(k)}
\]

(21)

Maximization of \( I(y) \)

Following the same steps as before, one can easily derive a stochastic algorithm maximizing \( I(y) \):

\[
h_{tm}(k+1) = h_{tm}(k) + \mu \epsilon e_m[2|y_t|^2 E[|y_i|^2] E[|y_i|^2] - E[|y_i|^4]]
\]

\[
- \alpha \sum_{i \neq t=1}^{n} \left( \frac{|y_i|^2 y_t - y_i E[y_i y_t]}{E[|y_i|^2 E[|y_i|^2] - E[|y_i|^4]]} \right) \quad (22)
\]

The unknown moments are estimated like in (19).

Computer simulations:

In Fig. 4 (resp. 5) are depicted the performances of the adaptive algorithms (18) with AGC and (22) without AGC for two sources of type 1 (resp. type 2). Both algorithms achieve source separation because \( p(S) \) reaches zero. Comparing Fig. 1 and 2 to Fig. 4 and 5 respectively, it is clear that the new contrast functions permit a better achievement than the HJ algorithm. It is, however, at the price of a higher computational complexity.

References


Figure 3: Direct separation with output AGC’s

Figure 1: Complex HJ algorithm, sources of type 1

Figure 2: Complex HJ algorithm, sources of type 2

Figure 4: contrast algorithms, sources of type 1

Figure 5: contrast algorithms, sources of type 2
A general adaptive algorithm for nonGaussian source separation without any constraint

F. HARROY (1), J.L. LACOUME(1), M.A. LAGUNAS (2)

(1) CEPhAG - ENSIEG INPG CNRS URA 346
BP 46
38 402 Saint-Martin d’Hères Cedex
France
Fax : (33):76:82:63:84
(2) E.T.S.I. Telecm
Apdo 30.002
68080 Barcelona
Spain
Email : harroy@cephag.observ-gr.fr

Abstract. This paper deals with the blind source separation. The task consists in separating some independent and linearly mixed signals called sources. After some general remarks, the model is recalled and our approach based on the Maximum-Likelihood principle and on the higher-order statistics (HOS) is introduced. The main stages of the calculation are presented leading to the criterion of the separation based on a sum of squared cumulants of the sources at the fourth order. The second part is devoted to the adaptive implementation which is in opposition to the block treatment. The procedure using the gradient calculus is described.

Some results obtained in simulations are shown, they correspond to the case of a mixture of two real valued sources. Finally, an example of a possible integration in a communications system based on multidimensional beamformers is briefly shown. But some tests on real data should be carried out beforehand.

INTRODUCTION

In many physical problems, the recovering of signals corrupted by noise and mixed anyway is the main task, generally named: Filtering. This study belongs to the general issue of blind source separation. It means that no knowledge is available concerning the mixing process and the nonGaussian source signals. Then, anyone must take in mind that the sources cannot be reached in a physical way and only the outputs of a sensor array are available.

For a long time, the Gaussianity assumption, as a derivation of the central limit theorem, was adopted as a last resort, but nature rarely delivers such kinds of signals. Moreover, the recent introduction of the higher-order statistics has brought a new renewal of interest to go further in the characterization of non Gaussian processes. Hence, a variety of methods have appeared with two different levels. The first concerns the theoretical background that can be heuristic [4] or based on statistical ideas, even on algebraic basis [2], whereas the second is relevant to the type of implementation: block treatment or adaptive way. Though, for all kinds of treatment, a performance analysis is useful: the Cramer-Rao bounds for the block treatment and the asymptotic behaviour of the estimators [1] for adaptive treatments.

The first part of this paper is devoted to the explanation of the adopted physical model and the major principles supporting the point of view. Then, the second section is devoted to a method to solve the question. We present a complete adaptive implementation of an algorithm without constraint on the kind of sources (short-tailed or long-tailed random sources). Simulations that have been carried out illustrate the work in the third section. In the last part which concerns the conclusions and the forecasts, an example of a possible integration in a communication system is briefly shown.

1. The theoretical backgrounds

1.1 The physical model

Having an \( n \) sensor array and \( p \) statistically independent sources mixed linearly, the most general writing of the linear model is the following:

\[
x_i(t) = \sum_{j=1}^{P} a_{ij}(t) s_j(t) + b_i(t) \quad i \in \{1, \ldots, n\}
\]

where \( a_{ij}(t) \) represent the transfer functions between sources and sensors and \( b_i(t) \) the \( i \)-th component of the corrupted noise assumed to be spatially white and Gaussian and equi-powerful (\( \Rightarrow \sqrt{E[b_i^2]}=c \)).

Choosing a treatment in time domain for no-delay mixture or in frequency domain for convolutive mixture, the model can be summarized with a linear relation: \( x(n) = A x(n) + b(n) \) between the observations \( x(n) \) and the sources \( s(n) \). Most of the methods uses the higher-order statistics, more precisely the fourth-order cumulant-tensor. Because of an assumption on the even source probability density function (pdf), the third-order cumulants are null. Thus, the elements of the cumulants at the fourth order in the real case are equal to:

\[
\kappa_{i,j,k,l}(X) = E[X_i X_j X_k X_l] + E[X_i X_j] E[X_k X_l]
\]

The notation between brackets [3] holds for all permutations over the indexes.

Now, many approaches have been applied. One among the first was a heuristic one [4] consisting in maximising the inverse of the sum of the squared source cross-cumulants. Afterwards, an original approach has been introduced by I.F. Cardoso, using the multi-linear algebra on the fourth-order cumulant-tensor named Quadrat-
covariance. Theoretically, this method permits to estimate more sources than the number of sensors. Others founded on statistical tools like the Maximum Likelihood (ML) have also appeared [10].

1.2 A method using the ML principle

One of them [Gaeta-Lacourme] use this principle to estimate the parameters of the problem. Before evoking the steps of the calculation, it is important to remember the asymptotic properties of the ML-estimators. They are asymptotically consistent, efficient, unbiased, Gaussian. At first, the mixture matrix \( \mathbf{A} \) is decomposed as a general product of three matrices: \( \mathbf{A} = \mathbf{V} \Sigma \mathbf{P} \), where \( \mathbf{V} \), \( \Sigma \), and \( \mathbf{P} \) are the unitary matrices and \( \mathbf{A} \) a diagonal matrix. The two first correspond to a pre-whitening of the data and are derived from the eigenvalue-decomposition (EVD) of the spectral matrix (the correlation matrix in a no-delay mixture).

As the third matrix vanishes when the spectral matrix is estimated, the use of the HOS is a mean to solve the question. It is a classical result that the eigen-vector associated to the largest eigen-values of the spectral matrix span the signal subspace and correspond to a maximum for the likelihood function in the Gaussian case.

Now, it is interesting to explain how the Log-likelihood is obtained with the restricted model taking place in the signal subspace: \( \mathbf{x}(p) = \sum \mathbf{h}(p) \). Then \( \mathbf{P} \) corresponds to a rotation in the signal subspace and the vectors \( \mathbf{h}(p) \) are the projections of, respectively, the observation \( \mathbf{h}(n) \) and the noise \( \mathbf{h}(n) \). In the sequel, they will be denoted \( \mathbf{x} \) and \( \mathbf{h} \). A scale operation with the inverses of the \( p \)-largest eigenvalues leads to consider the new fictitious observations as white vectors with unitary power and uncorrelated components.

\[
\text{ML}_4(\mathbf{x}(1), \ldots, \mathbf{x}(N); \Pi) = \frac{1}{4!} \sum_{i, \alpha, \beta, \gamma, \delta} \left( \mathbf{P} \mathbf{B} \mathbf{P} \right) \left( \mathbf{x}(1), \ldots, \mathbf{x}(N) \right)^{2}
\]

This result is important because of its similarity with another obtained independently by P. Comon His theory has reinforced the weight of these 'nice' ML-estimators obtained by Gaeta, by mixing the above-mentioned properties and a rigorous theoretical support. Here, the concept of contrast function was introduced [3].

Using this set of tools, the following sections are dedicated to one way to perform the blind source separation where the aim is to maximize the function

\[
\sum_{1 \leq i < p} (\kappa_i)_{i,i,i,i}^2 \text{ (equivalent to } \text{ML}_4) \text{ with respect to the set of unitary matrices.}
\]

2. The algorithm without any constraint

In this part, we explain an endevour to practise adaptively the source separation. The main purpose is to use the previous criterion explained above. Some authors [6,7] have presented adaptive algorithms based on objective functions which are the sum of the absolute value of the source cumulant. The main drawback of this procedure is that it is required a priori knowledge on the sources by imposing the same sign to the cumulants. Using the Comon-Lacourme criterion, this constraint vanishes keeping a whole freedom.

Taking account that the remaining matrix belonging to \( \mathbb{R}^{m \times m} \) is orthogonal, it can be expressed as a product of \( r = \frac{n(n-1)}{2} \) Givens matrices:

\[
\Pi = \prod_{l<s} R_l(\theta_l)
\]

The determination of \( \Pi \) with the unitary constraint corresponds to the knowledge of a set of angles with a smaller size. The objective function remains:

\[
J(Y) = \sum_{1 \leq i < p} (\kappa_i)_{i,i,i,i}^2 (Y_i)^2
\]

knowing that:

\[
\mathbb{E}[Y Y^+] = \mathbf{I}, \quad \{\kappa_i\}_{i,i,i,i} = \mathbb{E}[|Y_i|^4] - 3 \mathbb{E}^2[|Y_i|^2]
\]
and becomes:

\[ J(\mathbf{Y}) = \sum_{1 \leq i \leq p} E[Y_i^4 - 3]. \]

The gradient with respect to \((\theta_\alpha)_{1 \leq \alpha \leq r}\) gives:

\[ \frac{\partial J(\mathbf{Y})}{\partial \theta_\alpha} = 8 \sum_{1 \leq i \leq p} (E[Y_i^4] - 3) E[\frac{\partial Y_i}{\partial \theta_\alpha} Y_i^3]. \]

So, the relation \( \mathbf{Y} = \prod_{i=1}^{p} R^{-1}(\theta_i) \mathbf{X} \) implies:

\[ \frac{\partial \mathbf{Y}}{\partial \theta_\alpha} = \prod_{i=1}^{p} R_i^{-1}(\theta_i) \frac{\partial}{\partial \theta_\alpha} R_i^{-1}(\theta_\alpha) \prod_{i=1}^{p} R_i^{-1}(\theta_i) \mathbf{X}. \]

From this, several choices are possible to write a "stochastic" version of this problem, often without any theoretical support. The way chosen by O. Macchi permits to derive easily the stochastic version and perhaps, to perform the asymptotic study. For our case, among the different possibilities, one of them is considered here by evaluating the different expectations in (A) with moving averages

\[ E[X^\alpha Y^\beta](n) = (1-\delta)E[X^\alpha Y^\beta](n-1) + \delta x^\alpha(n) y^\beta(n) \]

Practically, as at the step \( n \), the observation \( x(n) \) and the latter estimate \( \theta(n-1) \) are known, we decide to use the source signal being equal to:

\[ y(n|n-1) = \prod_{i=1}^{p} \theta(n-1) x(n). \]

As a summary, the algorithm has been implemented with the general way:

\[ \theta(n) = \theta(n-1) + \mu F(\theta(n-1), X_n) \]

where \( F = -\nabla J \) and \( X_n \) is the state vector equal, here, to the new observation \( x(n) \). It means:

\[
\begin{align*}
F(\theta(n-1), x(n)) &= 8 \sum_{1 \leq i \leq p} (E[Y_i^4](n-1) - 3) E[\frac{\partial Y_i}{\partial \theta_\alpha} Y_i^3](n) \\
E[Y_i^4](n) &= (1-\delta)E[Y_i^4](n-1) + \delta y(nln-1) Y_i^3(n) \\
E[\frac{\partial Y_i}{\partial \theta_\alpha} Y_i^3](n) &= (1-\gamma) E[\frac{\partial Y_i}{\partial \theta_\alpha} Y_i^3](n-1) + \gamma y(nln-1) \frac{\partial y(nln-1)}{\partial \theta_\alpha} Y_i^3(n-1)
\end{align*}
\]

The study with the technique of the ODE seems too difficult since the ODE appears as a trigonometric polynomial.

Moreover, this part has been introduced in a whole algorithm including the pre-whitening stage. This latter uses a procedure implemented by Ortigueira-Lagunas [8]. It is based on the updating of the EVD of the correlation matrix. So, the step using at the second-order using the correlation matrix has the following way:

\[
\begin{align*}
R_n &= (1-\beta) R_{n-1} + \beta \chi(n) \chi(n)^+ \\
EVD of R_n: & R_n = V_n \Lambda_n V_n^+ \\
& \Rightarrow R_n = V_{n-1} \Lambda_n V_{n-1}^+ \\
& \Rightarrow M_n = (1-\beta) \Lambda_{n-1} + \beta \Lambda_{n-1}^+ \chi(n) \chi(n)^+ V_{n-1} \\
EVD of M_n: & M_n = U_n D_n U_n^+ \\
& \Rightarrow \Lambda_n = (V_{n-1} U_n) D_n (V_{n-1} U_n)^+ \\
& \Rightarrow V_n = V_{n-1} U_n \text{ and } \Lambda_n = D_n
\end{align*}
\]

We carried out some simulations yielding to promising results. That is why it can be attempted to include this algorithm in an operational system.

3. Results in simulations

The tests that were carried out, stress the question of the simultaneous regulation of the different parameters. Generally, the evolution can be based on local dynamics or can be chosen to obtain first a fast convergence and then afterwards to fluctuate slowly.

On Figure 1, we show results obtained for a mixture of a Laplacian source \((\kappa_4=3)\) and a B-PSK signal \((\kappa_4=1.5)\). One can see the mixture as a product of a first rotation \((25^\circ)\) with a diagonal matrix and with another rotation which can not be observed at the second order. Curves (A) and (B) on the second graph correspond to the sets of parameters (A) and (B) (in left top of the figure, \(\alpha=\beta=\gamma\)). The parameter \(\alpha\) must be greater than \(\mu\) to obtain a convergence of the corresponding estimated term faster than one of the searched angle \((37^\circ)\). Of course lower are the gain \((\alpha, \mu)\) (B) \((<A)\), better and smoother is the result. (Compare curve (B) with (A)). We have noted that the parameter \(\beta\) leading to the uncorrelation of signals has not a great influence except at the beginning of the tracking. We can note the small dispersion obtained with the algorithm at the second order. Here, the curve fluctuates rapidly with very small dynamics. The histograms of the residual errors plotted with the 20,000 last estimated values look like to a Gaussian shape only for the first one which illustrates the behaviour of the algorithm at the second order.

As different but equivalent attractors exist, we have forced the estimates in the interval \([0, 90^\circ]\). Here, the problem of the regulation of the whole parameter set is emphasized. As no theoretical result exist to guide the choice, we move forward by trial.

CONCLUSION and FORECASTS

After having spoken about generalities, a method has been explained and tested to perform blind source separation. It is founded on the use of adaptive algorithm
which maximizes the criterion found independently by P. Comon and J.L. Lacoune. Simulations have been performed with satisfying results. Now, we can still improve this algorithm by testing another implementations and by carrying out some experiments with real signals. Moreover, its behaviour must be checked in a noisy environment to try afterwards an integration in a whole system of communications as explained now. Figure 2 illustrates the different parts of the process. The fundamental basis of this interesting issue is explained at greater length in another paper [9]. To roughly summarize, the problem appears when several users ask for the use of the channel at the same time. Then, it is necessary to separate their own contributions to adapt another treatment of the procedure in the best way.

References


[7] Moreau E., Macchi O., Séparation auto-adaptative de sources sans blanchiment préalable, Proc. 14ème GRETSI, Juan-les-pins (Fr), Sept 1993, pp 325-328


[10] Pham D.T., Garat P., Jutten C., Separation of a mixture of independent sources through a maximum likelihood approach, Proc. EUSIPCO'92, Brussels (Bel.)

Figure 1. Curve [I] is for the estimation of the angle 25° at the second order, curve [II] is for the second angle 37° obtained with the criterion. [III] and [IV] are the corresponding histograms of the errors.

Figure 2. The whole system for the application in communications
NONLINEARITY ON THE CONTROL ERROR
FOR RECURSIVE ADAPTIVE FILTERS

S. Cherif (1), Ch. Vignat (2), M. Jaidane (1), S. Marcos (2)

(1) Laboratoire des Systèmes de Télécommunications, ENIT/ESPTT, Campus Universitaire, BP 37 le Belvédère, Tunis, TUNISIA. Tel/Fax : (1) 762499/762819, E-Mail : eleclab@espttn.esptt.tn
(2) Laboratoire des Signaux et Systèmes, ESE/CNRS, Plateau de Moulon, 91192 Gif sur Yvette, FRANCE. Tel/Fax : (1) 69418040/69413060, E-Mail : vignat@univ-mlv.fr

Abstract. This paper proves that the introduction of a nonlinearity into the structure of a recursive filter has stabilizing and convergence properties compared to the linear case. The effect of this nonlinearity on the performances of fixed and adaptive recursive filters is presented. The stability domain is shown to expand, for a stochastic input, as a function of a so-called saturation degree of the nonlinearity. The modification of the error surface as a function of the nonlinearity is also investigated. It thus appears that the inversion of a non-minimum phase filter is made possible by the introduction of a nonlinearity. The nonlinearity also modifies the gradient-based optimization methods like the LMS and FMRLMS algorithms. Simulation results are given to show a better convergence compared to the linear structure case.

1 Introduction

There has been a growing interest in adaptive recursive filtering for applications such as teleconference, hands-free and radiomobile telephones, where the acoustic channel is characterized by a very long and time varying impulse response [1]. Unfortunately the theory of adaptive recursive filters is still incomplete (see for example [2]). For fixed (not adaptive) recursive filters the non quadratic and possibly multimodal nature of the mean square error may lead to suboptimal performances. Moreover, the filter parameters must be situated within a stability domain to guarantee BIBO stability. The characterization of this stability domain may be complicated.

In the adaptive context, due to the infinite memory of the filter, it is not possible to rigorously implement gradient-based optimization methods. The extended LMS as well as the SHARF algorithms are based on a rough estimation of the error and of its derivatives w.r.t the parameters [3][4]. Moreover the adaptive recursive filter with its updating algorithm is a nonlinear system which may have a complicated behaviour, possibly chaotic, concerning its stability [5].

Some systems using adaptive filters may include a quantizer-type nonlinearity into the structure. this is the case of systems processing digital signals as decision feedback equalizers which implement a decision device in the recursive part of the equalizer or as the ADPCM encoder which implements a predictor in conjunction with a quantizer. It is also the case when the updating algorithm is implemented with finite-length words.

The influence of this kind of nonlinearity on the adaptation of transversal filters has been analysed in [6]. It has been shown that a larger saturation degree of the nonlinearity leads to a slower convergence speed of the algorithm.

In this paper, we introduce a nonlinearity in the structure to improve the performances of the adaptive recursive filter. The positive effects of the introduction of a quantizer-type nonlinearity introduction is analysed in the context of a stochastic input. Section 2 is devoted to the presentation of the nonlinear recursive filter. Section 3 deals with the modification of the stability domain compared to the linear case. Section 4 presents the modification of the error surface due to the nonlinearity. Section 5 analyses some convergence properties of gradient based optimization methods like the LMS and FMRLMS algorithms [7-9] for the nonlinear recursive filter.

2 The nonlinear recursive filter

We introduce a nonlinearity $Q(\cdot)$ in the recursive structure (figure 1), such that:

$$e_n = y_n - H^T Q(E_n)$$

(1)

where $H$ is the filter weight vector of dimension $N$, $E_n = (e_{n-1}, \ldots, e_{n-N})^T$ and $Q(E_n) = q(e_{n-1})$.

For example, the structure depicted in figure 1 can be viewed as a nonlinear prediction which is a part of the ADPCM encoder. In this case, $y_n$ is the signal to be predicted and $e_n$ is the prediction error.

For our analysis we will consider the following nonlinearity (error function referred to as the erf function)

$$q(x) = A \int_0^x e^{\frac{u^2}{2}} du$$

(2)
where \( \sigma^2 \) is referred to as the saturation degree and \( A \) is used to fix the saturation level. The behaviour of this nonlinearity \( q(.) \) can vary between that of a linear device \((q(x) = x)\), when \( \sigma^2 \to \infty \) and \( A = 1 \), and that of a hard limiter \((q(x) = \text{Bsign}(x))\), when \( \sigma^2 \to 0 \) and \( A = \sqrt{2B/\sigma^2} \), by changing \( \sigma^2 \). The significance of \( B \) will be seen later. In our following analysis, \( q(.) \) will also be used as a continuous approximation of a hard limiter appearing in a quantizer.

### 3 Characterization of the stability domain of the fixed nonlinear recursive filter

In this paper, we consider a first-order filter

\[
\epsilon_n = y_n - h q(\epsilon_{n-1})
\]

with input \( y_n \) generated by:

\[
y_n = \epsilon_n + b \epsilon_{n-1}
\]

where \( \epsilon_n \) is a zero-mean Gaussian white noise with variance \( \sigma^2 \). For simplicity, we also assume that \( \epsilon_n \) is Gaussian.

In order to analyse the stability of the system defined by (3) and (4), we evaluate the evolution of \( E(\epsilon_n^2) \) as a function of time (transient rate). We obtain for the nonlinearity defined in (2):

\[
x_{n+1} = h^2 A^2 \text{asin}^{-1} \left( \frac{1}{1 + \frac{x_n}{\alpha}} \right) + \\
L + b^2 - \frac{2hbA}{\sqrt{1 + \frac{x_n}{\alpha}}} \right]
\]

where \( x_n = E(\epsilon_n^2) \) and \( \alpha = \frac{\sigma^2}{\sigma^2} \).

The stability domain of the recursive nonlinear filter corresponds, in fact, to the domain of convergence of the recurrence \( x_{n+1} = f(x_n) \), defined by (5). We consider at first the two extreme cases: the linear case \((q(x) = x)\) and the sign nonlinearity case \((q(x) = \text{Bsign}(x))\).

In the linear case \((\sigma^2 \to \infty \) and \( A = 1 \)\), the equation (5) becomes:

\[
x_{n+1} = h^2 x_n + (1 + b^2 - 2hb)
\]

\( x_n \) is then bounded if \( h \) lies in the stability domain \([-1, 1]\).

For the sign nonlinearity \((\sigma^2 \to 0 \) and \( A = \sqrt{2B/\sigma^2} \)), the equation (5) becomes:

\[
x_{n+1} = h^2 \frac{B^2}{\sigma^2} + \left[ 1 + b^2 - 2hb \frac{B}{\sigma} \sqrt{\frac{2}{\pi} \frac{1}{\sigma} x_n} \right]
\]

The stability domain of the sign nonlinear filter is defined by the set of \( h \), for a given \( B \), which permits the convergence of the recurrence (7). \( B = \sigma^2 \) corresponds to a saturation degree which will be matched to the power of the signal \( \epsilon_n \). We choose \( B = \sigma^2 \) in the following.

The classical stability, as defined in the linear case, is related to the stability of the fixed point solution of \( f(x) = x \) of the recurrence (5). The stability is then obtained for \( |\lambda| < 1 \), where \( \lambda = \frac{d(x_{n+1})}{d(x_n)} \) is the multiplier.

In the linear case, we find \( \lambda = h^2 \).

For the sign nonlinearity, we prove that:

\[
\lambda = hb \sqrt{\frac{2}{\pi} \frac{1}{x_n}}
\]

where \( x_n \) is the fixed point. In this case, the exact calculus of \( \lambda \) is possible. Figure 2 exhibits \( \lambda \) as a function of \( h \) for different values of \( b \). We see that \( |\lambda| < 1 \), \( \forall h, \lambda \). We conclude that the stability domain is all \( R \).

It is interesting to note that for \( b = 1 \), \( \lambda \) reaches a maximum \((\lambda_{max, 1})\), at the vicinity of \( h = \pm 1 \) (the boundary of the linear stability domain), smaller than 1 and which decreases when \( |h| \) increases. For the other values of \( b \), \( \lambda \) reaches a maximum \( \lambda_{max, b} < \lambda_{max, 1} \). It is thus possible to consider the cases \( |b| > 1 \) with a nonlinear recursive structure. Figure 3 presents the evolution of \( \lambda_{max, b} \) (relative to equation (8)) versus \( b \). The values of \( \lambda_{max, b} \) are smaller than 1, for all the values of \( b \). For communications applications like equalization, it is thus possible to inverse a non-minimum phase channel.

When we consider an \( erf \) nonlinearity, an exact resolution of the problem is difficult. We try to observe the effect of this nonlinearity on the stability domain by simulations. Figure 4 exhibits \( E(\epsilon_n^2) \) as a function of \( h \), for \( b = 0.5 \) and for the \( erf \) nonlinearity \((\sigma^2 = 0.01, 0.1 \) and \( 1 \)\) compared to the linear case. Unlike the linear case, the curves related to the \( erf \) nonlinearity do not tend to infinity at \( h = \pm 1 \). Moreover, figure 4 proves that the stability domain expands when \( \sigma^2 \) decreases, i.e. when the nonlinearity tends to the sign nonlinearity. Then the introduction of the nonlinearity enlarges the stability domain which thus becomes, for a hard limiter, all \( R \).

A previous study [7] has proved that for a constant input, the error \( \epsilon_n \) (eq 3) has a complex behaviour, possibly chaotic, related to the bifurcation parameter \( h \).

In the context of a stochastic input, it is interesting to analyse the nonlinear behaviour through the deterministic quantity \( x_n \). It is then possible to define a possible chaotic behaviour of a stochastic nonlinear system.
4 Optimality of the fixed non-
linear recursive filter

For the sign nonlinearity, we can choose the optimal weight \( h_{\text{opt}} \) that leads to the minimum Mean Square Error MSE \( (\sigma_{\text{mse}}^2 \text{ or } x_{\text{mse}}) \). We establish that:

\[
x_{\text{min}} = \frac{1 + b^2 + \sqrt{b^4 + 2(1 - \frac{1}{x})b + 1}}{2}
\]

which is greater than 1.

In the linear context, the optimal value of \( h \) is \( b \) and the minimum mean squared error is \( \sigma_{\text{mse}}^2 = \sigma_\eta^2 \). It is clear that the introduction of the sign nonlinearity degrades the performances in the stochastic context, considered here. Note that other nonlinearity functions can yield a minimum MSE close to the minimum MSE of the linear case.

The degradation is due to the particular input which is chosen here. Indeed, it has been found in [8], that for a constant input, the introduction of a nonlinearity may decrease the minimum MSE. Figure 5 shows that, for the sign nonlinearity and for a constant input (equal to 1), \( \sigma_{\text{mse}}^2 \) is lower than in the linear case. The introduction of the nonlinearity may also improve the minimum mean square if the signal \( y_n \) is generated by a nonlinear filter. This is interesting if we consider an adaptive equalizer in the presence of a nonlinear channel.

5 Effects of the nonlinearity in the updating context

Theoretically, the strict implementation of the LMS algorithm for the optimization of a recursive filter should require an infinite length memory. The Extended LMS as well as the t SHARF algorithm [3][4] are approximated versions of the actual LMS algorithm [9]. Indeed, the LMS algorithm is defined by:

\[
h(n + 1) = h(n) - \mu e_n \left. \frac{d e_n}{d h} \right|_{h=h(n)}
\]

where:

\[
\left. \frac{d e_n}{d h} = -q(e_{n-1}) \right|_{h=h(n)} - h(n) \left. \frac{d p(e_{n-1})}{d h} \right|_{h=h(n)}
\]

To implement this algorithm, we have to solve two difficulties. The first one is due to the recursive form of (11) and the second one is due to the fact that both \( e_n \) and \( \frac{d p}{d h} \) should be computed with the same value \( h(n) \) of \( h \). The latter then involves a memory growing with \( n \). This is not feasible in practice. The FMRMLS algorithms [9] which implement a finite-memory take these two difficulties into account.

The extended LMS uses two approximations: (i) it considers that the filter has a one-length memory, (ii) the second term in (11) is neglected. In the following, we consider that:

\[
q(x) = x + \eta
\]

where \( \eta \) is the quantizer noise of variance \( \sigma_\eta^2 \), is an approximation of a hard limiter appearing in a quantizer. We evaluate the asymptotic ratio of the two components of equation (11) for the linear case, which corresponds to \( q(x) = x \), and for the nonlinear case with (12). Figure 6 gives this ratio for different values of \( \sigma_\eta^2 (\sigma_\eta^2 = 1, 0.1, 0.01) \), compared to the same ratio for the linear case. We see that this ratio decreases towards zero as \( \sigma_\eta^2 \) increases (hard limiter). This shows that the nonlinearity makes the second component of (11) negligible compared to the first component.

We conclude that the approximation used by the Extended LMS is more valid for a nonlinear system than for a linear system. We thus expect that the introduction of a nonlinearity in the structure of a adaptive recursive filter gives a more accurate convergence than in the case of a linear structure.

6 Conclusion

The effect of a quantizer type nonlinearity on the performances of fixed and adaptive recursive filters is investigated through comparison with the linear structure case. We show (for a first-order filter) that the nonlinearity enlarges the stability domain compared to the linear case as the saturation degree of the nonlinearity increases. The nonlinear behaviour of the fixed filter is analysed through the (deterministic) mean square error evolution. Moreover, we show that the introduction of a nonlinearity on the control error confers a finite memory property to the output of the filter and its derivative. It is thus expected that the FMRMLS algorithms [9] used for the optimization of a nonlinear recursive filter require a shorter memory than in the case of a linear filter.

Acknowledgement : The authors wish to thank Dr. Saphia Belghit for fruitful discussions.

References

Figure 1: The nonlinear recursive filter structure.

Figure 2: \( \lambda(h) \) for \( b=0.5, 1 \) and 2, for the sign nonlinearity compared to the linear case.

Figure 3: \( \lambda_{max, b}(b) \) for the sign nonlinearity.

Figure 4: \( \langle E(\epsilon_n^2) \rangle(h) \) for \( b=0.5, \) for the erf nonlinearity (\( \sigma^2=0.01, 0.1, 1 \)) compared to the linear case.

Figure 5: \( \langle E(\epsilon_n^2) \rangle(h) \) for the sign nonlinearity compared to the linear case, with a constant input.

Figure 6: Extended LMS implementation for the linear and nonlinear structure (\( \sigma_n^2=0.01, 0.1, 1 \))
Variable Rate Speech Coding

Allen GERSHO and Erdal PAKSOY

Center for Information Processing Research, Dept. of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106 U.S.A.
Tel/Fax +1 (805) 893-2037 / 893-3262, E-Mail: gersho@ece.ucsb.edu

Abstract. Variable rate speech coding has become a very active and important topic in the last few years. Recent activities and advances in variable-rate coding have been motivated in large part by CDMA wireless networks. Effective variable-rate coders based on CELP and MBE have been developed with good communications quality and average rates ranging from 1.5 to 4 kbps. These coders usually include both voice activity detection and phonetic classification to achieve an effective tradeoff between quality and average bit rate. Both source control and external control of rate is considered in this review of recent work.

1. Introduction

Most speech coders are designed to generate a constant-rate bit stream for digital transmission or storage. While a constant rate is well-suited for many digital communication systems, speech is by nature intermittent and has a short-term statistical character that varies greatly with time. Speaking intuitively, we can think of speech as having a short-term entropy that varies widely with time. Thus, encoding speech at a constant bit rate implies that individual segments may be inefficiently coded, either by giving more bits than necessary for a particular quality objective or by degrading quality due to an insufficient bit allocation. For this reason, it is more natural and efficient to code speech with a variable bit rate. Certain multiple-channel communication systems, are able to take advantage of the statistical variation of the bit rate from multiple variable rate signal sources. Most notable are wireless cellular networks based on code-division multiple access (CDMA). Recent work in variable-rate speech coding has been motivated to a large degree by the emerging importance of CDMA.

For the purpose of this paper, we take the definition of a variable bit rate (VBR) speech coder to mean a coder that can select a different rate for each coding unit, which is of short duration, typically 10 to 40 ms. This distinction excludes coders that essentially operate at fixed rates but whose rate can be manually or automatically reset by external control at relatively infrequent intervals, say less than once per minute on the average. Coders of the latter type, sometimes called multi-rate coders, are not able to exploit the varying statistical character of speech and therefore are not classified here as VBR coders.

For a given level of quality, speech can be compressed at a lower average rate by a VBR coder than by a fixed rate coder. This is done by taking advantage of the statistical variations in the character of the speech signal. One side of a typical telephone conversation consists of bursts of speech, called talk spurts, separated by pauses or gaps during which the speaker is silent. Since the speech information is present only intermittently, there is no need to transmit speech data at a fixed high bit rate throughout the duration of the telephone call. Thus, an important component of a VBR coding system is a voice activity detection (VAD) algorithm, designed to detect the beginning and end of these talk spurts. Some VBR coders exploit only the on-off character of the speech and employ a constant rate during active speech segments. However, more can be done to improve the effectiveness of VBR speech coding. It is known that active speech segments have highly time-varying statistics. Properties such as energy levels, spectral shape, and the degree of voicing exhibit large amounts of time-variation. For a given reproduction quality, some speech segments often require fewer bits than others so that the average bit rate of a VBR coder operating on active speech can be substantially reduced over that of a fixed rate coder while maintaining the same or similar quality.

It is natural to vary the encoding rate of the input signal based on the local behavior of the speech source. This is called source-controlled VBR speech coding. However, the dynamic assignment of bit rate can also be made to depend on the instantaneous status and needs of the overall network. This is called network-controlled VBR coding and will also be discussed in this paper.

VBR coding has been studied for many years in the context of digital circuit multiplication equipment
(DCME). In DCME, source-controlled VBR coding is based on VAD to reduce the total bit rate needed to collectively represent a large set of voice channels at any one time. Such systems also use network controlled VBR coding by reducing the bit rate of all voice channels when the total traffic at any instant exceeds the capacity of the channel. Another application has been in packetized voice transmission where VAD is used so that voice packets do not need to be sent during silent intervals. Both these applications have generally employed traditional waveform coders, most commonly ADPCM, and they did not exploit the statistical variations in active speech to reduce the average bit rate.

The interest in VBR speech coding has been renewed recently primarily due to the development of new wireless multiple access technologies such as Packet Reservation Multiple Access (PRMA) [1], Extended TDMA (E-TDMA) [2], and especially Code Division Multiple Access (CDMA), which, unlike established methods like TDMA and FDMA, are designed to support VBR transmission. The emergence of CDMA in particular, has lead to standardization efforts in the U.S. and Europe. The North American Telephone Industry Association (TIA) has adopted a new standard in 1993, known as IS-95 for digital cellular telephony based on CDMA and VBR coding. The European Community's current program for Research on Advanced Communication in Europe, called RACE-II, currently has a Code Division Testbed (CODIT) project dedicated to studying the use of CDMA and VBR coding for the Universal Mobile Telephone System (UMTS).

The paper is organized as follows. In Section 2 we address the important issue of VAD which alone can help us achieve significant savings in average bit rate. In Section 3, we discuss various VBR coding techniques for the speech signal itself. Section 4 explains the new wireless multiple access schemes and other new technologies that provide motivation for using VBR coding of speech. In Section 5 we describe the issue of network control of coding rate.

2. Voice Activity Detection

Various studies made on voice activity patterns [4], [5] have shown that one side of a two-way telephone conversation typically consists of intermittent talk spurts separated by silences which occupy on average 60 to 65% of the duration of the signal. These silences correspond to relatively long time intervals during which the speaker is listening, or shorter intervals when the speaker is pausing between sentences or words. Very short pauses within words also occur.

Effectively detecting the presence or absence of speech is a very important part of a VBR speech coding system. The task of a VAD algorithm is to identify when talk spurts occur. Portions of the input signal classified by the VAD algorithm as containing speech are called active speech, whereas segments containing solely background noise or silence can be called nonspeech. VAD provides the basis for the simplest form of VBR speech coding, simply turning a fixed rate speech coder on and off depending on the presence of speech. It may also be the basis for somewhat more sophisticated VBR coders which are able to represent and reproduce the background noise during nonspeech. In such coders the background noise can either be coded at a very low bit rate, or it can be directly reproduced without transmitting any information to the receiver, which can generate an approximation to the original noise, called comfort noise using some parameters obtained from prior knowledge of the acoustical environment.

The percentage of the time that the VAD algorithm declares the presence of active speech is called the voice activity factor (VAF). This quantity also determines the increase in capacity (or reduction in bit rate) that the VAD algorithm can provide. The sensitivity and the reliability of the VAD algorithm is crucial to the performance of a VBR speech coder. If nonlinear is detected as speech, the capacity is reduced. If active speech is detected as nonspeech, the quality of the reconstructed signal is adversely affected due to clipping, the process where active speech segments are deleted when misclassified as nonspeech.

The task of VAD is rendered even more difficult by the presence of different types of background noise, such as engine noise in a moving vehicle or conversational (babble) noise in a restaurant. In such circumstances certain speech units can more readily be confused with background noise. Hence the VAD algorithm must be designed more conservatively so as to minimize the degrading effect of clipping.

The Global System for Mobile Communications (GSM) subcommittee of the European Telecommunications Standards Institute (ETSI) has adopted a standard for digital cellular networks which includes a recommendation for a VAD algorithm designed to identify speech in the presence of background vehicle noise [6]. In this method, the spectral shape of the noise is dynamically estimated from nonspeech segments by updating an FIR noise suppression filter, whose coefficients are the LPC coefficients of the noise spectral envelope. Each frame of input speech is passed through this filter, and the power at the output of the filter is compared to an adaptive threshold to detect the presence of speech. A fixed hangover period is also included in the algorithm. This means that the declaration of a transition from speech to silence is delayed, and as such clipped during or at the end of a talk spurt.

New techniques for improving the performance of the GSM VAD algorithm at very low SNR values were proposed in [7]. These features are: energy level com-
parisons in individual frequency subbands, determination of the spectral flatness measure and the use of an adaptive hangover period. This work also describes an algorithm for the more challenging task of detecting babble noise and a method for combining the VAD methods for both babble and vehicle noise.

3. VBR Coding Of Active Speech

While a substantial capacity increase can be obtained with VAD alone, the highly time-varying perceptual entropy of active speech can be further exploited to achieve additional savings in average bit rate. A VBR speech coding algorithm often divides the speech signal into fixed length frames and for each frame, selects an appropriate bit allocation and coding technique based on a rate decision mechanism. For each bit rate option, the coding method can be an entirely distinct algorithm or simply a different mode of a given type of coder. A special case of particular interest is embedded coding, where a single coding algorithm generates a single fixed rate data stream from which one of several reduced rate data signals can be extracted by a simple bit-dropping procedure. Embedded coding can be used for multi-rate coding as well as VBR coding.

The rate determination algorithm is a critical part of a VBR coder. This mechanism can be of one of two main types. In open-loop classification, the rate and coding method are selected based on an analysis of the properties of the current segment or frame of the speech signal, prior to encoding this segment. In closed-loop rate selection, the speech frame is encoded using a number of different methods with different rates, and the resulting speech quality and rate are evaluated in order to select the best available method based on examining the rate-quality trade-off of each option. In this section, we will briefly review some speech coding algorithms for source-controlled speech coding.

An example of a closed-loop rate selection algorithm is finite state CELP (FS-CELP) introduced by Vaseghi [8], [9]. Here, among the coding configurations available for coding a frame, the one which achieves a predetermined weighted SNR level with the smallest bit rate is selected. This is an attractive idea, however the lack of a suitable objective distortion measure of coded speech quality renders it rather inefficient. In particular, frames corresponding to unvoiced speech are often coded at a high bit rate, whereas from a perceptual standpoint, they require fewer bits than voiced speech frames. Another closed-loop approach was presented by Eriksson and Sjoberg [10], here the variable-rate coder is created by first designing a large number of coders and by then going through a "natural selection" process where the candidates with unsatisfactory performance are gradually eliminated and the best coding configurations are retained. The coding itself is performed by optimizing a joint cost function which takes into account the rate and the weighted SNR.

QCELP [3], [11], is the speech coding algorithm developed by Qualcomm and proposed for use with the North American IS-95 standard for digital cellular systems by the Telephone Industry Association (TIA). It is based on a simple open-loop rate decision technique which makes use of the wide dynamic range of the speech signal. In QCELP, the energy of the background noise is monitored and used to adapt a set of three time-varying thresholds which float above the noise level. The energy of the speech signal is compared to these three thresholds. If it exceeds all three, speech is coded at the highest available rate of 8 kb/s, if it is lower than all three, then the lowest rate of 1 kb/s is selected. If the frame energy falls below the thresholds, one of the intermediate rates of 2 kb/s or 4 kb/s is used. The four rates available in QCELP essentially correspond to scaled versions of a basic CELP coder. Despite its good performance, QCELP does not make maximally efficient use of all the available bit rate options, since it mostly operates at 1 or 8 kb/s and rarely uses the intermediate rates.

An important class of open-loop rate decision algorithms is based on phonetic classification where the coder can adapt to the needs of each coding frame in a perceptually meaningful manner [12], [13], [14], [15], [16], [17]. It should be noted that the type of phonetic classification depends on the type of coding algorithm to be used. For instance, in [15], the coding is CELP-based and active speech is classified into categories such as voiced, unvoiced, and onset. The onset category corresponds to a transition from unvoiced to voiced speech, which is a perceptually important feature in low bit-rate CELP coding of speech. On the other hand, in the work by Das and Gersho, the coding is based on an enhanced version of multiband excitation (MBE) coding and the categories are fully unvoiced, fully voiced, and mixed voiced [17]. Recently, McClellan and Gibson have proposed the use of spectral entropy for open-loop rate determination in VBR coders [18].

In [19] and [20], Cellario et al. describe a hybrid algorithm which combines open-loop and closed-loop rate decision mechanisms. Each frame of the input signal is first classified as nonspeech, voiced speech or unvoiced speech. Then, for each frame within a broad class, the coder selects a suitable configuration based on weighted SNR calculations. The coding options within each broad class follow an embedded structure.

4. Applications

The different types of variable-rate coders described above can be used in various applications, ranging from voice storage to cellular telephony. These applications will be presented in the following.
4.1 Multiple access for cellular telephony

In traditional TDMA and FDMA systems, a fixed time or frequency slot is assigned to each user in a two-way conversation, regardless of the nature of the signal transmitted. In such a system, for the entire duration of the call, a fixed amount of channel resources are held by each side of the conversation, even during periods when they are silent. Such systems are not designed to handle VBR transmission, and must be modified to do so.

There have been efforts to modify these more traditional multiple access schemes so that they can handle VBR and achieve increased capacity by incorporating the concept of statistical multiplexing as in digital speech interpolation (DSI). DSI exploits the on-off nature of speech by allocating channel resources to a user only for the duration of a talk spurt. DCME products are based on this concept. A proposal for applying DSI to TDMA systems, called extended TDMA (E-TDMA) was introduced by Kay in [2]. A mobile-base transmission protocol called packet reservation multiple access (PRMA) is described by Goodman in [1]. In both E-TDMA and PRMA, the capacity of the system is increased significantly by transmitting data only when speech is present. On the other hand, the rate variation is limited to turning a fixed rate coder on and off and a significant amount of overhead is required in the form of either dedicated control channels or packet headers to specify the source, destination, and nature of the transmitted data channels.

These problems are avoided in code division multiple access (CDMA) [21], which is based on direct-sequence spread spectrum. Here, the signal coming from each user is spread over the entire available frequency band by modulating it on a pseudo-random carrier signal. Reducing the bit rate for each user reduces the interference that is seen by the others, and improves the signal-to-interference ratio. This leads to higher quality received speech. Such a method offers a simple and natural way of benefiting from VBR coding since the savings in average bit rate are directly translated into an improvement in quality.

4.2 Other applications

The application of variable-rate speech coding is not limited to cellular telephony. In fact, in cases where transmission delay is not a serious concern, the use of variable-rate coding is considerably less complicated. These uses include telephone answering machines, voice mail, and multi-media applications where VBR speech coding ensures efficient utilization of available storage resources. Already mentioned are DCME, used in submarine and satellite communication systems, and packetized voice systems which are currently used in private digital networks.

5. Network Control of Coding Rate

In Sections 2 and 3, we saw how the speech coder rate can be made to vary as a function of the properties of the input signal. The instantaneous rate can also made to depend upon the traffic level in a cellular network. The speech coders of mobile units can be instructed by a central control unit to adjust their rate according to the needs of the network, irrespective of the current state of the signals to be coded.

A simple example of network control of rate is present in the North American IS-54 CDMA system, where a "dim-and-burst" or a "blank-and-burst" mode can be selected in order to allow for the transmission of signaling information. In dim-and-burst mode the coder rate is halved, in blank-and-burst mode, it is reduced to zero. The frequency with which these operations are performed needs to be limited since they may introduce audible clicks, clipping, and other degradations.

In [22] Kawashima et al. investigate a more sophisticated strategy for network control applicable in the case of a heavily loaded network. In an extreme case, the number of active mobiles in a given cell may be very high, for a certain amount of time. If they all transmit at the peak bit rate at the same time, the interference seen by each receiver may be excessive and could lead to serious or catastrophic degradations in speech quality. The paper shows that, under such circumstances, it may be beneficial to reduce the source coding bit rate of some or all coders, thus reducing the interference in the network and improving the average received speech quality.

Another interesting application of network-controlled VBR coding is to allow a dynamic tradeoff of rate between source and channel coding where the total rate is fixed but varying radio channel conditions may require a varying amount of error correction coding. Such a scheme, reported by Yuen et al. [23], is applicable to conventional TDMA based cellular networks where a constant overall bit-rate is required for the combined source/channel coding of a speech signal.

6. Conclusion

In this paper, we have examined the main issues related to variable bit rate coding of speech signals. The principal advantage of VBR coding is the increase in capacity that it offers for a given quality, when used in applications designed to support variable-rate transmission or storage. It appears that VBR speech coding is likely to be an important requirement for several upcoming standardization activities in many areas, including CDMA based cellular systems for mobile telephony and personal communication systems of the future.
7. Acknowledgment
This work was supported in part by Fujitsu Laboratories Ltd., the National Science Foundation, the UC Micro program, Rockwell International Corporation, Hughes Aircraft Company, Echo Speech Corporation, Signal Technology, Inc.

References


A Signal Subspace Approach for Noise Reduction of Speech Signals

Søren Holdt JENSEN, Per Christian HANSEN, Steffen Duus HANSEN and John Aasted SØRENSEN

ESAT—Department of Electrical Engineering, Katholieke Universiteit Leuven
Kardinaal Mercierlaan 94, B-3001 Heverlee, Belgium
E-mail: Soren.Jensen@esat.kuleuven.ac.be

†UNiC, Building 305, Technical University of Denmark, DK-2800 Lyngby, Denmark
E-mail: Per.Christian.Hansen@uni-c.dk

‖Electronics Institute, Building 349, Technical University of Denmark, DK-2800 Lyngby, Denmark
E-mail: sdh@ei.dtu.dk and jaas@ei.dtu.dk

Abstract. We present a novel algorithm for reduction of colored broad-band noise in speech signals. The algorithm is based on the signal subspace approach and is formulated by means of the quotient-singular-value-decomposition. With this formulation, a pre-whitening operation becomes an integral part of the algorithm. We demonstrate that this is essential in connection with updating issues in real-time recursive applications. We also illustrate by experiments that we are able to achieve a satisfactory quality of the reconstructed signal.

1 Introduction

At a noisy site, e.g. in the cabin of a moving vehicle, speech communication is affected by the presence of acoustic noise. This effect is particularly serious when Linear Predictive Coding (LPC) is used for the digital representation of speech signals at low bit rates as in digital mobile communication. Low-frequency acoustic noise severely affects the estimated LPC spectrum in both the low- and high-frequency regions. Consequently, the intelligibility of digitized speech using LPC often falls below the minimum acceptable level.

Several approaches for noise reduction of speech signals have been proposed over the last two decades and perhaps the most popular approach is the one based on spectral subtraction [1, 2]. However, in a mobile environment algorithms based on this approach often suffers from the fact that either the noise reduction is not sufficient or the speech distortion is too large [3]. Recently, a new approach for noise reduction of speech signals based on subspace decomposition have been proposed [4, 5]. The key idea here is to consider the noisy speech signal as a vector in N-dimensional space, and to separate the speech signal and the noise into two mutually orthogonal components lying in different subspaces; the signal subspace and the noise subspace.

The noise reduction algorithm in [4] is based on the Singular Value Decomposition (SVD). In essence, the algorithm first arranges the data in a Toeplitz matrix, then computes a Least-Squares (LS) estimate of the signal-only data matrix, and finally restores the Toeplitz structure of the computed LS estimate. The LS estimate of the signal-only data matrix is computed by neglecting 'small' singular values, and the number of retained singular values is determined adaptively from the data in each speech segment. The algorithm in [5] is very similar, except that it uses the Karhunen-Loève transform instead of the SVD. There are two immediate drawbacks of these noise reduction algorithms. The first is that the algorithms deal only with white noise, i.e., in case of colored broad-band noise a pre-whitening is necessary. Pre-whitening is difficult to treat in a recursive algorithm which is desirable in real-time speech processing applications. Secondly, the LS estimate requires a good judicious concerning the number of retained singular values, and much computational effort is spent in computing this number.

In this paper, we present a novel noise reduction algorithm based on the signal subspace approach. Our algorithm is based on the Quotient Singular Value Decomposition (QSVD), and it deals with colored broad-band noise. Moreover, by using a Minimum-Variance (MV) estimate of the signal-only matrix (instead of the LS estimate), our algorithm is less sensitive to the choice of retained singular values [6].

2 Signal and Noise Model

We consider a noisy signal vector of N samples:

$$x = [x_0, x_1, \ldots, x_{N-1}]^T,$$

(1)

and we assume that the noise is additive and uncorrelated with the signal, i.e.,

$$x = \hat{x} + n,$$

(2)

where \(\hat{x}\) contains the signal component and \(n\) represent the noise. From \(x\) we can construct the following \(L \times M\) Hankel matrix \(H\), where \(M + L = N - 1\) and \(L \geq M\):

$$H = \begin{bmatrix}
x_0 & x_1 & \cdots & x_{M-1} \\
x_1 & x_2 & \cdots & x_M \\
\vdots & \vdots & \ddots & \vdots \\
x_{L-1} & x_L & \cdots & x_{N-1}
\end{bmatrix}.$$

(3)

We can always write \(H\) as

$$H = \tilde{H} + N,$$

(4)

1174
where $\hat{H}$ and $N$ represent, respectively, the Hankel matrices derived from $x$ and $n$ in (2). Moreover, we assume that $\hat{H}$ is rank deficient, rank($\hat{H}$) = $K < M$, and that $\hat{H}$ and $N$ have full rank, rank($\hat{H}$) = rank($N$) = $M$. This assumption is, e.g., satisfied when the samples $\hat{z}_i$ of $x$ consist of a sum of $K$ sinusoids and the samples $n_i$ of $n$ consist of broad-band noise. Such a model has often been attributed to speech signals, see e.g. [7]. We remark that one can also choose to work with Toeplitz matrices instead of Hankel matrices. There are no fundamental differences between these two approaches.

3 SVD and Signal Subspace Estimation

A detailed derivation of the algebraic and geometric conditions that allow us to derive the signal model from the SVD of the data matrix (5) is given in [8]. Using a multivariate version of the classical Pythagorean lemma for triangles it is shown that the signal matrix $\hat{H}$ cannot be recovered consistently when the signal matrix is perturbed by additive noise.

3.1 White Noise Case

Partition the SVD of $\hat{H}$ as follows:

\[
\hat{H} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix},
\]

(5)

where $U_1 \in \mathbb{R}^{K \times K}$, $V_1 \in \mathbb{R}^{M \times K}$, $\Sigma_1 = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_K)$ and $\Sigma_2 = \text{diag}(\sigma_{K+1}, \ldots, \sigma_M)$, and partition the SVD of $\hat{H}$ accordingly:

\[
\hat{H} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}.
\]

(6)

If the following three assumptions:

1. The signal is orthogonal to the noise in the sense:
   $\hat{H}^T N = 0$,

2. The noise is white, i.e., the noise matrix $N$ has orthogonal columns and every column of $N$ has norm $\sigma_{\text{noise}}$:
   $N^T N = \sigma_{\text{noise}}^2 I$,

3. The smallest singular value of $\Sigma_1$ is strictly larger than the largest singular value of $\Sigma_2$ in the SVD of $\hat{H}$, i.e.,
   $\sigma_K > \sigma_{K+1}$,

are satisfied, then the rank of $\hat{H}$, the row space of $\hat{H}$ (represented by $V_1$), and the null space of $\hat{H}$ (represented by $V_2$) can be estimated consistently, while the column space of $\hat{H}$ (represented by $U_1$) cannot be recovered from the SVD of $\hat{H}$—not even asymptotically.

In practice, Assumptions 1 and 2 are never satisfied exactly. A nice feature of the SVD, however, is its robustness with respect to mild violations of these assumptions.

Despite the fact that $U_1$ cannot be consistently estimated from the SVD of $\hat{H}$, it is possible to find an LS or MV estimate of $\hat{H}$ from the SVD of $\hat{H}$ [6]:

\[
\hat{H}_{LS} = U_1 (F_{LS} \Sigma_1) V_1^T,
\]

(7)

\[
\hat{H}_{MV} = U_1 (F_{MV} \Sigma_1) V_1^T,
\]

(8)

where $F_{LS}$ and $F_{MV}$ are $K \times K$ filter matrices defined by, respectively,

\[
F_{LS} = I_K,
\]

(9)

\[
F_{MV} = \text{diag} (1 - \frac{\sigma_{\text{noise}}^2}{\sigma_i^2}, \ldots, 1 - \frac{\sigma_{\text{noise}}^2}{\sigma_{K}^2}).
\]

(10)

The LS and MV estimates $\hat{H}_{LS}$ and $\hat{H}_{MV}$ of $\hat{H}$ spoil the Hankel structure. So, to obtain a signal vector $\hat{x}$ corresponding to the LS and MV estimates we need to make a Hankel matrix approximation to $\hat{H}_{LS}$ or $\hat{H}_{MV}$. A simple way to compute a Hankel matrix approximation is to arithmetically average every antidiagonal of these matrices, and put each average-value as a common element in the corresponding diagonal of a new Hankel matrix of the same dimension [9]:

\[
\tilde{\hat{H}} = \begin{bmatrix} \hat{x}_0 & \hat{x}_1 & \ldots & \hat{x}_{M-1} \\ \hat{x}_1 & \hat{x}_2 & \ldots & \hat{x}_M \\ \vdots & \vdots & \ddots & \vdots \\ \hat{x}_{L-1} & \hat{x}_L & \ldots & \hat{x}_{N-1} \end{bmatrix}.
\]

(11)

where

\[
\hat{x}_i = \frac{1}{\beta - \alpha + 1} \sum_{k=1}^{\beta} \hat{H}(i - k + 2, k),
\]

(12)

\[
\alpha = \max(1, i - L + 2),
\]

(13)

\[
\beta = \min(M, i + 1).
\]

(14)

3.2 Colored Noise Case

If the noise is not white, $N^T N \neq \sigma_{\text{noise}}^2 I$, then a pre-whitening matrix $R^{-1}$ can always be applied to $\hat{H}$ in (4). If the noise covariance matrix of the type $N^T N$ is known or can be estimated, then we can factor $N^T N = R^T R$, where $R$ is the Cholesky factor of $N^T N$. Moreover, if the noise matrix $N$ is directly available, then we can factor $N = Q R$, where $Q^T Q = I$ and $R$ is the same Cholesky factor as before. In the colored noise case, we thus consider the matrix

\[
X = \hat{H} R^{-1}.
\]

(15)

Substituting (4) into (15) yields

\[
X = \hat{H} R^{-1} + N R^{-1}.
\]

(16)

This transformation does not change the nature of the linear model while it diagonalizes the covariance matrix of the noise as shown by $(NR^{-1})^T(NR^{-1}) = Q^T Q = I$. That is, (16) is now in the same form as (4), and Assumptions 1–3 apply to the transformed signal matrix $\hat{H} R^{-1}$ and the transformed noise matrix $N R^{-1}$. We can therefore apply the technique from (5)–(14) to the transformed matrices. The only modification is that the LS or MV estimates of $\hat{H} R^{-1}$ should be de-whitened before restoring the Hankel structure.

The requirement that the noise covariance matrix $N^T N$ or an estimate should be known is not always trivial. In speech processing, we are in the "lucky" situation that $N$ is known for non-speech periods and that this can be used to estimate the noise covariance matrix for data where speech is present.

4 Algorithm

In this section we outline the noise reduction algorithm. For convenience, we first describe the algorithm using pre-whitening and then point out how the pre-whitening step can be incorporated by using a generalization of the SVD to two matrices, namely, the QSVD.
4.1 Formulation of the Algorithm

The following steps comprise our noise reduction algorithm:

1. Obtain a pre-whitening matrix $R^{-1}$ in non-speech sections as follows:
   - Form a $L \times M$ Hankel matrix $N$ from $n$.
   - Compute the QR decomposition of $N$:
     \[ N = QR. \]  
     (17)

2. Form a $L \times M$ Hankel matrix $H$, $L \geq M > K$ and $L + M = N - 1$ from $x$.

3. Perform a pre-whitening of $H$:
   \[ X = HR^{-1}. \]  
   (18)

4. Modify the matrix $X$ as follows:
   - Compute the SVD of $X$:
     \[ X = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}. \]  
     (19)

   - Truncate to rank $K$ and correct the retained singular values by applying a filter matrix $F$:
     \[ X_K = U_1 (F \Sigma_1) V_1^T. \]  
     (20)

5. Perform a de-whitening of $X_K$:
   \[ Z_K = X_K R. \]  
   (21)

5. Compute the elements $k_i$, $i = 0, 1, \ldots, N-1$ of $k$ from $Z_K$ by arithmetic averaging along its antidiagonals:
   \[ z_i = \frac{1}{\beta - \alpha + 1} \sum_{k=0}^{\beta} z_{k}(i - k + 2, k), \]  
   (22)

where $\alpha = \max(1, i - L + 2)$ and $\beta = \min(M, i + 1)$.

A challenging problem in the noise reduction algorithm is the determination of $K$, i.e., the effective rank of the data matrix. In a recursive algorithm it is convenient if a fixed effective rank can be used. We return to the actual choice of the rank $K$ in a practical algorithm in Section 5.5.

4.2 Implementation by QSDV

The explicit use of the matrix $R$ may result in loss of accuracy in the data. This can be avoided by working directly with $H$ and $R$, i.e., by using the QSDV of the pair of matrices $(H, R)$ which delivers the required factorization without forming quotients and products. Another issue, which may be even more important in practise, is that it is very complicated to update the matrix $X = HR^{-1}$ when $H$ and $R$ are updated, e.g., in a recursive application. Therefore, it is advantageous to use a decomposition of the matrix pair $(H, R)$ instead, which allows each matrix to be updated individually.

In the QSDV the matrix pair $(H, R)$ is written as follows:

\[ H = U \Delta \Theta^{-1}, \quad R = V M \Theta^{-1}, \]  
(23)

where $U \in \mathbb{R}^{L \times M}$ and $V \in \mathbb{R}^{M \times M}$ are matrices with orthonormal columns, $\Theta \in \mathbb{R}^{M \times M}$ is nonsingular, and $\Delta = \text{diag}(\delta_1, \delta_2, \ldots, \delta_M)$ and $M = \text{diag}(\mu_1, \mu_2, \ldots, \mu_M)$ are diagonal matrices.

Using (23) we may write

\[ HR^{-1} = U \Delta \Theta^{-1} (V M \Theta^{-1})^{-1} = U \Delta \Theta^{-1} \Theta M^{-1} V^T \]
\[ = U (\Delta M^{-1}) V^T, \]  
(24)

i.e., $U$, $\Delta M^{-1}$ and $V$ in the QSDV of $(H, R)$ are identical to the SVD of $X = HR^{-1}$, with $\Sigma = \Delta M^{-1}$. Hence, working with the QSDV of $(H, R)$ is mathematically equivalent to working with the SVD of $HR^{-1}$.

Using the QSDV formulation, Steps 3-5 of the noise reduction algorithm reduces to the following two steps:

(i) Compute the QSDV of $(H, R)$:

\[ H = U \Delta \Theta^{-1}, \quad R = V M \Theta^{-1}. \]  
(25)

(ii) Obtain $Z_K$ as follows:

\[ \Delta_1 = \text{diag}(\delta_1, \delta_2, \ldots, \delta_K), \]  
(26)

\[ Z_K = U \begin{bmatrix} F \Delta_1 & 0 \\ 0 & 0 \end{bmatrix} \Theta^{-1}. \]  
(27)

We see that the pre-whitening is now an integral part of the algorithm, and we refer to the complete algorithm based on the QSDV implementation as the Truncated QSDV algorithm. A similar use of the truncated QSDV is suggested in [10] in connection with regularization problems.

4.3 Updating Issues

In real-time speech processing applications, one would like to update the QSDV instead of recomputing it. In connection with the Truncated QSDV algorithm, we see from (27) that it is only the matrices $U$, $\Delta$ and $\Theta^{-1}$ that we need to update. However, it may be more convenient to update the matrices $M$ and $V$ instead of the matrix $\Theta^{-1}$, because $M$ is diagonal and $V$ is orthogonal, while $\Theta^{-1}$ is merely non-singular. In this case, one must compute the intermediate quantity $X_K$ (20) and then compute $Z_K = X_K R$. That is, the cost of this approach is an extra matrix multiplication (with $R$). Updating of the Truncated QSDV algorithm is a topic of current research.

5 Experiments

In this section we present some experiments from applying our Truncated QSDV algorithm to speech signals corrupted by additive noise. The algorithm was programmed in MATLAB, and the QSDV step was implemented along the lines described in [11]. Moreover, the speech sentences were uttered by two male and two female speakers and the noise source had the following spectral density:

\[ S(\omega) = \frac{1}{\frac{1}{2} - \cos(\omega T_s)}. \]  
(28)

All the signals were sampled at 8 kHz; $T_s = 125$ ms. The segment length, $N_s$, was set to 160 samples (20 ms) which is a standard length in many communication systems; and the number of columns, $M$, in the Hankel data matrix was set to 20. This choice of $M$ is based on experiments, and it results in a quite efficient implementation (the "skinnier" matrices, the faster the algorithm) while not degrading the reconstructed speech significantly.
noise-free speech segment much more closely in the regions near the three lowest formants than the noisy one does.

6 Summary

A signal subspace approach for noise reduction of speech signals is presented and applied successfully to speech corrupted by colored broad-band noise. The algorithm is formulated by means of the QSVD. With this formulation, the pre-whitening becomes an integral part of the algorithm and this is essential in connection with updating issues in real-time recursive applications. Updating of the Truncated QSVD algorithm is a topic of current research.

Acknowledgements

Søren Holdt Jensems research was supported in part by Fabrikant Otto Johannes Brunn Fonds, in part by UNiC, and in part by the European Community Research Program HCM, contract no. ERBCHBI-CT92-0182.

References

Bandwidth Enhancement of Narrow-Band Speech Signals

Holger CARL
Grundig Professional Electronics GmbH
Grundig Office, Kurgartenstr. 37
D-90762 Fürth, Germany
Tel/Fax: +49 [011] 703 114 / 703 8272

and

Ulrich HEUTE
Institute for Network and System Theory
Christian Albrechts University
D-24143 Kiel, Germany
Tel/Fax: +49 [0431] 77572 265 / 77572 281

Abstract. Although there is an increasing interest in wide-band speech coding, up to now most of the existing telephone systems are designed for transmitting narrow-band speech signals. So, to avoid the quality degradation due to the bandwidth limitation, it could be desirable to perform some kind of post-processing at the receiver of a narrow-band transmission system in order to enhance the decoded speech signal in such a way that it sounds as if it would have been transmitted directly from the speaker via a channel wider than 4 kHz. In this contribution, a novel bandwidth-enhancement scheme is presented. It allows to improve the quality of narrow-band speech signals by adding synthetic components outside the telephone-frequency range based on a receiver-only type of processing. An LPC-based method is used for the spectral shaping of the added components while, e.g., a simple BB-RELP HFR method serves for generating the spectral fine-structure. An efficient way of integrating the scheme into the receiver of an LP coding system is presented.

1. Introduction

In many speech transmission systems, there is a particular degradation of the transmitted signals: Due to the bandwidth limitation of the original speech signal to the frequency range up to at most 4 kHz, narrow-band speech signals exhibit a muffled character, and in the case of missing essential low-frequency components they sound tinny as well.

In terms of intelligibility this is not too severe a drawback. However, especially when directly compared to high-quality speech signals as, e.g., during a TV or broadcast interview with the interviewer talking from the studio to a partner on the phone, the lack of naturalness becomes apparent.

In the early seventies the BBC searched for enhancement schemes which were to avoid the unnatural sound of narrow-band speech signals [1]. They performed some experiments to reconstruct the spectral components lost on the narrow-band transmission path, e.g., by introducing a noise generator or by using a non-linearity. Patrick worked on the bandwidth enhancement problem as well [2], [3].

However, none of their approaches is capable of enhancing narrow-band speech signals, unless additional information extracted from the original wide-band signal is used.

Dietrich succeeded in enhancing the bandwidth of speech signals sampled at 16 kHz, but his scheme fails to yield improved quality when applied to narrow-band speech signals [4].

In this contribution, a new algorithm for enhancing narrow-band speech signals is presented. It allows to improve the quality of narrow-band speech signals by adding synthetic components outside the telephone-frequency range based on a receiver-only type of processing. Wide-band speech signals sampled at 16 kHz are derived from narrow-band speech signals which have been sampled at a rate of 8 kHz.

The problem of generating synthetic components outside the telephone frequency band can be subdivided into two tasks: i) a suitable short-time spectral envelope of the synthesized components has to be achieved and ii) the spectral fine-structure in the synthetic band(s) has to be made reasonably similar (in terms of perception) to that of the (unknown) original wide-band speech signal.

In this paper, the first task shall be addressed in more detail. The solution to this problem may consist of a system which, after a training with wide-band speech signals, is capable of detecting the spectral envelope of a speech sound in the frequency range outside the telephone band based on the narrow-band portion of the short-time spectrum. To achieve this, an LPC-codebook method was developed which is outlined in the following section. The spectral distance measures used therein are given. When applied to the bandwidth enhancement task, they have to include frequency weighting. In the context of bandwidth enhancement, proper scaling of the added spectral components is a crucial point. Therefore, not only reasonable spectral shaping but also applying the optimum amount of energy is important for achieving good quality. For this scaling problem, an efficient solution is presented.

The spectral fine-structure for the synthetic components can be generated in several ways. When synthesizing these components using an LPC all-pole filter, HFR methods as known for BB-RELP coding [5] can be adopted. Also, Harmonic Modelling [6] can be employed. In section 3, a synthesis arrangement according to the common LPC filter model is presented along with information on the performance of different methods for deriving an appropriate excitation. This synthesis method is shown to be directly implementable in the receiver of a linear-predictive speech transmission system. In this contribution, the reconstruction of high-frequency components will be addressed more deeply than the extension of the frequency range towards the lower-frequency end.

Finally, in section 4 the results obtained from listening tests are presented.
2. Envelope Extrapolation by LPC—Codebook Approach

A simple approach to extrapolate the spectral envelope of a narrow-band speech signal is as follows: An LPC codebook consisting of wide-band LPC vectors must be available. During the enhancement process, the short-time spectra of the narrow-band speech signal to be enhanced are computed at a constant rate (e.g., every 20 ms). A comparison, restricted to the frequency range of the narrow-band speech signal, is carried out between the narrow-band spectral envelope and the corresponding parts of the envelopes stored in the codebook. In each frame, the best-fitting entry is selected as the desired wide-band envelope estimate.

Usually LPC codebooks, if they shall remain tractable in size, include just different shapes of the spectral envelopes; the gain factor is treated elsewhere (gain/shape VQ). The shape codebook has to be generated and searched using a spectral distance measure which accounts for the differences in shape, only. A well-known distance measure of this type is the gain-optimized version of the Itakura-Saito distance measure, known as the Itakura distance measure. An equivalent (in the nearest-neighbor sense) measure is the gain-normalized version of the Itakura-Saito distance measure, the likelihood-ratio distance measure [7].

The distance measures given in the following indicate the distance of two power spectra represented by all-pole models:

$$d \left( \frac{\sigma^2}{|A(e^{j\Omega})|^2}, \frac{g^2}{|A(e^{j\Omega})|^2} \right)$$

is the distance of a test model

$$\frac{g^2}{|A(e^{j\Omega})|^2} = \frac{\sigma^2}{1 + \sum_{i=1}^{p} \tilde{a}_i e^{-j\Omega i}}$$

from a reference model

$$\frac{\sigma^2}{|A(e^{j\Omega})|^2} = \frac{\sigma^2}{1 + \sum_{i=1}^{p} a_i e^{-j\Omega i}}.$$  \hspace{1cm} (1)

The Itakura-Saito distance including a frequency weighting function is given by

$$d_{IS}^{(W)} = \int_{-\pi}^{\pi} \frac{W(e^{j\Omega})^2}{|A(e^{j\Omega})|^2} \left( \frac{\sigma^2}{g^2} - \ln \frac{\sigma^2}{g^2} \right) |W(e^{j\Omega})|^2 \frac{d\Omega}{2\pi}$$

with

$$\int_{-\pi}^{\pi} |W(e^{j\Omega})|^2 \frac{d\Omega}{2\pi} = 1.$$  \hspace{1cm} (2)

Gain optimization ($\sigma$ is chosen such that for given parameters $\sigma$, $a_i$, and $\tilde{a}_i$, $d_{IS}^{(W)}$ is minimum) leads to a frequency-weighted equivalent to the Itakura distance measure:

$$d_{I}^{(W)} = \ln \frac{\left| \frac{\hat{A}(e^{j\Omega})}{A(e^{j\Omega})} \right|^2}{|W(e^{j\Omega})|^2} \left| W(e^{j\Omega}) \right|^2 \frac{d\Omega}{2\pi}$$

and

$$\int_{-\pi}^{\pi} |W(e^{j\Omega})|^2 \frac{d\Omega}{2\pi} = 1.$$  \hspace{1cm} (3)

Using one of these measures (5) or (6), the envelope comparison to be performed during the codebook search can be restricted to any desired frequency range by appropriately choosing $W(e^{j\Omega})$. In general the evaluation of the distance measure will have to be performed by means of FFT computations, i.e., according to

$$d_{I}^{(W)} \approx \sum_{\mu \neq \mu_0} \frac{\hat{A}_F(\mu)^2|SF_{FT}(\mu)|^2}{\mu_0 - \mu + 1}$$

with

$$\hat{A}_F(\mu) = FFT_M \{ \hat{a}_i \} = \sum_{i=0}^{p} \hat{a}_i e^{-2\pi i \mu}$$

and

$$SF_{FT}(\mu) = \frac{\sigma^{(NB)}}{\text{FFT}_M \{ \hat{a}_i^{(NB)} \}} = \sum_{i=0}^{p} \frac{\hat{a}_i^{(NB)} e^{-2\pi i \mu}}{\sigma^{(NB)}}.$$  \hspace{1cm} (4)

This corresponds (approximately) to choosing $W(e^{j\Omega})$ to be constant inside and zero outside the frequency range of the narrow-band speech signal. Here, the suffix (NB) indicates that the variable belongs to narrow-band speech signals ($f_s = 8$ kHz) while the other variables refer to wide-band signals ($f_s = 16$ kHz). The different sampling rates are accounted for by the different FFT lengths ($M$ and $M/2$) used in (6) and (9), respectively. The summation limits $\mu_0$ and $\mu_u$ correspond to the frequency range the comparison shall be restricted to — usually the telephone frequency range.

As compared with the efficient manner that can be used to evaluate the unwighted likelihood-ratio distance measure, this FFT-based computation is by far more complex.

This conceptually simple envelope extrapolation scheme has been realized as follows: The likelihood-ratio distance which is consistent with the classical LPC analysis was selected so that the codebook generation consisted of a standard task. A 12-leaf tree-structured codebook (2 layers with 64 branches existing each node) was created based on roughly 80000 frames (of 25 ms length each) extracted from the TIMIT database (wide-band speech). The LPC order was chosen to be 16. The codebook search was carried out using (7) with $p^{(NB)} = 10$.

This initial approach was refined: Our so-called extended codebook approach allows for a more efficient codebook search and leads to improved quality of the resulting envelope (in terms of perception).
The method described above is based on a codebook which consists of a representative collection of spectral envelopes of wide-band speech signals. For a given narrow-band speech signal, the best-fitting (within the telephone band) wide-band envelope is selected from this collection. Since the basis for this selection is the narrow-band speech signal to be enhanced, it is preferable to tackle the problem starting from this side: First, from a representative collection of spectral envelopes of narrow-band speech signals the best-fitting one is identified. Afterwards, the wide-band envelope typically related with this particular narrow-band codebook entry has to be retrieved from another table — the so-called shadow codebook.

Thus, the codebook search problem reduces to a standard task: the vector quantization of the narrow-band LPC information. So, in this case the (unweighted) likelihood-ratio distance can be used allowing for a highly efficient codebook search.

To put this idea into action, an LPC codebook had to be generated using narrow-band speech data. Once again, the training material was taken from the TIMIT data base by (zero-phase) filtering and downsampling the wide-band speech signals contained therein. Again, a 12-bit tree-structured codebook (2 layers) was created, this time for an LPC order of 16. A record was kept of the training results, in particular of the assignment of every single training frame to the clusters of the final codebook. After training completion, this enabled us to build up a shadow codebook using the wide-band counterparts of the training frames. This basic generation procedure is illustrated in figure 1.

![Figure 1: Generation of narrow-band LPC codebook with associated shadow codebook consisting of the corresponding wide-band envelopes](image)

The correspondence of narrow- and wide-band versions of the (in this illustration example 12) training frames is shown. Their spectral envelopes were stored using LPC representations of orders 10 and 16, respectively. The narrow-band training material is used to generate an LPC codebook consisting of (in this example 4) narrow-band envelopes. Subsequently, the clustering resulting from the training procedure is projected to the wide-band counterparts of the training frames. In the exemplar of example 1: The training ends up with the second code vector n2 of the narrow-band codebook being the centroid of a cluster consisting of the narrow-band training frames 5, 10, and 12. Thus, in the shadow codebook a wide-band envelope w2 corresponding to n2 is entered which is the centroid of the wide-band training frames 5, 10, and 12.

There is only one drawback of this extended codebook approach using a narrow-band LPC codebook together with an associated shadow codebook consisting of wide-band LPC envelopes: A further codebook (for LPC order 10) has to be stored.

This method yields the shape of the desired wide-band envelope. For finally fixing the extrapolated envelope according to (1), the gain factor \( \hat{\sigma} \) remains to be determined. For optimizing \( \hat{\sigma} \), the distance measures given above can be used, again. So, the optimal gain factor according to the Itakura-Saito distance measure (3) is given by

\[
\hat{\sigma}^2_{\text{opt}} = \sigma^2 \int_{-\pi}^{\pi} \left( \frac{\hat{A}(e^{j\omega})}{|\hat{A}(e^{j\omega})|^2} \right)^2 |W(e^{j\omega})|^2 \, d\omega.
\]  

Here, once again we can use an FFT-based evaluation:

\[
\hat{\sigma}^2_{\text{opt}} \approx \frac{1}{\mu - \mu/L + 1} \sum_{\mu = \mu/L}^{\mu/L \cdot L} |\hat{A}_{\text{FFT}}(\mu)|^2 |S_{\text{FFT}}(\mu)|^2.
\]

The computational complexity can be reduced here, too, virtually without any loss in quality: The method outlined above derives, in a first step, the shape of the wide-band envelope from that of the narrow-band spectral envelope. In a second step, e.g. (10) is used to determine the gain factor. That means that, in addition to adapting to the gain factor of the narrow-band spectrum, the gain factor optimization accounts for the envelope shape, once again. As, however, the shape has already been accounted for in the codebook search, it should be possible to merge these two steps into one.

If we look at the normalized optimal gain factors

\[
g(\omega)^2 = \frac{\hat{\sigma}^2(\omega)}{(2\sigma_{\text{opt}}^2(\omega))^2}
\]

for all those training frames \( \nu \) that have been assigned to a specific cluster during codebook training, a narrow distribution around a cluster-specific mean value can be observed. Therefore, it is possible to assign to each cluster (i.e. to each codebook entry) a fixed representative normalized optimal gain factor \( g(\nu)^2 \) (for the \( n \)-th cluster) which can be retrieved from a table after finishing the codebook search. This saves computational effort as the evaluation of (10) is replaced by a simple table look-up.

When optimizing the normalized gain factor for a cluster, a frequency weighting function \( W(e^{j\omega}) \) can be used which accounts for deviations outside the frequency range of the narrow-band speech signals. So, for each cluster the total
error
\[ D^{(n)} = \sum_{\nu=1}^{N} d(W) \left( \frac{g^{(n)}(\nu)^2}{|A^{(n)}(\nu)|^2} + 2 \frac{\sigma^{(n)}(\nu)}{|A^{(n)}(\nu)|^2} \right) \]

has to be minimized with respect to \( g^{(n)}(\nu)^2 \). Using the frequency weighted Itakura–Saito distance measure this leads to
\[ g^{(n)}(\nu)^2 = \frac{1}{N} \sum_{\nu=1}^{N} \frac{\sigma^{(n)}(\nu)}{|A^{(n)}(\nu)|^2} \int_{-\infty}^{\infty} \frac{|A^{(n)}(\nu)e^{(i\omega)}}{|A^{(n)}(\nu)|^2} W(e^{(i\omega)}) \frac{\sigma^{(n)}}{2\pi} \, d\omega . \]


Synthesis of the spectral components to be added to the narrow-band speech signal can easily be performed by applying an LPC synthesis–filtering arrangement. Using the LPC codebook approach outlined above, we obtain information concerning the spectral envelope (including the energy information) described by a set of LPC parameters. So, what remains to be determined for the generation of a wide-band signal is a suitable excitation signal for the given synthesis filter. The spectral components of the generated signal outside the frequency range of the narrow-band speech signal are the components to be added in order to achieve the enhancement.

The basic arrangement is presented in figure 2 for the case of integrating the bandwidth enhancement scheme into the receiver of a linear–predictive speech transmission system such as, e.g., the GSM full-rate codec.

![Figure 2: Receiver of a linear–predictive speech coding system with integrated bandwidth enhancement](image)

The block HFR (High–Frequency Regeneration), in the most simple form a spectral folding function, generates the wide-band excitation signal from the narrow-band (coded) LPC residual. The problem of generating a wide-band excitation signal is well–known from Base–Band RELP coding [5]. In this context several HFR methods have been proposed. In our experiments Makhoul’s spectral–folding method [9] turned out to be a good choice for extending bandwidth towards the higher–frequency end. Other HFR schemes such as to application of non–linearities achieve an extension towards lower frequencies, also. For this purpose, however, Harmonic Modelling proved to be superior. Upsampling and interpolation of the decoded narrow-band speech signal is performed using output filter 1 before the synthetic spectral components are added after being selected by output filter 2.

4. Results

As mentioned earlier, part of our experiments consisted of enhancing low–pass filtered narrow–band speech signals, i.e. the frequency range was to be extended towards higher frequencies, only. For this case, informal listening tests employing more than ten listeners resulted in a 90 % preference of the processed speech signals when compared directly with the narrow–band "originals". The increase in naturalness due to the removal of muffling when enhancing low–pass filtered narrow–band speech signals was considered perceptually so important that some high–frequency background noise introduced by the bandwidth enhancement scheme was accepted. Furthermore, the discriminability of /s/ and /f/ sounds was improved, thus allowing for increased intelligibility of the processed speech signals.

The regeneration of low–frequency components turned out to be somewhat more difficult, but especially when using the Harmonic Modelling approach our experiments resulted in improved speech quality here, too.

References


Spectral Subtraction Based on Minimum Statistics

Rainer MARTIN

Institute for Communication Systems and Data Processing (IND), Aachen University of Technology, Templergraben 55, 52056 Aachen, Germany, Phone: +49 241 806984, Fax: +49 241 8888186, E-Mail: martin@ind.rwth-aachen.de

Abstract. This contribution presents and analyses an algorithm for the enhancement of noisy speech signals by means of spectral subtraction. In contrast to the standard spectral subtraction algorithm the proposed method does not need a speech activity detector nor histograms to learn signal statistics. The algorithm is capable to track non stationary noise signals and compares favorably with standard spectral subtraction methods in terms of performance and computational complexity. Our noise estimation method is based on the observation that a noise power estimate can be obtained using minimum values of a smoothed power estimate of the noisy speech signal. Thus, the use of minimum statistics eliminates the problem of speech activity detection. The proposed method is conceptually simple and well suited for real time implementations. In this paper we derive an unbiased noise power estimator based on minimum statistics and discuss its statistical properties and its performance in the context of spectral subtraction.

1. Introduction
Spectral subtraction is a well known speech enhancement technique which has by now developed numerous facets [1]. At the heart of a spectral subtraction algorithm is a noise power estimator and a subtraction rule which translates the subband SNR into a spectral weighting factor, such that subbands with low SNR are attenuated and subbands with high SNR are not modified. Both the noise power estimator and the subtraction rule have significant impact on the audible residual noise [2]. The basic spectral subtraction algorithm which requires only one microphone employs a speech activity detector to update noise statistics. Therefore, tracking of varying noise levels might be slow and confined to periods of no speech activity.

In this paper we address the problem of noise power estimation and develop an algorithm which essentially eliminates the need for explicit speech pause detection without an substantial increase in computational complexity. While the conventional approach to spectral subtraction employs a speech activity detector we here use the minimum of the subband noise power within a finite window to estimate the noise floor. The algorithm is based on the observation that a short time subband power estimate of a noisy speech signal exhibits distinct peaks and valleys (see Figure 1). While the peaks correspond to speech activity the valleys of the smoothed noise estimate can be used to obtain an estimate of subband noise power. To obtain reliable noise power estimates the data window for the minimum search must be large enough to bridge any peak of speech activity.

The remainder of this paper is organized as follows. In section 2 we will present the spectral subtraction algorithm and the noise power estimation method. In section 3 we will discuss the statistical properties of minimum subband noise power estimates. Section 4 will present experimental results.

2. Description of Algorithm
A block diagram of the basic spectral subtraction method is shown in Figure 2. The algorithm appropriately modifies the short time spectral magnitude of the disturbed speech signal such that the synthesized signal is perceptually as close as possible to the undisturbed speech signal. The optimal weighting of spectral magnitudes is computed using a noise power estimate and a subtraction rule.

Figure 1: Short time subband power and estimated noise floor of noisy speech signal (f = 8kHz, W_{DFT} = 256, subband k=8)

2.1 Spectral Analysis/Synthesis
We assume that the bandlimited and sampled disturbed signal s(t) is a sum of a zero mean speech signal s(t) and a zero mean noise signal n(t), s(t) = s(t) + n(t), where t denotes the time index. We further assume that s(t) and n(t) are statistically independent, hence E[s^2(t)] = E[s^2(t)] + E[n^2(t)]. Spectral processing is based on a DFT filter bank with W_{DFT} subbands and with decimation/interpolation ratio R [3]. The phase of the disturbed signal is not modified. We denote the data window by h(t) and the DFT of the windowed disturbed signal s(t) by

\[ X(\lambda,k) = \sum_{\mu=0}^{W_{DFT}-1} z(\lambda R - \mu) \cdot h(\mu) \cdot \exp \left( -j \frac{2\pi \mu k}{W_{DFT}} \right) \] (1)
\( \lambda \) and \( k \) refer to the decimated time index and the DFT frequency bins \( \Omega_m = 2\pi k / W_{DFT} \), \( k \in 0, 1, ..., W_{DFT} - 1 \), respectively. Typically we use a DFT length of \( W_{DFT} = 256 \) and decimation ratio \( R = 64 \). The improved subband signals are converted back to the time domain using an inverse DFT. The synthesized improved speech signal is denoted by \( y(t) \), the corresponding spectral magnitude by \( |Y(\lambda, k)| \).

![Figure 2: Block diagram of spectral processing](image)

### 2.2 Subtraction Rule

Let \( P_n(\lambda, k) \) and \( |X(\lambda, k)|^2 \) denote the estimated subband noise power and short time signal power, respectively. To obtain the short time signal power subsequent magnitude squared input spectra are smoothed with a first order recursive network (\( \gamma \leq 0.9 \))

\[
|X'(\lambda, k)|^2 = \gamma \cdot |X(\lambda - 1, k)|^2 + (1 - \gamma) \cdot |X(\lambda, k)|^2
\]

Following the proposal of Bercuti et. al. [4] we subtract spectral magnitudes with an oversubtraction factor \( osub(\lambda, k) \) and a limitation of the maximum subtraction by a spectral floor constant \( subf \) (0.01 \( \leq \) subf \( \leq \) 0.05)

\[
|Y(\lambda, k)| = \begin{cases} subf \cdot P_n(\lambda, k) & \text{if } |X(\lambda, k)| \cdot Q(\lambda, k) \leq subf \cdot P_n(\lambda, k) \\ [X(\lambda, k)] \cdot Q(\lambda, k) & \text{else} \end{cases}
\]

where \( Q(\lambda, k) = \left( 1 - \sqrt{osub(\lambda, k) / |X(\lambda, k)|^2} \right) \)

While a large oversubtraction factor \( osub(\lambda, k) \) essentially eliminates residual spectral peaks ('musical noise') it also affects speech quality such that some of the low energy phonemes are suppressed. To limit this undesirable the oversubtraction factor is computed as a function of the subband signal-to-noise ratio \( SNR_s(\lambda, k) \) and the frequency bin \( k \), i.e. \( osub(\lambda, k) = f(\lambda, k, SNR_s(\lambda, k)) \). In general we use less oversubtraction for high SNR conditions and for high frequencies than for low SNR conditions and for low frequencies.

### 2.3 Subband Noise Power and SNR Estimation

We first compute the short time subband signal power \( P_s(\lambda, k) \) using recursively smoothed periodograms. The update recursion is given by eq.(4). The smoothing constant is typically set to values between \( \alpha = 0.9...0.95 \).

\[
P_s(\lambda, k) = \alpha \cdot P_s(\lambda - 1, k) + (1 - \alpha) \cdot |X(\lambda, k)|^2
\]

The noise power estimate \( P_n(\lambda, k) \) is obtained as a weighted minimum of the short time power estimate \( P_s(\lambda, k) \) within a window of \( D \) subband power samples [5], i.e.

\[
P_n(\lambda, k) = \min \cdot P_n(\lambda, k)
\]

\( P_n(\lambda, k) \) is the estimated minimum power and \( \min \) is a factor to compensate the bias of the minimum estimate. In section 3 we show that \( \min \) depends only on known algorithmic parameters.

For reasons of computational complexity and delay the data window of length \( D \) is decomposed into \( W \) windows of length \( M \) such that \( M \cdot W = D \). For a sampling rate of \( f_s = 8 \) kHz and a decimation ratio \( R = 64 \) typical window parameters are \( M = 25 \) and \( W = 4 \), thus \( D = 100 \) corresponding to a time window of \( (D-1) \cdot R / W_{DFT} \).

To determine the minimum of \( M \) consecutive subband power samples at time \( \lambda = \lambda_1 \) we initialize a variable \( P_{sub}(\lambda = \lambda_1, k) \) to the first of the \( M \) samples \( P_{sub}(\lambda = \lambda_1, k) = P_s(\lambda = \lambda_1, k) \). The minimum of the \( M \) samples, \( P_{sub}(\lambda_1, k) \), is then found by a samplewise comparison of the actual minimum \( P_{sub}(\lambda_1, k) \) with the short time power \( P_n(\lambda, k) \). Whenever \( M \) samples have been read, i.e. \( \lambda = \lambda_1 + M - 1 \) we store the minimum power of the last \( M \) samples \( P_{sub}(\lambda = \lambda_1 + M - 1, k) = P_{sub}(\lambda = \lambda_1 + M - 1, k) \) and the search for the minimum begins over again. The minimum power of the length \( D \) window is now easily obtained as the minimum of the last \( W \) minimum power estimates \( P_{sub}(\lambda = \lambda_1 + qM - 1, k) \) with \( q = 1, 0, ..., (2 - W) \). The decomposition of the length \( D \) window into \( W \) subwindows has the advantage that a new minimum estimate is available after already \( M \) samples without a substantial increase in compute operations.

If the actual subband power \( P_s(\lambda, k) \) is smaller than the estimated minimum noise power \( P_{sub}(\lambda, k) \) the noise power is updated immediately independent of window adjustment: \( P_{sub}(\lambda, k) = \min \cdot P_s(\lambda, k) \). Thus in case of decreasing noise power we achieve a fast update of the minimum power estimate. In case of increasing noise power the update of noise estimates is delayed by \( D + M \) samples. Finally, to control the oversubtraction factor \( osub(\lambda, k) \) we compute the SNR in each subband

\[
SNR_s(\lambda, k) = 10 \cdot \log \left( P_s(\lambda, k) / \left( \min \cdot P_n(\lambda, k) \right) \right)
\]

Figure 1 plots the short time power estimate and the estimated noise floor for a noisy speech sample. The window length \( D \) must be large enough to bridge any peak of speech activity, but short enough to follow non stationary noise variations. Experiments with different speakers and modulated noise signals have shown that window lengths of approximately \( 0.8s \) - \( 1.4s \) give good results.

### 3. Statistical Properties of the Minimum Power Estimate

In this section we derive bias and variance of the minimum estimate with the aim to develop an unbiased noise power estimator and to evaluate its statistical efficiency. As a result we will be able to compute the overestimation factor \( \min \). To facilitate the analytical evaluation of minimum estimates we assume that the noise process \( n(i) \) is stationary and that no speech is present, i.e.
\[ \pi(i) = n(i). \quad S_{xx}(\Omega) \text{ will denote the power spectral density of signal } z(i). \]

We also assume that the computation of the short time subband power estimate is done by means of non recursive smoothing of \( K \) successive magnitude squared spectra, i.e.

\[ P_s(\lambda, k) = \frac{1}{K} \sum_{m=0}^{K-1} |X(\lambda - m, k)|^2 \]  

(7)

If successive spectra \( X(\lambda, k) \) are independent the subband power estimate \( P_s(\lambda, k) \) is asymptotically chi-square distributed with mean \( \sigma^2(k) = S_{xx}(\Omega_b = 2\pi k/W_{DFT}) \cdot \sum_{m=0}^{W_{DFT}-1} h^2(m) \) and \( N = K \) degrees of freedom for DC and Nyquist frequency bins and \( N = 2K \) degrees of freedom for all other frequency bins.

\[ f_P_s(y, k) = \frac{1}{\sqrt{2\pi}\sigma^2(k)} \cdot e^{-y^2/2\sigma^2(k)} \cdot U(y) \]  

(8)

\( \Gamma() \) and \( U() \) denote the Gamma function and the unit step function, respectively. Since the data segments from which the periodogram is computed usually overlap the approximating chi-square distribution has less than \( N \) degrees of freedom. Following the approach by Welch [6] the equivalent degrees of freedom for overlapping data segments are estimated by fitting a chi-square distribution with the same mean and variance as the smoothed periodogram in eq. (7). For non recursive (eq. (7)) and for recursive smoothing (eq. (4)) the equivalent degrees of freedom are given by \( N_{rec} \) and \( N_{rec} \), respectively.

\[ N_{rec} \approx \frac{2K}{a(K)} \quad N_{rec} \approx 2 + \frac{1}{\alpha} \]  

(9)

where \( a(K) \) and \( b(\alpha) \) are functions of \( K \) or \( \alpha \), the data window \( h(i) \), and the frame rate \( R \). Typical values are given for a Hamming window of length 256 in the Appendix. Thus, regardless of what kind of smoothing is employed the noise power estimate in eq. (5) can be modelled as the minimum of \( D \) (approximately) chi-square distributed power estimates \( P_s(\lambda, k) \).

### 3.1 Minimum of uncorrelated power estimates

In this section we assume that the minimum power estimate \( P_{\text{min}}(\lambda, k) \) is based on \( D \) independent power estimates \( P_s(\lambda, k) \). This is clearly not the case if we use successive smoothed estimates \( P_s(\lambda, k) \) but with non recursive smoothing and a suitable decimation we will be able to approximate this condition. The density of the minimum of \( D \) independent power estimates is given by [7]

\[ f_P_{\text{min}}(y) = D \cdot (1 - F_P(y))^{D-1} \cdot f_P(y) \]  

(10)

where \( f_P(y) \) denotes the distribution function of the chi-square density which becomes after repeated partial integrations of eq. (8).

\[ f_P(y) = 1 - e^{-y/2\sigma^2} \cdot \sum_{m=0}^{N/2-1} \frac{1}{m!} \cdot \left( \frac{yN}{2\sigma^2} \right)^m \cdot U(y) \]  

(11)

We use eq. (10) to compute mean and variance of the minimum power estimate. Note that mean and variance of the minimum estimate are proportional to \( \sigma^2(k) \) and \( \sigma^4(k) \) respectively.

It follows that the bias of the minimum subband power estimate is proportional to the noise power \( \sigma^2(k) \) and that the bias can be compensated by multiplying the minimum estimate with the inverse of the mean computed for \( \sigma^2(k) = 1 \)

\[ \alpha_{\text{min}} = \frac{1}{E\{P_{\text{min}}\}} \cdot \sigma^2(k) = 1 \]  

(12)

The upper graph in Figure 3 plots the mean \( E\{P_{\text{min}}\} \) versus

![Figure 3: Mean (upper graph) and variance (lower graph, solid line) of minimum power estimate computed for uncorrelated data, \( \sigma^2(k) = 1 \), \( K=20 \), and \( D=1-100 \). (Dashed line gives minimum variance achieved by non recursive smoothing over \( K \cdot D \) segments).](image)

D for \( \sigma^2(k) = 1 \), \( K=20 \), and \( a(K)=2 \). To evaluate the statistical efficiency we compare the variance of the unbiased estimate \( P_n(\lambda, k) = \text{amin} \cdot P_{\text{min}}(\lambda, k) \) with the variance of a periodogram smoothed over \( K \cdot D \) data segments ('conventional' noise estimation). In this comparison the minimum subband power estimate and the conventional smoothed periodogram use the same amount of data. The lower graph in Figure 3 plots the variance for \( \sigma^2(k) = 1 \). It is immediately obvious that the variance of the minimum estimate is much larger than the variance of the smoothed periodogram. Since we typically use \( D=80-140 \) the minimum power estimation method is certainly not attractive for uncorrelated data. We now turn to case of correlated data and show that there the situation is quite different.

### 3.2 Minimum of correlated power estimates

Clearly, successive values of \( P_s(\lambda, k) \) are correlated. In the case of correlated data there is, however, no closed form solution for the probability density of the minimum or for its mean and variance available. We therefore generated data of variance \( \sigma^2(k) = 1 \), computed the smoothed periodogram (eq. (4)), and evaluated the mean and the variance of the minimum estimate. Figure 4 presents these simulation results for recursive smoothing with \( \alpha = 0.95 \). Again, the dashed line in the lower graph gives the variance for non recursive smoothing of \( D \) successive values of \( P_s(\lambda, k) \). In the region of interest, i.e. \( D=80-140 \), we now have only a small deviation of the variance of the unbiased minimum estimator with respect to the variance achieved by a conventional estimator.
During speech activity we have noise power information only within the narrow valleys of the short time power of the speech signal. Thus, the effective window length $D$ is much shorter than for speech pause. The reduced window length and a possible distortion of the noise power estimate by low energy phonemes imply to use a smaller overestimation factor $\alpha$ during speech activity. However, as the experiments show, an overestimation factor adjusted according to Figure 4 gave good results also during speech activity.

Figure 4: Mean (upper graph) and variance (lower graph, solid line) of minimum power estimate computed for correlated data with $\sigma^2(\hat{e}) = 1$, $\alpha = 0.55$, and $D=1-160$ (Dashed line gives variance for non recursive smoothing of D short time power estimates).

4. Experimental Results
The proposed algorithm was compared to a 'conventional' spectral subtraction algorithm which uses the same subtraction rule and approximately the same amount of data to estimate the subband noise power. The 'conventional' spectral subtraction was equipped with an ideal speech activity detector (i.e. manually segmented speech files) and a noise power estimator equivalent to eq. (4). The speech material consisted of 8 phonetically balanced German speech samples recorded by two male and female speakers. After adding car noise at two different levels the segmental SNR during speech activity was 6dB and 0dB.

After the first experiments it became clear that the most crucial compromise is that of power smoothing (controlled by $\alpha$) versus window length for minimum search (controlled by $D$). As it is evident from Figures 3 and 4 smoothing is a statistically more efficient procedure than the minimum search. With too much smoothing, however, the valleys of power (see Figure 1) are not pronounced enough to warrant reliable noise estimates. We therefore use an additional first order network with a pole at 0.9 to smooth the minimum estimates. With this additional measure and the other parameters set to $D = 100$, $\alpha = 0.95$, $\gamma = 0.9$, $\text{omin} = 1.5$ the speech quality and noise suppression was almost identical to the 'conventional' spectral subtraction. Occasionally, however, a phoneme is somewhat attenuated. The limitus test for our noise estimation method is of course non stationary noise and high speech activity. To investigate the performance under non stationary noise conditions the noise signal was multiplied with an increasing or decreasing ramp function prior to adding it to the clean speech signal. In these experiments the SNR of the speech samples thus varied between 6dB and 0dB and as expected, the performance of the proposed method was indeed superior to the performance of the 'conventional' spectral subtraction.

5. Conclusion
The proposed minimum subtraction method eliminates the need for a speech activity detector by exploiting the short time characteristics of speech signals. For stationary noise the performance of our method is very close to the performance of spectral subtraction with conventional noise power estimation and ideal speech activity detection. For non stationary noise the method is at a clear advantage. The theoretical analysis shows that for typical parameter settings, the variance of the minimum estimator is less than twice as large as the variance of the conventional noise power estimator. However, more experiments with different speaking situations and different languages are necessary to improve the trade-off between the smoothing constant and the window length for minimum search.

Appendix: Equivalent Degrees of Freedom (see [6])

<table>
<thead>
<tr>
<th>$K$</th>
<th>$a(K)$</th>
<th>$N_{rec}$</th>
<th>$\alpha$</th>
<th>$b(\alpha)$</th>
<th>$N_{rec}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.98</td>
<td>10.1</td>
<td>0.7</td>
<td>1.75</td>
<td>6.48</td>
</tr>
<tr>
<td>20</td>
<td>2.04</td>
<td>19.6</td>
<td>0.9</td>
<td>1.98</td>
<td>18.2</td>
</tr>
<tr>
<td>100</td>
<td>2.09</td>
<td>95.6</td>
<td>0.95</td>
<td>2.04</td>
<td>38.2</td>
</tr>
</tbody>
</table>

Table 1: $a(K)$, $N_{rec}$, $\alpha$, $b(\alpha)$ and $N_{rec}$ for various values of $K$ and $\alpha$ computed for a Hamming window with $W_{DFT} = 256$ and $R=64$.

References
Speech Enhancement Algorithms Using MMSE Estimation with Interframe Constraint

Zhigang Cao, Bin Tian and Xiaowen Wang

State Key Lab. on Microwave and Digital Communications, Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, China,
Tel: [861]2594884, Fax: [861]2564176, E-mail: cao@ehepc2@slacvex.bitnet

Abstract. This paper mainly discusses the speech enhancement algorithms based on the Minimum Mean Square Error (MMSE) estimation of speech short time spectrum. On the basis of estimation of the general MMSE estimation, a new algorithm using interframe distribution constraints in frequency domain in MMSE estimation is proposed, if take into account the correlation between interframe spectra and incorporates them into the estimation. Due to the fact that the distribution of speech spectra is unknown, two solutions are adopted: Gaussian distribution assumption and statistic approximation obtained by real speech materials. The new algorithm has been evaluated in our speech enhancement experimental system, and compared with other algorithms. The results show the new algorithm is more effective to reduce the background noise, and there is no significant speech degradation.

1. Introduction

The effect of additive background noise on speech compression coder and speech recognizer performance can be substantial, even overwhelming when the signal-to-noise ratio is low. The objectives of speech enhancement are to develop some noise suppression algorithms and techniques to improve the overall quality of speech compression and recognition, to increase intelligibility and to reduce listener fatigue.

Generally speaking, there are two classes of speech enhancement techniques which are most widely used and proved to be efficient currently. The first one considers that speech is produced by an autoregressive model. Lim&Oppenheim proposed an iterative algorithm[1], which estimates the parameters of the autoregressive from the noisy speech then formulate a filter with these parameters to enhance the speech. This class of method has been developed a lot after that[2,3]. Another class of techniques is based on short time spectrum estimation. The main idea of this class is to try to restore each frequency channel (generally DFT coefficients) from the noisy speech. As human hearing is insensitive to the speech phase, the short time spectrum amplitude(STSA) is mostly concerned in estimation. The main advantage of this class is its relatively low computation complexity. In [4] and [5], Y.Ephraim and S.F.Boll separately proposed the algorithm of Minimum Mean Square Error (MMSE) estimation of speech short-time spectral amplitude. Both algorithms were reported to achieve good results.

This paper mainly discusses the speech enhancement algorithms based on the Minimum Mean Square Error (MMSE) Short Time Spectrum estimation which has been proved to be successful. The previous MMSE estimations developed by [4,6] assumed the spectral components between adjacent frames are independent. In fact, the assumption does not reflect the nature of speech, especially for voiced segments. Figure 1 shows the conditional probability of the normalized spectrum amplitude of the present frame when the distribution of the spectrum amplitude within preceding frame is already known. The number by the curve is the normalized spectrum amplitude(dB) of the preceding frame. The figure illustrates clearly that the conditional distribution of the spectrum amplitude focuses in three different ranges within present frames, which means the spectra between two frames are correlated to some extent. Moreover, we often use frame overlap in practical processing which increases the correlation of spectra between two adjacent frames. In this paper, a new type of algorithms using interframe constraints in frequency domain is proposed, which takes into account the correlation between the interframe spectra and incorporates them into MMSE estimator.

![Figure 1. Conditional Probability of Normalized Speech Spectral Amplitude within Present Frame](image-url)
This paper describes and derives the estimation algorithms and the real-time realization. Comparing with previous MMSE algorithms and other enhancement algorithms, practical noisy speech experiments show that the proposed algorithms obtain significant improvement.

2. Estimation Algorithms

2.1 General MMSE-STSA Algorithm

Only noisy speech itself is available in a single-microphone acquisition enhancement system. The noisy speech can be described as \( y(t) = x(t) + d(t) \) where \( x(t) \) and \( d(t) \) denote the speech and the noise process, respectively. Let \( Y_k = R_k \exp(j\theta_k) \), \( X_k = A_k \exp(j\alpha_k) \) and \( D_k \) denote the \( k \)th spectral component of the noisy speech \( y(t) \), the pure speech \( x(t) \) and the noise \( d(t) \), respectively. The goal is to estimate \( A_k \) from \( Y_0, Y_1, \ldots, Y_N \).

We define the nonlinear gain function of the \( k \)th spectral component by \( G_k = \frac{|X_k|}{R_k} \). On the assumptions that noise is of Gaussian distribution and the spectral components are independent each other according to the Bayes formula the MMSE gain function can be described as[7]

\[
G_k = \frac{\int_{-\infty}^{\infty} b_k p(b_k) q(b_k, \xi_k, \gamma_k) db_k}{\int_{-\infty}^{\infty} p(b_k) q(b_k, \xi_k, \gamma_k) db_k} (1)
\]

\[
q(x, y, z) = \exp(-x^2 y) I_0(2x \sqrt{yz})
\]

\[
\hat{A}_k = R_k \cdot G_k
\]

Where \( \gamma_k = \frac{R_k^2}{\lambda_e(k)} \) is a posteriori signal-to-noise ratio and \( \xi_k = \frac{\lambda_e(k)}{\lambda_d(k)} \) is a priori signal-to-noise ratio, in which \( \lambda_e(k) \) denotes the variance of the speech distribution \( p(\alpha_k) \) and \( \lambda_d(k) = \mathbb{E}[|D_k|^2] \) denotes the variance of the noise obtained from the statistical analysis when speech is absent. Let \( b_k \) and \( p(b_k) \) denote the normalized amplitude of speech spectrum and its distribution. \( I_0(\cdot) \) denotes the modified Basal function of zero order.

2.2 MMSE-STSA estimation with distribution constraints in frequency domain (M-FDDC)

Let \( Y_k = R_k \exp(j\theta_k) \), \( X_k = A_k \exp(j\alpha_k) \) and \( D'_k \) denote the \( k \)th spectral components of the noisy speech, the noise and the pure speech process of the preceding frame, respectively. According to different parameters used as constraints in the preceding frame, three types of M-FDDC estimation can be derived. In the following the subscript \( k \) which represents the \( k \)th spectral component is omitted for the convenience of writing.

2.2.1 M-FDDC-1

In this method, we use the complex spectrum of the noisy speech of the preceding frame to be the constraint parameter. Thus the MMSE-STSA estimation formula becomes

\[
\hat{A} = \mathbb{E}[A|Y, Y'] \cdot \frac{\int_{-\infty}^{\infty} \frac{1}{2\pi} p(a, a', \alpha, \alpha') p(Y, Y'|a, a', \alpha, \alpha') da da' d\alpha d\alpha'}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} p(a, a', \alpha, \alpha') p(Y, Y'|a, a', \alpha, \alpha') da da' d\alpha d\alpha'} (3)
\]

This is a four-dimension integral and is hard to calculate.

2.2.2 M-FDDC-2

To overcome the difficulty of calculation, considering the characteristics of human hearing, the amplitude of the preceding frame spectrum of the noisy speech is used as the constraint. The gain function is obtained as

\[
G = \frac{\mathbb{E}[A|R, R']}{\mathbb{E}[R]}
\]

\[
= \frac{\int_{-\infty}^{\infty} \frac{1}{2\pi} p(b, b') q(b, \xi, \gamma) q(b', \xi', \gamma') db db'}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} p(b, b') q(b, \xi, \gamma) q(b', \xi', \gamma') db db'} (4)
\]

Where \( p(b, b') \) denotes the normalized spectral amplitude distribution of the speech in the preceding and the present frames. \( \xi, \xi', \gamma, \gamma' \) denote the priori and the posteriori signal-to-noise ratios of the present and the preceding frames, respectively.

2.2.3 M-FDDC-3

We use the spectral amplitude of the pure speech of the preceding frame as constraint here. The gain function can be reformed as

\[
G = \frac{\int_{-\infty}^{\infty} \frac{1}{2\pi} p(b, B') q(b, \xi, \gamma) db}{\int_{-\infty}^{\infty} p(b, B') q(b, \xi, \gamma) db} (5)
\]

\[
B' = \frac{A'}{\sqrt{\alpha_m}}
\]

Since the spectrum of the pure speech is not known, in practice, we substitute it by the enhanced speech processed from the preceding frame.
2.3 Short Time Log-spectral MMSE Estimation

In above discussions, we use MMSE estimation mainly based on the short-time spectral amplitude. In fact, the perception of human ear is proportion to the log-spectral amplitude. Furthermore, the practice of speech processing also shows that using a distortion measure of the log-spectra is more suitable. Then we can obtain the MMSE estimator based on the short-time log-spectra as the extension of above-mentioned MMSE estimation. The short time log-spectral MMSE estimator with distribution constraints in frequency domain is obtained as

\[ \hat{A}_k = \exp\left\{E[\ln(A_k)|R_k, R_k]\right\} \] or

\[ \hat{A}_k = \exp\left\{E[\ln(A_k)|R_k, A_k]\right\} \] (6)

3. Concrete Implementation

Each formulation of estimation in above derivation will encounter two main problems in concrete implementation: one is how to obtain the distribution of the speech spectra; the other is whether the implementation can satisfy the requirement of the real-time realization for computation complexity.

There may be two resolutions to obtain the unknown distribution of the speech spectra: to assume a rational probability model or to derive it by statistic method from the practical speech materials. What should be pointed out is that these two resolutions only reflect the characteristics of speech spectra to some extent, because speech is a nonstationary and nonergodic stochastic process, the precise distribution is actually unobtainable.

The distribution of speech spectra is often assumed as a Gaussian distribution. The fact that a central limit theorem exists encourages the use of a Gaussian model. Under this assumption the above-given estimation formulas all can be further simplified to one-dimension integral computation (except M-FDDC-1). However some parameters of the Gaussian distributed model such as variance and correlation must be estimated previously.

In the statistical method, what can be obtained is the appearing frequency of one value in speech spectra from the statistics, not the real probability. However, the statistic reflects the distributed features to some extent. The effectiveness of statistics is determined by whether the result of statistics can reappear or varies unobtrusively with the training materials and the request for quantities of the speech materials. Our practice proved the distribution derived by the statistic method based on the normalized data of the speech has some stability and is not sensitive to the sources or quantities of the speech materials. So the distribution can be used for the estimation.

In a practical speech enhancement system, a gain table can be used to lookup for \( G_k \), then we can use

\[ \hat{A}_k = R_k \cdot G_k \] (7)

to get the estimator. \( G_k \) can be calculated previously by several input parameters such as \( \xi, \gamma \). The real-time estimation can be realized easily.

4. Experiment and Result

We realize the above-mentioned algorithms with speech enhancement system developed by ourselves.

The input speech first passes a 3.4kHz low-pass filter and is sampled by 8kHz, then mixed with the stationary white Gaussian noise to generate the noisy signals with the input signal-to-noise ratios of 10dB～10dB. The original speech material is a segment of Mandarin spoken by a female for 6 seconds. We use general MMSE estimation algorithm and M-FDDC-3 algorithm based on log-spectrum in our experiment. In order to make comparison, we also use spectral subtraction and Kalman filter enhancement algorithms to process the same speech materials. The evaluation of the performance is in terms of SNR and Itakura distance improvement and informal listening. Figure 2 shows the evaluation curves of the SNR values and Itakura distances. Figure 3 shows the three-dimension speech spectra of the speech before and after M-FDDC-3 algorithm based on log-spectrum processing. To compare, the general MMSE-processed three-dimension spectra are also given.

From the above results and the informal listening, we can find that some obvious performance improvements are obtained by interframe constraints. Among all the enhancement algorithms M-FDDC-3 based on log-spectrum is the best. For general MMSE estimation algorithm, the listening effect is better when Gaussian model is used, while the distribution derived from practical statistics is more suitable for the new method we propose. The reason for this is that there are too many parameters to be estimated if a Gaussian model is used in the new method. In the two pragmatic constraints we use in experiments, the use of the spectral amplitude of pure speech during the preceding frame is better, which can be illustrated not only in the computation complexity but also in the result of process.

5. Conclusion

In this paper we describe the MMSE estimations based on short-time spectral amplitude (log-spectra) of speech, and propose two new MMSE estimations with interframe constraints in frequency domain, and discuss the concrete implementation.

These methods are proved to be effective in speech enhancement through the realization on the speech enhancement system developed by ourselves. The interframe constrained MMSE estimation based on a real distribution and the short-time log spectra present the best performance.
References


Trans. on ASSP., December, 1984, 1109.


Figure 2. Evaluation of enhancement algorithms (O: Kalman filter algorithm, V: Spectral subtraction algorithm, □: General MMSE algorithm, Δ: M-FDDC-3 algorithm based on log-spectrum, +: Noisy speech)

Figure 3. Three-dimension spectra (The SNR of noisy speech is 5dB)
Speech Enhancement Based on an Auditory Model

Tore Fjällbrant and Fisseha Mekuria

Dept. of Electrical Eng., University of Linköping, SE-581 83 Linköping, Sweden
Te/Fax: +46 (13) 281304 / 139282, E-Mail: toref@isy.liu.se, fisseha@isy.liu.se

Abstract On the basis of recent observations in auditory physiology, a signal processing model of the peripheral auditory system has been simulated and a signal processing method analyzed. Auditory images with one tonotopical and one periodototopical axis are created. Results show that the developed auditory model has interesting properties with regard to separating two concurrent vowels of natural speech by a female and a male speaker. Corresponding auditory images have a high frequency resolution determined not by the bandwidth of but by the pronounced shape of auditory filter magnitude and phase characteristics.

1. Introduction

Much attention is now being paid to the development of signal processing models of the peripheral auditory system. One reason for this interest is the desire to increase our understanding and to apply methods of auditory processing, which still outperform most man-made systems with regard to recognizing and separating speech and other sounds. A healthy auditory system has, for example, the remarkable ability of selective attention in situations with a background of simultaneous voices, which is a very common form of background disturbance. Simulation of such effects has been studied in several models. A review of, and further references to, some of such monaural studies are given in [1].

Different models focus on different properties of auditory systems. A fundamental part in most models is a linear filter bank, which, in combination with non-linear operations, represents the combined processing up to the auditory nerve. In the model described in this and a previous paper [2], some very pronounced properties of filter characteristics observed in physiological measurements but overlooked in most other models are taken into account. This is combined with the simulation of the processing of amplitude modulation products and the existence of a periodototopical axis perpendicular to the tonotopical axis in accordance with recent findings [3,4,5]. It is shown that, with such a combination, auditory images can be constructed having unique potential with respect to separating two concurrent vowels from natural speech. The filter characteristics include a notch [2,6,7] (with corresponding phase-jump [2,6] or knee [8] in otherwise almost linear phase characteristics) on the high-frequency side for lower centre frequencies (CF's) and the presence of an extremely steep high-frequency edge at somewhat higher CF's [7], as well as a relatively large number of parallel channels and channels with different band-widths for the same CF [9]. The processing of amplitude modulation products includes enhancement [3,5], detection [4], band-pass [3,5] and low-pass [3,5] filtering of such products. The model also makes use of the auditory property of high phase discrimination ability [10].

2. Description of the auditory model. Result with single vowel inputs.

A block diagram of the model is shown in figure 1 and the time and frequency representation of the swedish vowel /æ/ in a 16 ms segment of a female utterance is shown in fig.4. Figures 5 to 8 depict the signal processing of the signal from this vowel. The first filter bank which is used for the female utterence consists of 26 overlapping band-pass filters equally spaced on an ERB-rate [11] scale corresponding to the tonotopical y-axis in figures 5 to 8 and with the bandwiths of the filters chosen from physiological measurements (the upper bandwidth range for a given CF [9]). For lower CF's, the filters have a notch (with corresponding 180 degree phase jump in otherwise linear phase characteristics) on the high-frequency side, figure 2, and for higher CF's, the magnitude characteristics have an extremely steep high-frequency edge, as shown in figure 3. The spatio-temporal pattern of the signals in 26 channels after the half-wave rectification of the IHC's (level A in figure 1) is shown in Figure 5, together with the averaged (over a time window of 16 ms) magnitude profile in the filters on the tonotopical axis.

As there are more than one harmonic component within the same filter channel, amplitude modulation (AM) product frequency components in the pitch frequency region will be present after the non-linear action of the IHC's. Figure 6 shows the spatio-temporal pattern of these products in the channels after suppression of all components above 500 Hz in a low-pass filter for each channel (level B in figure 1). The rectified and averaged (over a time window of 16 ms) signal magnitude profile across channels at this level is also shown in figure 6 (but not being used in the proceeding signal processing). This profile is somewhat different from that in figure 5, because when more than two harmonics components are present in the same filter channel, the magnitude of the AM-products is dependent on the relative phases between the components.

For simulation of the auditory band-pass filtering in the pitch frequency region on a periodotopical axis [3,5], an FFT analysis with computation of both magnitude and phase is performed for each channel. The auditory image shown in figure 7 is derived from this analysis by presenting the relative profile of the magnitudes obtained by the FFT analysis on the horizontal, periodotopical, axis for each channel. The relative profile on the tonotopical, y-axis, is that obtained from the time averaged magnitude profile derived from the signal in the channels before low-pass filtering (in the cochlea part of the model). The resolution on the periodotopical ("pitch") axis is determined by the analysis window, 16 ms in this case. The individual harmonic components are not resolved on the tonotopical axis because of the large relative width of the filters in the filter bank but the formant structure is represented.

The spatio-temporal pattern of the AM products after low-pass filtering (figure 6) contains, however, more information. Significant phase discontinuities between the AM products in adjacent channels occur at the "place" on the tonotopical axis of the harmonic components. This is due to the shapes of the band-
pass filters in the cochlea filter bank. When one compares channels for lower CF's, one harmonic component can be in the side lobe for one channel and in the main lobe for an adjacent channel (Figure 2). This harmonic will then be 180 degrees out of phase in the comparison, leading to a significant phase difference also between the filtered AM products. When comparing channels with higher CF's, one harmonic component can be just outside the steep edge of the filter lobe in one channel but still inside in the other channel (Figure 3). Because of the different shapes of the low- and high-frequency edges there will be one harmonic component less in one of the channels and this will also lead to a significant difference in phase as well as in magnitude of the AM-products. What is presented in the image in Figure 8 is the product of the sum of the phase and magnitude differences and the sum of the magnitudes in adjacent channels (Each next to adjacent channel is being compared in the model to avoid dead sectors). Here the harmonic components are resolved in spite of the large bandwidth of the filters in the cochlea filter bank (large bandwidth ensures good resolution in the time domain). This high frequency resolution is given by the relatively high number of channels in combination with the sharp notches and phase-jumps, respectively the extremely steep high-frequency edges. A similar effect is likely to be obtained if the phase characteristic has a sharp knee [8] instead of a phase jump. This will also be tried.

The magnitude profile on the tonotopical axis of the peaks in the image in Figure 8 is (as in Figure 7) derived from the magnitude information in the channels before low-pass filtering.

It should be possible to reconstruct the signal in the time domain using the information from the auditory image combined with phase information from the signals in the cochlea filter bank (feasible information also in auditory systems due to phase-locking at CF's below 5kHz).

Figure 10 shows the auditory image for a 32 ms sequence of the English vowel /æ/ in “theatre” for a male speaker (Figure 9). This has been processed using a second filter bank, which is a scaled version (with half the bandwidth and half the CF values) of the first filter bank. The bandwidths in this second bank correspond more to the lower bandwidth range for a given
Figure 5. Spatio-temporal response at level A (fig.1) for vowel in fig.4

CF [9]. Due to the increased time window (32 ms in this case instead of 16 ms), the resolution on the periodotopical axis is also doubled. Male speech with a lower pitch frequency needs a longer time window for analysis, but the corresponding filter bank has longer impulse responses and is not suitable for following rapid variations in female speech with higher pitch frequencies. The bandwidth of the low-pass filters is in this case 250 Hz (as compared with 500 Hz in combination with the first filter bank).

3. Auditory images for concurrent vowels

Figures 11 a-b show the resulting auditory images when two signals of approximately equal strength corresponding to the two vowels /æ/ and /e/ from the female and male speakers are added together. Despite the complete overlap of the spectra from the two vowels in the region of the spectrum of the male speech, the two vowels are represented with surprisingly little distortion. There is no doubt about the presence of two vowels and of the value of their pitch frequencies. The harmonic components are resolved with slight distortions in the overlap region. A reconstruction of both vowels is quite feasible using the information in these images in combination with phase information from the signals in the cochlea filter banks.

Figure 7. Auditory image of vowel in fig.4 with unresolved harmonics

Figure 6. Spatio-temporal response at level B (fig.1) for vowel in fig.4

The image in figure 12 is equivalent to that in figure 11 a when only the signals for one periodicity have been selected. This is very close to the undistorted image in figure 8.

4. Conclusion and discussion

It has been shown that auditory filter banks with pronounced characteristics from physiological measurements can be used together with simulated processing of amplitude modulation products in accordance with recent findings, to construct auditory images with one tonotopical and one periodotopical axis. Individual harmonic components are resolved despite the presence of several harmonics in each channel. These images have very interesting potential with respect to selective attention to simultaneous speakers. The ability of separating the signals from concurrent vowels is obtained in very short time windows applicable to natural speech, and without any previous knowledge about the signals or need for any separate pitch analysis. Some main differences compared with the model by Shamma [12], are that in this model comparisons between channels are performed using AM products after low-pass filtering and that sharply defined filter characteristics are responsible for the high frequency selectivity. Real time implementations of this model are feasible, particularly with a combination of dedicated VLSI-chips and conventional programmable DSPs.

Figure 8. Auditory image of vowel in fig.4 with resolved harmonics
References


**Figure 9.** Male vowel /el/, frequency and time representations

**Figure 10.** Auditory image of vowel in fig.9

**Figure 11a-b.** Auditory images for concurrent vowels

**Figure 12.** Image in figure 11a with signals for one periodicity selected.
Some Robust Speech Enhancement Techniques using Higher Order AR Estimation

Josep M. SALAVEDRA*, Enrique MASGRAU**, Asunción MORENO*, Joan ESTARELLAS*

* Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Apartat 30002, 08080-Barcelona, Spain. Tel/Fax: +34-3- 4017404 / 4016447. E-mail: mia@tsc.upc.es
** Department of Electrical Engineering and Computers, Universidad de Zaragoza, María de Luna, 3. 50015-ZARAGOZA, SPAIN

Abstract. We study some speech enhancement algorithms based on the iterative Wiener filtering method due to Lim-Oppenheim [2], where the AR spectral estimation of the speech is carried out using a 2nd-order analysis. But in our algorithms we consider an AR estimation by means of cumulant analysis. This work extends some preceding papers due to the authors, where information of previous speech frames is taken to initiate speech AR modelling of the current frame. Two parameters are introduced to design Wiener filter at first iteration of this iterative algorithm. These parameters are the Interframe Factor IF and the Previous Frame Iteration PFI. A detailed study of them shows they allow a very important noise suppression after processing only first iteration of this algorithm, without any appreciable increase of distortion. Two different ways to combine current and previous frame AR modelling are evaluated.

1. Introduction

It is well known, that many applications of speech processing that show very high performance in laboratory conditions degrade dramatically when working in real environments because of low robustness. The solution we propose here concerns to a preprocessing front-end in order to enhance the speech quality by means of a speech parametric modelling insensitive to the noise. The use of HO cumulants for speech AR modelling calculation provides the desirable uncoupling between noise and speech. It is based on the property that for Gaussian processes only, all cumulants of order greater than two are identically zero [1]. Moreover, the non-Gaussian processes presenting a symmetric p.d.f. have null odd-order cumulants. Considering a Gaussian or a symmetric p.d.f. noise (a good approximation of very real environments) and the non-Gaussian characteristic of the speech (principally for the voiced frames) it would be possible to obtain an spectral AR modelling of the speech more independent of the noise by using, e.g., 3rd-order cumulants of noisy speech instead of common 2nd-order cumulant.

2. Iterative Wiener Algorithm

In the original Lim-Oppenheim Method [2], noisy speech is enhanced by means of an iterative Wiener filtering that is defined as:

\[ W'(a) = \frac{P_s}{P_s + P_r} \]  

(1)

where Ps is the spectrum of the noise signal r(n), estimated in non-speech frames, and Pr is a spectrum estimation of the unavailable clean speech signal. So, both speech and noise spectra estimation must be available to design the Wiener filter at every frame. We talk over signal estimation because both signals are not available and only noisy speech signal can be processed. An iterative Wiener filtering is used to obtain a better estimation of the AR speech modelling (figure 1.). At first sight an improvement of performance can be expected after every iteration since this current AR speech estimation is carried out from a cleaner speech signal than filter estimation of the preceding iteration. But other factors sidetrack this iterative algorithm and a limitation in the number of iterations must be taken in account. Clearly the filtered speech signal contains a smaller residual noise but it presents a larger spectral distortion. Therefore, increasing the number of iterations doesn't always give a better speech estimation. It is well known that this algorithm leads to a narrowing and a shifting of the speech formants [3], providing an unnatural sounding speech. In [4] a detailed convergence analysis of this algorithm is carried out. It is proved that this estimated Wiener filter tends to cancel all signal frequencies with SNR lower than 4.77dB, and an additional attenuation, proportionally to the noise level, affects signal frequencies with higher SNR, in comparison to the optimum Wiener filter. Only the non-contaminated speech frequencies undergo a null attenuation.

![Figure 1. Scheme of the iterative Wiener algorithm.](image)

3. The Parameterized Algorithm

A parameterized Wiener filtering has been considered to have a better control over noise suppression, intelligibility loss and computational complexity, by adding two parameters \( \alpha \) and \( \beta \) in the Wiener filter computation (1). So, we consider now the following equation:

\[ W'_i(w) = \left( \frac{P_y}{P_y + \beta P_r} \right)^\alpha \]  

(2)

By varying these parameters \( \alpha \) and \( \beta \), filters with different characteristics can be obtained. Thus, if \( \alpha=\beta=1.0 \) then expression (2) corresponds to the general Wiener filter equation

This work was supported by TIC 92-0800-C05-04
AR modelling (figure 1.) of the speech spectrum estimation is computed from 3rd-order cumulants that are calculated using the covariance case:

$$C_k(i,j) = \sum_{n=p+1}^{N} x(n-k)x(n-i)x(n-j), \quad 0 \leq k, j \leq p$$  \hspace{1cm} (3)

where \( p=10 \) is the order of the filter. Then speech AR modelling coefficients \( a_k \) are computed by solving the following equations [1]:

$$\sum_{k=0}^{P} a_k \cdot C_k(i,j) = 0, \quad 1 \leq i \leq p; \quad 0 \leq j \leq i$$  \hspace{1cm} (4)

As discussed in preceding works due to the authors [5,6], we obtain a twofold benefit by considering this 3rd-order AR modelling: Firstly, an accelerated convergence of the iterative algorithm and so a reduction of both computational complexity and intelligibility loss; Secondly, achievement of a non polluted AR speech parameterization. In comparison to 2nd-order statistics estimation we obtain a good improvement but the price we pay for these advantages is a higher distortion. Thus a higher "peaking" or "narrowness" effect of the speech formants is brought about [4].

When additive noise is AWGN (SNR=0dB) the improvement over second-order algorithm is very appreciated for any number of iterations (see Table 1). While the improvement of second-order approach increases gradually, but slowly, iteration by iteration, 3rd-order one gets a very good improvement, about 34B, after only two iterations and thus it obtains a faster convergence. Furthermore, in comparison to 4th-order algorithm, third-order one also obtains better results and its computational complexity is much lower. Therefore, 3rd-order cumulants lead to a faster noise reduction because of its higher aggressiveness with respect to both 4th-order cumulants and autocorrelation function [5].

4. The Interframe Factor IF

In table 1, we may appreciate an improvement that increases gradually by iteration. Both main part of noise suppression is obtained after processing two iterations. Third-order cumulants obtain an appreciable noise suppression (about 2dB in Cepstrum distance) after first iteration (see Table 1.b) and then this speech modelling is enhanced a lot in the second iteration because it estimates Wiener filter from cleaner speech signal. At first iteration, speech AR modelling is computed from noisy signal without any initial information about the features of speech signal corresponding to the current frame. However, we now some information of the current speech frame by considering that speech signal features don't vary a lot between two consecutive overlapped frames. Therefore, we propose to obtain the first iteration AR coefficients as a combination between current frame AR estimation and previous frame AR coefficients. Thus, we design the non-causal Wiener filter, corresponding to first iteration, as a linear combination of coefficients \( a_k \) belonging to two consecutive frames, calculated as follows:

$$A_k(n,l) = IF \cdot a_k(n,l) + (1 - IF) \cdot A_k(n-1,PFJ)$$  \hspace{1cm} (5)

$$0 \leq k \leq P; \quad 1 \leq PFJ \leq \text{MAXITER} \quad 0 \leq IF \leq 1$$

where \( n \) is the current frame, \( IF \) is the Previous Frame Iteration that we consider to help first iteration of the current frame and IF is the Interframe Factor. We write \( a_k \) when coefficients are estimated directly from a noisy speech frame and we note capital letter \( A_k \) when coefficients are coming from a linear combination of \( a_k \). At the beginning of every speech activity we set parameter: IF=1 because information of last speech frame is not related to the current speech frame. Wiener filter designs corresponding to the remaining iterations of the algorithm are estimated over a cleaner speech signal coming from Wiener filtering Output of previous iteration of the same frame:

$$A_k(n,iter) = a_k(n,iter) \quad 2 \leq iter \leq \text{MAXITER}$$  \hspace{1cm} (6)

where \( iter \) is the iteration number of the current frame. We have two parameters to control this linear combination. First parameter is the Interframe Factor IF that represents the amount of current speech AR estimation \( a_k(n,1) \) we put in the AR modelling \( A_k(n,1) \) of the filter. The interframe factor is the main parameter to control linear combination (5) because parameter IF=1 represents that only current AR estimation is considered to design Wiener filter at first iteration of current frame and then parameter IF has no sense to be considered. Thus, parameter IF=1 refers to a situation where no interframe factor is defined. If we decide to consider previous frame information (IF<1) we need consider parameter IF to answer the following question: Which iteration number (PFJ) of preceding frame must we take to obtain a reliable speech AR modelling? Preceding works [5,6] have shown that it has no sense to process more than 5 iterations while third-order statistics are considered. Therefore, parameter MAXITER=5 has been fixed in all our tests.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>SNR</th>
<th>SEGGEN</th>
<th>ITAKU</th>
<th>COSTH</th>
<th>GEPFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Iter</td>
<td>0.00</td>
<td>0.79</td>
<td>9.57</td>
<td>11.67</td>
<td>12.02</td>
</tr>
<tr>
<td>1 Iter</td>
<td>1.34</td>
<td>1.57</td>
<td>5.00</td>
<td>7.82</td>
<td>7.99</td>
</tr>
<tr>
<td>2 Iter</td>
<td>1.90</td>
<td>5.56</td>
<td>5.03</td>
<td>7.54</td>
<td>8.08</td>
</tr>
<tr>
<td>3 Iter</td>
<td>2.91</td>
<td>5.88</td>
<td>5.03</td>
<td>7.43</td>
<td>7.94</td>
</tr>
<tr>
<td>4 Iter</td>
<td>1.68</td>
<td>5.78</td>
<td>4.86</td>
<td>7.45</td>
<td>8.16</td>
</tr>
</tbody>
</table>

Table 1. Distance measures using algorithms based on: a) second order statistics; b) parameterized third order cumulants; c) parameterized third order with interframe factor IF=0.7, considering parameter PFJ=3; d) fourth order cumulants at SNR=0dB.
When parameter $PFI=1$ and parameter $IF<1$, we may evaluate in (5) that current AR coefficients $a_k(n,1)$ to design Wiener Filter are coming from a linear combination of AR estimations belonging to previous frames:

$$A_k(n,1) = IF \sum_{r=2}^{n} (1-IF)^{n-r} \cdot a_k(r,1) + (1 - IF)^{n-1} \cdot a_k(1,1)$$

where $n$ is the distance in number of frames since last non-speech activity frame appeared. We have a combination of coefficients $a_k$ coming from a lot of different frames and therefore the estimation window may contain some phonemes and it doesn't fulfill our initial assumption. The worst of it is that parameter $IF=0$ means AR estimation of first frame after non-speech activity is assigned to all the frames belonging to the same speech activity. So, distortion effect may increase a lot, specially when parameter $IF$ takes small values.

To avoid this problem, another linear combination to design Wiener Filter is proposed:

$$A_k(n,1) = IF \cdot a_k(n,1) + (1-IF) \cdot a_k(n-1,PFI)$$

This estimation (8),(6) is referred as Method B while previous one (5),(6) is referred as Method A in Figure 2. A comparison of performance of both methods is shown. Cepstrum distance after first iteration of the parameterized third order algorithm is depicted with a value of parameter $PFI=1$. It can be appreciated that Method B give better estimations when parameter IF is lower than 0.5 and similar performance is achieved when the information coming from previous frame is less significant.

On the other hand, parameter $IF=0$ represents that the coming noisy speech frame is filtered by means of a filter estimation coming from preceding speech frame. Two different situations may be distinguished: $PFI=1$ and $PFI>1$. When information proceeding from first iteration of previous speech frame ($PFI=1$) is considered, no better results than before ($IF=1$) are expected, because the speech AR estimator is looking at the same noisy speech signal, but in the previous frame, and performance therefore decreases when parameter $IF$ decreases to zero (see 1st iteration line in figure 4.). However, a good improvement (about 1.5dB in Cepstrum distance) is obtained.

---

Figure 2. Comparison of two possible methods to combine AR coefficients coming from current and previous speech frames when first iteration is processed and $PFI=1$.

---

Figure 3. Distortion Effect produced by first iteration processing when some different speech AR estimations belonging to different iterations of previous speech frame are considered.

---

Figure 4. Noise Suppression after processing first iteration of current frame when some different speech AR estimations belonging to different iterations of previous speech frame are considered.
when parameter PF1=1 but distortion effect increases more than 2dB in Cepstrum distance (see figure 3) because current Wiener filter is designed with speech AR estimation proceeding from the preceding frame over a cleaner speech signal.

In figure 3., Cepstrum distance corresponding to first iteration of current frame has been represented and some different iteration numbers of preceding frame have been evaluated. Clean speech has been processed by this system and so distortion effect corresponding to the iterative algorithm has been depicted. To avoid an appreciable increase of distortion effect all values of parameter IF lower than 0.6 must be discarded. In figure 4., first iteration of current frame corresponding to speech signal disturbed by AWGN at SNR=0dB has been processed and some different speech AR estimations of previous frame have been evaluated (ranging PF1 from 1 to 5). We may come to the conclusion that values of parameter IF ranging from 0.6 to 0.8 represent a good trade-off between distortion and noise suppression. Therefore, we may achieve an improvement of 2dB in Cepstrum distance by introducing parameter IF (PF1=3 and IF=0.7) to estimate current speech AR modelling without any noticeable increase of distortion (0.25 dB). Thus, we may obtain an improvement higher than 4 dB in Cepstrum distance after processing only first iteration of the iterative Wiener filtering.

As it is depicted in Figure 5, values of parameter PF1=2 give a similar performance. After second iteration most part of linear combinations leads to similar levels of Cepstrum Distance but, in listening tests, it may be appreciated a less distortion effect when parameter PF1=2 and furthermore the best performance is achieved after 3 iterations of Lim-Oppenheim algorithm while 4 iterations are necessary when parameter IF=1, to obtain a similar quality. Therefore value PF1=3 may be considered as a good trade-off among computational complexity, distortion effect and noise suppression. This fact may be justified looking at Itakura Distance where a very important reduction (about 4.5dB) with only first iteration is achieved and therefore formants estimation in voiced sounds is clearly improved by introducing these two parameters at first iteration of every frame. Obviously this linear combination of coefficients as tends to improve quality inside of voiced sounds. Some constraints have been added to the algorithm to protect unvoiced frames against this linear combination and so parameter IF=1 is set inside of unvoiced frames and first voiced frame. Similar performance is achieved when diesel engine noise is considered, although differences between AR weighting method (IF<1) and no interframe weighting method (IF=1) are smaller.

5. Conclusions

A speech enhancement method based on an iterative Wiener filtering have been proposed. Spectral estimation of speech is obtained by means of an AR modelling based on third-order cumulant analysis to provide the desirable noise-speech uncoupling. Two parameters, IF (Interframe Factor) and PF1 (Previous Frame Iteration), have been considered to take advantage of previous speech spectrum estimations to initiate AR modelling corresponding to first iteration of the current speech frame. This approach achieves a noise suppression about 4dB (Cepstrum Distance) after processing only first iteration of the algorithm. This fact represents an improvement about 2dB (Cepstrum Distance) in relation to parameterized third-order algorithm (IF=1). Finally, the convergence of the iterative algorithms based on cumulant AR estimation is strongly accelerated. Therefore, a good reduction of computational complexity and processing delay are achieved, while no appreciable increase of distortion effect is generated. All these features are specially esteemed when low and medium SNR are considered.

References

Channel Noise Detection and Suppression in G721 decoded speech for CAI CT2 applications

Dimitris J. Hiotakakos, Costas S. Xydeas, Mark A. Ireton and Saf M. Asghar

Abstract. This paper describes a simple but effective burst error detection technique which is used to provide adequate noise suppression when the CCITT G721 32 kb/s digital speech codec is exposed to channel errors in a CAI-CT2 environment. An estimate of the energy of the decoded speech signal is determined using linear prediction, and the error between the actual and the predicted values is formed. If this error exceeds a pre-defined threshold, then a channel error burst is detected by the system and the decoder output is appropriately suppressed. This noise suppression algorithm has been adopted by Advanced Micro Devices (AMD) Inc. in their recently released controller chip Am79C410 for digital cordless telephones.

1.0 Introduction

With the ever increasing demand for personal speech communication systems, the CAI CT2 (Common Air Interface Cordless Telephone, 2nd generation) initiative has been directed at the provision of next generation cordless telephones. However, channel interference can result into unpleasant squawks in the recovered speech, particularly when the base station is relatively far from the handset and is temporarily obscured. Within this framework, we have designed an efficient burst error detection mechanism (BED) [1], which is used to provide adequate noise suppression when the CCITT G721 narrowband 32 kb/s digital speech codec standard [2] is exposed to channel errors. This noise suppression algorithm has been adopted by Advanced Micro Devices (AMD) Inc. in their recently released controller chip Am79C410 for digital cordless telephones [3]. The BED algorithm is incorporated within the device, among other operations performed by the controller such as: data formatting, audio processing and peripheral functions for CAI-CT2 cordless phones. These new generation CT2 cordless telephones employ the G721 ADPCM coding standard and are capable of serving both incoming and outgoing calls, within the vicinity of the base station, and outgoing calls only elsewhere [4]. Furthermore, channel noise conditions are relatively mild when compared to the Rician/Rayleigh fading conditions found in mobile phones. Due to its inherent robustness [5] the G721 codec performs well in the presence of random errors at a Bit Error Rate (BER) of $10^{-3}$ and performs acceptably well at a BER of $10^{-5}$. There are, however, occasions when bursts of channel errors occur, particularly when the base station becomes temporarily or permanently occluded from the cordless phone. Unfortunately such bursts create invariably a high energy “squawk” at the decoder output. Notice that it is possible to detect these events using the protected side data that is transmitted along with the speech data. This takes 32 or 64 ms on average, for a burst to be detected, depending upon whether 4 or 2 bit data signalling is being used. Following detection, the received signal is automatically muted in order to minimise the effects of unpleasant squawks. The BED scheme described in this paper operates directly on the decoded speech signal and detects with minimum delay the occurrence of these undesirable events. The method proposed is based on the fact that the energy of speech signals varies relatively slowly with time, following a trajectory which can be predicted fairly accurately.

2.0 Channel Characteristics

Numerous studies have been carried out in order to investigate the nature of the cordless channel [6,7]. In general, for a fixed radio link the error statistics of the channel can be described by a Gaussian random process where the probability density function (pdf) of the noise is given by:

$$p(n) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(n-\mu)^2}{2\sigma^2}}$$  \hfill (1)

where $\mu$ and $\sigma$ are the mean and variance of the noise process respectively. For a mobile link in the anticipated CAI environment the noise pdf is approximately modelled by the Rician distribution, defined as:

$$p(r) = \frac{r}{\sigma^2} e^{-\frac{(r^2+\mu^2)}{2\sigma^2}} I_0\left(\frac{\mu}{\sigma^2}\right)$$  \hfill (2)

where $\mu^2$ is the noncentrality parameter, $I_0(x)$ is the zero-th order modified Bessel function of the first kind and $\sigma$ is the standard deviation.

Given these two generalisations it is still not straightforward to model the CAI channel since the model will be greatly influenced by how the system is being used. In general, a pedestrian can be stationary when using this type of system, or have a conversation whilst moving. In each of these cases the “speed” of movement greatly affects the fade rate of the channel. Thus expressions are only theoretical approximations and an accurate CAI channel model which includes the effect of multiple reflections, scattering, movement of other objects in the vicinity of the mobile station and the base station,
and the possible motion of the mobile station, should be defined experimentally.

We have already pointed out that the G721 speech codec can cope reasonably well with Gaussian channel noise generating an error rate of the order of that envisaged in the CAI specification. Furthermore, such random errors are very difficult to detect from their effect on the decoded speech signal. We have therefore concentrated our efforts on the effect of bursts of errors. These have a more serious subjective effect on the recovered speech and can be detected and suppressed.

3.0 Preliminary Investigation

The G721 decoder appears to become unstable during error bursts thus creating undesirable high levels of output signal energy. During the course of our investigations our aim was to explore the behaviour of certain G721 parameters or signals associated with the decoder, first in the absence and then in the presence of error bursts, and thus to seek possible correlations between their values and the state of the channel. We have considered three different methods based on: i) the Predictor filter Stability, ii) the Slope of the quantiser step size, and iii) Energy variations in the reconstructed speech signal.

3.1 Predictor Filter Stability (PFS)

The predictor in the G721 codec is prevented from becoming unstable by "clipping" the filter parameters using appropriate thresholds (These tests are incorporated into the codec functions LMC and LMD-see [2]). We have therefore investigated whether channel error bursts would activate systematically the instability prevention mechanism. It was found however that this "criterion" does not correlate well with channel error events and PFS is not a useful BED indicator.

3.2 Quantiser Step-Size (QSS)

The rapid increase in signal energy which occurs at the output of the decoder, as a result of severe channel error bursts, is achieved by corresponding and rapid increments of the adaptive quantiser step size. The QSS rate of change (slope) was therefore tested as a possible BED indicator. However, although the step size slope becomes significant, soon after an error burst occurs in the channel, there are also instances where this QSS behaviour is adopted by the codec in order to track accurately certain rapid transitions in the input speech signal. This generates a relatively high rate of false burst detections which renders QSS as an inappropriate quantity for burst error detection.

4.0 Algorithm Description

The Burst Error Detection mechanism proposed here operates on G721 decoded speech samples \( \hat{X}(t) \) reconstructed by the decoder from the received bit stream. The method is based on the fact that the energy of a speech signal varies relatively slowly with time and, consequently, can be predicted fairly accurately, at clear channel conditions (noiseless transmission). During severe error burst conditions, however, the energy of the reconstructed speech is seen to rise very rapidly. Thus, if a prediction of the expected output energy is available, then the difference between the actual and predicted energy values will increase rapidly during a channel error burst. The BED algorithm is described as follows:

Consider that \( P_t \) is the current estimate of the output signal energy in dBs and \( \hat{P}_t \) is the predicted value of \( P_t \). In particular,

\[
P_t = 10 \log_{10} \frac{P_{\text{max}}}{P_{\text{max}}}
\]

where \( P_{\text{max}} = (2^{15})^2 \) is the maximum possible energy of a 13 bit signal, and

\[
P_t = \frac{S_t}{N}
\]

\( S_t \) is the signal energy calculated over a frame of \( N \) signal samples, i.e.

\[
S_t = \sum_{n=0}^{N-1} X^2(n)
\]

The predicted value \( \hat{P}_t \) of the estimate \( P_t \), is derived as the linear combination of past output signal energy values \( P_{t-m} \) and is given by:

\[
\hat{P}_t = \sum_{m=1}^{M} b_m P_{t-m}
\]

Given these estimates, a convenient measure for detecting the presence of a channel error burst is given by

\[
P_b = 10 \log_{10} \left( \frac{(P_b)^2}{(P_b - P_t)^2} \right)
\]

\( P_b \) can be thought as a signal to error ratio for the above linear prediction process, measured in dBs. In general, this measure will give high positive values except when an unexpected error burst occurs in the channel, in which case the prediction error is relatively large and \( E_b \) acquires low positive, or even negative values.

The generalised block diagram of the proposed BED mechanism is shown in Figure 5. If \( E_b \) exceeds a pre-defined threshold \( E_{\text{thr}} \) we mark this point in time as the start of a burst. The speech output (from that point onwards) is automatically scaled down by a value \( g_t \) (muting is applied when \( g_t = 0 \)) for a pre-defined period of time i.e. for \( k \) samples. The value of \( k = k_{\text{thr}} \) is determined experimentally according to the average expected burst length. Since \( E_b \) may exceed \( E_{\text{thr}} \) more than once, within the region of a certain burst, the \( k_{\text{thr}} \) samples interval is re-initialised every time the \( E_b < E_{\text{thr}} \) condition is satisfied. The interval is implemented in Fig 5 using an incremental counter \( k = k+1 \) and a \( k < k_{\text{thr}} \) comparison. When a new start of burst condition is detected, while \( k < k_{\text{thr}} \), the counter is simply reset to 0. The end of burst condition which restores the value of \( g_t \) to unity, occurs when \( k = k_{\text{thr}} \).

Notice that the first initialisation of the start of burst indicator (in the vicinity of the actual burst) and the corresponding \( g_t \) suppression, occurs with a delay of the order of 2 to 10 ms. This delay prohibits the system from suppressing completely the effect of channel error bursts from the decoded speech signal. However, this can be rectified by introducing an additional delay of \( M \) samples in the decoded speech signal (Fig 5).

5.0 Experimental Assessment

Investigations have shown that suitable values for \( N \) and \( E_{\text{thr}} \) are 10 and 8 dB respectively. Fig 1 shows the variation of \( E_b \) for a specimen speech segment of 3 secs. The fixed predictor coefficients \( b_m \) are pre-defined by a training process [8] from a set of \( E_b \) values derived by running the G721 codec over a large data base of male and female speech segments.

The BED algorithm has been tested under mild and severe channel conditions to ensure that it can provide an efficient detection mechanism under a wide range of channel noise conditions.
Specifically, the G721 encoded bit stream has been exposed to burst lengths of 64, 256 and 1024 bits (corresponding to 162, 64/8 and 256/32 samples/ch respectively). In addition, the algorithm has been tested with random errors, to ensure that the decoded signal is not adversely affected under random noise channel conditions.

\[ E_s (\text{dB}) \]

\[ Time (\text{sec}) \]

Figure 1. Plot of the \( E_s \) signal over derived over 3s of speech.

The objective and subjective results presented in this section are based on three different speech: one male, one female, and a conversation between a male and a female speaker (indicated by M, F and M+F respectively in Figs 2-4). The actual bursts detected are abbreviated as BD, and false detections are indicated as FD.

Fig 2 shows the number of channel error bursts that have been introduced into the bit stream of the above three speech files, for three channel burst error lengths. It also indicates how many of the introduced error bursts were judged in informal subjective listening tests to be serious, in the sense that they led to a noticeable click in the reproduced speech. During these listening tests two types of channel error bursts were counted according to their effect on the recovered speech. The most serious errors are labelled as "v. bad", and indicate serious loud clicks or squawks. Less serious errors are labelled as "poor" and indicate noticeable distortion which is not however serious enough to distract the listener from the conversation. Naturally the number of channel error bursts which lead to subjectively serious distortion increases, as the length of the error burst is increased.

Fig 3 shows the "objective" BED performance of the algorithm in terms of the number of times the mechanism has been activated, i.e. the number of times the \( E_s < E_{thr} \) condition is satisfied. The system detects the majority of error bursts but also provides a number of false positive indications. Notice however that all the perceptually damaging ("v. bad") error bursts of Fig 2 were detected in all the speech segments for all error burst lengths.

Fig 4 gives the number of channel error bursts which still have an effect on the decoded signal, after the application of the noise suppression algorithm with two different decoded signal delay values. It is clear that, when the algorithm introduces zero delay on the decoded speech, the subjective effect of the majority of severe error bursts, is eliminated. Furthermore, a delay of 120 samples (15 ms) enhances significantly the quality of the recovered speech since 85 to 100% of the error bursts are successfully suppressed (depending on the burst length). In addition, the number of false detections is very low.

Notice that the subjective performance of the burst error detection and suppression scheme has also been tested with random channel errors up to a maximum BER of \( 10^{-3} \). Results clearly indicate the ability of the scheme to improve under these random noise conditions the overall quality of the output decoded speech signal.
6.0 Conclusions
The error formed between the short-term energy value of the G721 decoded speech signal and its prediction estimate value has been used in this paper as a means of detecting bursts of channel errors within a CAI-CT2 application framework. The BFD algorithm presented is simple and yet effective in achieving a very high error burst detection rate, with a low count of false detection instances. Informal subjective tests at Manchester University and a real time hardware implementation at AMD, have shown the potential of the algorithm to provide the required channel noise suppression. As a result the algorithm has been incorporated in the Am79C41D device. Furthermore, the principle of using the speech energy signal trajectory to detect adverse channel conditions and resulting decoding instabilities, is general and can be used in conjunction with many medium and low bit rate coding systems.

Finally notice that the Fixed energy predictor employed in BED, can be combined with Adaptive energy prediction in order to reduce the system delay, without affecting the performance of the BED scheme.

References

Figure 5 The BED Algorithm
An Environment-Adaptive Noise Reduction Neural Network for Reliable Speech Recognition

Michael Trompf *, Harald Eckhardt * and Aliheb Mekhaiel **

* Alcatel SEL AG, Research Center, Lorenzstr. 10, 70435 Stuttgart, Germany
  email: mtrompf@rcs.sel.de, heckard@rcs.sel.de

** Universität Karlsruhe, Institut für Nachrichtensysteme, Kaiserstr. 12, 76128 Karlsruhe, Germany

Abstract. Voice input systems in telecommunication applications suffer from a decrease of recognition performance in instationary signal environment. Our approach to robust speech recognition consists of neural network-based reduction of the noise component in the feature vector domain. Based on our previous work with a noise reduction neural network (NRNN) trained offline [1], this current work is focused on the adaptation of the noise reduction mapping to instationary noise signals during speech pauses. This approach uses noise parameters as network input to control the mapping function. Furthermore, a comparison between the NRNN and spectral subtraction is made for an isolated word recognition task in noise-contaminated environment. Finally, a combination of spectral subtraction and the neural network-based noise reduction is investigated.

1. Introduction

Several publications about neural network-based approaches to noise reduction [2] [3] [4] refer to the ability of multilayer perceptrons (MLP) to approximate nonlinear functions when trained by error backpropagation (EBP) [5]. It can be considered as a nonlinear mapping from a space of noisy to a space of noise-free speech representation and can be realized in different processing domains. In general, a feature vector domain is preferred because of lower computational costs as compared to the time domain.

Our approach is similar to the one reported in [4] and lead to the development of a noise reduction neural network referred as NRNN [1]. In previous experiments we have shown that nonlinear mapping applied to the feature vector sequence outperforms a linear approach.

During training, the network learns an approximation to the unknown optimal noise reduction mapping while minimizing a mean squared error (MSE) criterion between the network output and the desired training target vector. The vector-based MSE over \( L \) noisy input and noise-reduced output vector pairs \((x, y)\) is given by

\[
MSE = \frac{1}{L} \sum_{k=0}^{L-1} \sum_{i=0}^{n} (y_{ki} - x_{ki})^2,
\]

with \( n \) denoting the dimensionality of the feature vectors. For a given a network topology, the \( MSE_T \) in the training set is minimized by weight modification with gradient descent. The \( MSE_V \) from a separate verification set is calculated for the generalisation test between successive training iterations.

Our multi-speaker database and the testbed for isolated word recognition were described in [1]. All experiments reported here were performed using the OFFICE multi-speaker database (30 words vocabulary, 5 repetitions of 30 words uttered by 10 speakers). After sampling with 8 kHz, 10 lpc-cepsrum coefficients are calculated every 20 ms from speech signal segments with a duration of 32 ms.

From this representation, the noise reduction experiments are performed. After the initial training using three recordings per speaker, the network maps the noisy test data on a frame-by-frame basis in subsequent forward-passes. The evaluation is done by means of the frame-based MSE as given in equation (1). In addition, the word accuracy is calculated by a DTW-based word recognizer from two recordings of each speaker and averaged over ten speakers, leading to a total of 600 test words.

The noise database includes printer, computer room and white noise recordings. They were added to the speech signal at different SNR-levels in 5 dB steps. Two types of experiments were performed: noise dependent tests refer to noise dependent training of the NRNN, and separate noise reduction and word recognition tests for each speaker and each noise signal. In noise-pooled experiments, NRNN training was done with all noise signals at the same time, but tested on each one separately. The MSE and recognition rates reported here were averaged over all speakers.

2. Feature-Based Neural Noise Reduction

If training and test data follow the same statistics, network training can be done offline in an initial training phase. In instationary noise signals or changing noise sources, a time-variant transfer function is required. Therefore concepts for continuous network adaptation become necessary, and the mapping function is time dependent.
2.1. Noise Reduction with Offline-Training

The design parameters of the NRNN described in [6] include temporal context from past and future signal frames in the input layer, cross validation as generalization test, and training on multiple SNR-ratios to enhance the robustness against variations of the signal energy. Best results were obtained from a fully connected three-layer perceptron with 50 input units according to the present and two past and future vectors with 10 coefficients each. The hidden and output layer contained 20 and 10 units, each of them with sigmoid activation function. After offline training on the OFFICE database using EBP, an average of 93.8% word recognition rate was obtained for speaker-dependent word recognition in 0 dB additive computer room noise [1].

2.2. Continuous Noise Adaptation

Continuous adaptation of the network transfer function can be done either by online modification of the neural model (e.g. [7] for chaotic time series and [9] for speaker adaptation) or by control information presented to designated input nodes. The latter approach was first described in [9] for the recognition of subword units and is modified for noise reduction in our current work. It is computationally inexpensive and allows for adaptation of the network transfer function according to speech pause parameters which were selected previously.

Based on noise parameter estimation during the adaptation intervals (e.g. speech pauses), the signal characteristics is described by a few parameters which are easy to recompute. For each speech segment, the parametric description for the noisy input signal is represented by a vector \( v = (x, n) \) composed of noisy speech coefficients in the subvector \( x \) and control parameters in \( n \). This concept is illustrated in figure 1. While the noise reduction mapping is performed continuously on the coefficients of subsequent feature vectors, the control parameters are updated from time to time, frozen, and presented to the control inputs of the network.

If \( F() \) denotes the network transfer function and \( w \) represents the weight vector, the output vector is given by

\[
y = F(w, n, x)
\]

After the initial training phase has been completed, the weights are fixed. Then the network output

\[
y = F_w(n, x)
\]

depends only on the input signal and the control inputs. During recognition, the noise reduction mapping of word \( k \) is modified by \( n_k \) and frozen afterwards, i.e.

\[
y_k = F_w(n_k, x_k)
\]

Hence, the adaptation of the mapping function is based on the sequence \( (n_1, n_2, \ldots, n_L) \) of subvectors calculated in the pauses before each word in the test data set.

Initial results with network adaptation were reported from the spectral moments and the SNR as control parameters in [10]. Improvements of a few percent in word recognition rate could be obtained when the test noise was a member of the training pool. Since the signal representation has not been optimized so far, additional investigations were made to select a good representation.

2.3. Noise Parameter Selection

The search for an appropriate parametric representation of the noise component was performed in four steps:

1. Different noise representations were extracted.
2. The best parameters were selected according to their normalized covariance with the error signal.
3. The NRNN was trained with the selected coefficients presented to the control inputs.
4. The residual MSE from the training set and the word recognition rate were used for evaluation.

Parameter extraction. The parameters investigated can be divided into three groups: signal energy related parameters such as the frame-based speech and noise energy with linear and logarithmic scales and the estimation of the SNR, spectral moments calculated from the short time energy spectra, and lpc-cepsrum coefficients. All parameters with only a few exceptions were calculated from the noise signal in the speech pauses and averaged over the pause segments. A total of 26 different parameters were extracted.

Parameter selection. The parameters were sorted according to their normalized covariance \( \sigma_{ie} \) with the absolute value of the error in the training and test sets. With \( \sigma_i^2 \), the variance of the i-th parameter as normalization factor, we obtain the regression coefficient \( r_i \) as defined in [11] by

\[
r_i = \sigma_{ie} / \sigma_i^2
\]

This normalization by \( \sigma_i^2 \) is necessary for the comparison of noise representations with different ranges of values. Table 1 shows a short description of the first 10 parameters sorted with respect to their value of \( r_i \). Among them were some of the 10 averaged lpc-cepsrum coefficients averaged over the pause signal as defined by
Table 1: Parameter selection according to r.

<table>
<thead>
<tr>
<th>no</th>
<th>symbol</th>
<th>description</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$E_W$</td>
<td>energy, word avg.</td>
<td>0.394</td>
</tr>
<tr>
<td>2</td>
<td>$C_{SP}$</td>
<td>ipc-cep, n=3, pause avg.</td>
<td>0.242</td>
</tr>
<tr>
<td>3</td>
<td>$C_{4P}$</td>
<td>ipc-cep, n=4, pause avg.</td>
<td>0.220</td>
</tr>
<tr>
<td>4</td>
<td>$M_{HP}$</td>
<td>spec. mom. high, p. avg.</td>
<td>0.218</td>
</tr>
<tr>
<td>5</td>
<td>$C_{4P}$</td>
<td>ipc-cep, n=1, pause avg.</td>
<td>0.218</td>
</tr>
<tr>
<td>6</td>
<td>$C_{6P}$</td>
<td>ipc-cep, n=6, pause avg.</td>
<td>0.208</td>
</tr>
<tr>
<td>7</td>
<td>$C_{SP}$</td>
<td>ipc-cep, n=5, pause avg.</td>
<td>0.204</td>
</tr>
<tr>
<td>8</td>
<td>$M_{LP}$</td>
<td>spec. mom. low, p. avg.</td>
<td>0.201</td>
</tr>
<tr>
<td>9</td>
<td>$C_{2P}$</td>
<td>ipc-cep, n=2, pause avg.</td>
<td>0.201</td>
</tr>
<tr>
<td>10</td>
<td>$M_{P}$</td>
<td>spec. mom., p. avg.</td>
<td>0.200</td>
</tr>
</tbody>
</table>

\[ c(n) = \frac{1}{P_k} \sum_{i=1}^{P_k} c_i(n) \quad \text{with} \quad (7) \]

\[ c_i(n) \quad \text{n-th cepstrum parameter} \]

\[ P_k \quad \text{number of segments in the k-th pause} \]

Details of ipc-cepstrum extraction can be found in [12]. Furthermore, the spectral moments described in [10] were among the first ten parameters. The coefficient with the highest $r_1$ in this list was the normalized energy of the succeeding word. Surprisingly, the SNR was not ranked among the first ten parameters.

Experimental Results. The following noise reduction experiments were performed:
1. noise dependent training without adaptation.
2. noise pooled training without adaptation.
3. noise pooled training with adaptation using the SNR- and spectral moment-parameters investigated in [10].
4. noise pooled training with adaptation using the 10 parameters from the table 1.

The MSE_T and the MSE_V values (not shown here) from the adaptive network were better than without adaptation when trained on the noise pool. As previously expected, noise dependent training resulted in lower MSE values as compared to noise-pooled training. The word recognition results are shown in figure 2. The pooled training with noise adaptation using the 10 parameters is denoted as pool+10 and yields better results than without adaptation. At some SNR-levels it performs even better than the noise dependent network. The reason might be the increased amount of training data for pooled training.

3. NN- and SS-based Noise Reduction

3.1. Noise Reduction by Spectral Subtraction

Spectral subtraction [13][14] uses the noise signal during speech pauses to get an estimate of the noise-free speech spectrum $S(l,i)$ from its noisy version $R(l,i)$, i.e.

\[ \hat{S}(l,i) = A(l,i) R(l,i) \quad , \quad (8) \]

with the frequency index $l$ and the segment number $i$. Two versions of SS with different system functions $A(l,i)$ were investigated. The original version, often referred as linear spectral subtraction (LSS), is described in [14]. At first, an estimate $\hat{N}(l,i)$ of the noise spectrum is calculated from subsequent short-time spectra $N(l,i)$ during the speech pauses. After weighting and subtraction from $R(l,i)$, an estimation of $S(l,i)$ is obtained. Improvements were achieved by modifying the transfer function $A(l,i)$ with an estimation factor $a$ and the spectral floor $c$ [13],[14] to

\[ A(l,i) = \begin{cases} 1 - \frac{\hat{N}(l,i)}{R(l,i)} & \text{for } A(l,i) > c \\ c & \text{else} \end{cases} \quad (9) \]

$a$ and $c$ were set to 1.25 and 0.1.

A more elaborated class of models are denoted as nonlinear spectral subtraction (NSS) [15]. In NSS, a SNR-dependent estimation factor and memorization of the noise spectrum maxima over past signal frames are applied. One representant of this class of system functions is given by

\[ A(l,i) = \begin{cases} \Phi\left(\frac{N_{\text{max}}(l,i), p(l,i)}{R(l,i)}\right) & \text{for } A(l,i) > c \\ c & \text{else} \end{cases} \quad (10) \]

The estimate of the noise spectrum in equ. (9) is replaced by $\Phi()$, which is a nonlinear function of the maxima in past noise spectra $N_{\text{max}}(l,i)$ and of $p()$, which represents a biased estimate of the SNR. For details see [15].
3.2. Comparison of SS and NRNN

A comparison of the NRNN-based noise reduction with LSS and NSS (figure 3) verifies, that the modifications in NSS lead to considerable improvements over LSS. The results show also, that the NRNN outperforms both SS-based approaches. This could be due to the nonlinear processing capabilities and the optimization of a global criterion during the NRNN-training in contrast to the more intuitive modelling of signal properties with NSS.

3.3 Combination of SS and NRNN

Encouraged by the results from speech pause adaptation, experiments were performed combining the advantages of the SS- and the NRNN-based approaches. At first, LSS was applied in the spectral domain. After feature extraction, the NRNN was trained to remove the residual error. The word recognition results in figure 4 show that particularly for low SNR levels better results were obtained. This was confirmed by lower residual MSE-values as compared to experiments with only one reduction step.

4. Summary

A new method for online-adaptation of the NRNN with control parameters was shown. We suggested a selection method for appropriate control parameters. An improvement of the recognition rate in noisy speech was achieved when the test noise was contained in the training pool. A comparison with two versions of SS yielded better results for the NRNN. The combination of SS for fast noise adaptation and the NRNN with nonlinear processing capability was successful. Future activities will include combinations of NSS and NRNNs as well as the investigation of alternative control parameters.

Acknowledgements. The authors wish to thank Prof. Dr.-Ing. K. Kroschel from the University of Karlsruhe for fruitful discussions about noise reduction. This work was sponsored by the German Ministry for Research and Technology (BMFT) under contract IT-01TBE-011/V102 L2.

References

Speech Enhancement Using Sub-Band Decomposition and Comparison with Full-Band Techniques

R. LE BOUQUIN-JEANNÈS, G. FAUCON, A. AKBARI AZIRANI, F. EHRMANN

Laboratoire de Traitement du Signal et de l’Image, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes cedex, FRANCE, Tel / Fax: (33) 99 28 69 19 / (33) 99 28 69 17, E-Mail: regine@lists.univ-rennes1.fr

Abstract. Noise reduction in degraded speech remains an important problem of signal processing in many fields of engineering. This paper deals with the enhancement of noisy speech signals for hands-free radiocommunications. It presents a two-channel technique based on a sub-band decomposition that takes the decorrelation of the disturbing noises into account. In this way, after decomposition, in the lower band the signal is filtered using short-time spectral amplitude estimation and in the upper-band a method based on the coherence function is carried out. In terms of results, sub-band and full-band techniques are compared according to their performance, constraints and robustness to non-stationary noise.

1. Introduction

This paper deals with the enhancement of noisy speech signals recorded by hands-free systems. There is actually a great interest in noise cancelling in a car with the development of mobile radio telephone technology. The objective of enhancement may vary and depend on the specific application. In our case, we focus on the problem of speech enhancement to the transmission to improve speech quality and get a signal more pleasant to listen to. Two microphones are placed in the vehicle and since the car environment is very restricted, we cannot fix the microphones in such a way that we form a reference input which is free of speech signal. In any case, each observation \( x_i \) we receive is composed of a speech signal \( s_i \) and an additive noise \( n_i ; \) we write the observations in the following form: \( x_i = s_i + n_i \) \( (i = 1, 2) \).

This paper introduces a noise reduction algorithm based on a sub-band decomposition and using the degree of decorrelation of the disturbing noises. In the low frequency band, techniques based on short-time spectral amplitude (STSA) estimation are used and in the upper band, the coherence function computed from the observations becomes a pertinent criterion to distinguish between useful signals and noises.

Section 2 is devoted to the noise and speech analysis. Then, the three noise reduction techniques we use are described in section 3. The fourth section presents our algorithm based on sub-band filtering. In terms of results, speech signals and noises have been recorded independently to conduct objective measures as well as informal listening tests that are detailed in section 5. Some concluding remarks are drawn by comparing the different methods, according to their performance, constraints and robustness to non-stationary noise.

2. Noise and signal characteristics

The aim of this analysis is to provide information to justify noise reduction algorithms that we can use in a car environment. Signals are recorded by two omnidirectional electret microphones, which are located on the ceiling of the car in front of the speaker and spaced by 73 cm. So, each microphone receives a noisy speech signal we write:

\[
\begin{align*}
    x_1(t) &= s_1(t) + n_1(t) \\
    x_2(t) &= s_2(t) + n_2(t),
\end{align*}
\]

and \( X_i(f,m), S_i(f,m), N_i(f,m) \) represent the spectral components of the signals \( x_i(t), s_i(t), n_i(t) \) in each frame \( m \); the sampling rate is 8 kHz. Firstly, we compute the power spectra of speech signals and noises using short time Fourier transforms and a Hamming window of 256 points long. These power spectra display some important noise components in the low frequencies. Since the telephone band is limited from 300 to 3400 Hz, we use a highpass FIR filter (256 taps) with a -3 dB cut-off frequency equal to 150 Hz, which enables to cancel a part of the disturbing noise without degrading the speech signal. Secondly, in order to evaluate the spatial correlation, we compute the coherence function from the two channels. This is a complex function of frequency and is defined as:

\[
\rho_{ac}(f) = \frac{E[U(f)V^*(f)]}{\sqrt{E[|U(f)|^2]E[|V(f)|^2]}}
\]
where \( U(f) \) and \( V(f) \) are the Fourier transforms of the signals \( u(t) \) and \( v(t) \) picked up at the same time by the two microphones. The coherence magnitude varies between 0 and 1 and gives, for each frequency, the percentage of signal energy coming from correlated sources. The coherence function computed from the speech signals only \( (s_1 \) and \( s_2) \) is close to 1 over all the frequency bandwidth while the coherence computed from the noises decreases quickly along with the frequency.

### 3. Noise reduction methods

In this section, we recall the three noise reduction methods used as full-band techniques or in the sub-band decomposition. The signal to be estimated is \( \hat{s}_1 \). Each method performs in the frequency domain using a 75% overlapping and the estimated signal is obtained in the time domain by inverse FFT and overlap-add.

- **The first method** is the method by Ephraim and Malah [1]. In this method, the clean speech estimator is obtained in two steps. First, a STSA estimator that minimizes the mean square error (MMSE) is computed. Then, it is derived under the uncertainty of speech signal presence in the noisy observations. By assuming the Fourier expansion coefficients of the speech and noise signals as statistically independent gaussian random variables, the MMSE STSA estimator is computed as:

\[
\hat{s}_1(f,m) = G_1(f,m)G_2(f,m)X_1(f,m) \tag{3}
\]

where \( G_1 \) is the gain function of the MMSE STSA estimator and \( G_2 \) represents the gain function due to the uncertainty of speech presence in the noisy observation \( X_1(f,m) \).

\[
G_1 = \frac{\sqrt{\text{E}}}{\sqrt{2}} \frac{\text{E}}{R_{ps}} \exp(\frac{V}{2}) \left[ (1 + V)I_0(\frac{V}{2}) + V I_1(\frac{V}{2}) \right] \tag{4}
\]

\[
I_0() \text{ and } I_1() \text{ denote the modified Bessel functions of zero and first order.} \quad V \text{ is a combination of } R_{ps}(f,m) \text{ and } R_{ps}(f,m): \tag{5}
\]

\[
V(f,m) = \frac{R_{ps}(f,m)}{1 + R_{ps}(f,m)} R_{ps}(f,m) \tag{5}
\]

\[
R_{ps}(f,m) = \frac{E[\{\hat{s}_1(f,m)\}^2]}{E[\{X_1(f,m)\}^2]} \quad R_{ps}(f,m) = \frac{X_1(f,m)}{E[\{N_1(f,m)\}^2]} \tag{6}
\]

where the expectation operator indicates a time averaging. \( R_{ps}(f,m) \) and \( R_{ps}(f,m) \) can be interpreted as the a priori and a posteriori signal-to-noise ratios (SNR) respectively.

\[
G_2(f,m) = \frac{A(f,m)}{1 + A(f,m)} \tag{7}
\]

\[
A(f,m) = \frac{1 - q(f,m)}{q(f,m)} \left( \frac{E[\{V(f,m)\}]}{1 + R_{ps}(f,m)} \right) \tag{8}
\]

\( q(f,m) \) is the probability of signal absence.

In practice, \( R_{ps}(f,m) \) can be computed using the Fourier transform of the noisy speech \( X_1(f,m) \) and the noise power spectrum \( P_n(f) \) estimated during speech pauses

\[
P_n(f) = \frac{1}{K} \sum_{k=1}^{K} [N_1(f,k)]^2 \tag{9}
\]

\( R_{ps}(f,m) \) can be computed from the amplitude estimator of the \((m-1)^{th}\) frame, \( P_n(f) \) and the a posteriori SNR in the \(m^{th}\) frame

\[
R_{ps}(f,m) = \lambda \left( \frac{\hat{s}_1(f,m-1)}{P_n(f)} \right)^2 + (1 - \lambda) p[R_{ps}(f,m-1)] \tag{10}
\]

where the parameter \( \lambda \) is a forgetting factor and the operator \( p[\cdot] \) is defined by:

\[
p[x] = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{11}
\]

- **The second method** takes advantage of the two channels to improve a mono-channel spectral subtraction technique. This technique, called PSI (Preprocessing + Signal Identification) performs first a noise reduction on each channel and then an identification. The three steps of this method are as follows:

  - on each microphone \((x_1 \text{ and } x_2)\) a noise reduction is performed using the magnitude spectral subtraction: on silent periods, we estimate the noise power spectra \( P_n(f), i = 1, 2 \) as in eq. (9). For each block \( m \) containing the useful signal, we compute [2]:

\[
D_1(f,m) = |X_1(f,m)| - \alpha \beta^{1/2}(f) \tag{12}
\]

then

\[
X_1'(f,m) = \begin{cases} D_1(f,m) & \text{if } D_1(f,m) \geq (\beta P_n(f))^{1/2} \\ \alpha P_n(f) \end{cases} \tag{12}
\]

The spectrum of each processed input is obtained using the phase of the noisy observation. The new observations \( x'_1 \) and \( x'_2 \) are estimates of \( s_1 \) and \( s_2 \) respectively and have a greater SNR than \( x_1 \) and \( x_2 \).

- the second step is an identification step. The supplementary information provided by the second channel preprocessed, \( x'_2 \), is used to give another estimate of \( s_1 \). \( x'_2 \) is applied to a filter \( H(f,m) \) to obtain a second estimate \( x''_1 \) of \( s_1 \):

\[
H(f,m) = \lambda H(f,m-1) + (1 - \lambda) \frac{X_1'(f,m)}{X'_2(f,m)} \tag{12}
\]

where \( \lambda \) is a forgetting factor.
the final spectral estimate is computed as:

\[
\hat{S}_1(f,m) = \frac{1}{2} \left( X_1^*(f,m) + X_1(f,m) \right).
\]  

(15)

- **The third method** is based on the coherence function. It assumes the speech signals are strongly correlated while the noises are spatially uncorrelated. By assuming \( \rho_{n_1 n_2}(f,m) = 0 \) for each block \( m \), the coherence function between the observations is:

\[
\rho_{n_1 n_2}(f,m) = \frac{\gamma_{\rho_{n_1 n_2}}(f,m)}{\sqrt{\left(\gamma_{\rho_{n_1 n_2}}(f,m) + \gamma_{n_1 n_2}(f,m)\right) \left(\gamma_{n_1 n_2}(f,m) + \gamma_{n_1 n_2}(f,m)\right)}}
\]

where \( \gamma_{uv}(f,m) \) is the cross spectral density between \( u \) and \( v \). The objective of the coherence function consists in turning off uncorrelated signals and passing correlated signals. To this end, we choose the following criterion:

- if \( \left| \rho_{n_1 n_2}(f,m) \right| \) tends to 0 \( \left( \left| \rho_{n_1 n_2}(f,m) \right| < T_{\text{min}}, T_{\text{min}} : \text{minimal threshold} \right) \), we assume there are only noises;
- if \( \left| \rho_{n_1 n_2}(f,m) \right| \) tends to 1 \( \left( \left| \rho_{n_1 n_2}(f,m) \right| > T_{\text{max}}, T_{\text{max}} : \text{maximal threshold} \right) \), the speech signal is predominant and must pass without any distortion; for intermediate values of \( \left| \rho_{n_1 n_2}(f,m) \right| \), there are speech and noise, and the coherence function is used to filter \( x_1 \), according to the following algorithm:

\[
\begin{align*}
\hat{S}_1(f,m) &= \left(\text{MSC}(f,m)\right)^2 X_1(f,m), \quad T_{\text{min}} < \left| \rho_{n_1 n_2}(f,m) \right| < T_{\text{max}} \\
\hat{S}_1(f,m) &= X_1(f,m), \quad \left| \rho_{n_1 n_2}(f,m) \right| \geq T_{\text{max}} \\
\hat{S}_1(f,m) &= \left(\text{T}_{\text{min}}\right)^2 X_1(f,m), \quad \left| \rho_{n_1 n_2}(f,m) \right| \leq T_{\text{min}}.
\end{align*}
\]

(17)

where MSC(.) represents the Magnitude Squared Coherence defined as:

\[
\text{MSC}(f,m) = \left| \rho_{n_1 n_2}(f,m) \right|^2 = \frac{\left| \gamma_{n_1 n_2}(f,m) \right|^2}{\gamma_{n_1 n_2}(f,m) \gamma_{n_2 n_2}(f,m)}.
\]

(18)

The spectral densities used in the evaluation of the coherence function are estimated using a time averaging by a simple recursive calculation [3]:

\[
\gamma_{n_1 n_2}(f,m) = \lambda \gamma_{n_1 n_2}(f,m-1) + (1 - \lambda) X_1(f,m)X_1^*(f,m), \quad (i,j = 1, 2).
\]

(19)

4. **Algorithm based on sub-band filtering**

The main idea consists in using two different techniques, the first one in the low frequencies (< 1kHz), the other one in the high frequencies (1kHz - 4kHz). The sub-band decomposition uses a CQF filter, that splits the signal into 4 sub-bands each of 1kHz, from 0 to 4kHz. Then, we consider two bands, the first one including frequencies from 0 to 1kHz, the second one including frequencies from 1 to 4kHz, and in each sub-band, we apply a noise reduction method.

In the low band, the speech signals are highly correlated, and the noises are rather correlated too. So, it is possible to discriminate between useful signals and noises with either a priori knowledge of the transfer function between signals (or between noises) or a priori knowledge of the noise characteristics. As a consequence, we investigate two methods derived from the STSA estimator. Nevertheless, these techniques require the hypothesis of noise stationarity.

The first one is the method by Ephraim and Malah applied on the observation \( x_1 \). The second one is the PSI method that takes advantage of the two channels to improve the mono-channel approach.

On the contrary, in the high frequency band, speech signals remain correlated while noises are slightly correlated or decorrelated. The method based on the coherence function, described in section 3, is all the more suited as it does not require any learning of noise characteristics and takes the noise non-stationarity into account.

5. **Results**

The methods using sub-band decomposition (composite methods) and the full-band methods have been compared. To conduct objective measures, speech signals and noises have been recorded independently in real conditions, in a car moving at 130km/h and 90km/h, and the corresponding SNR for these stationary noises are equal to 0dB and 3dB respectively. For the non-stationary noise condition, we use the stationary noise at 130km/h weighted by a variable coefficient \( \delta \) (Figure 1) to simulate the non-stationarity; the averaged SNR is close to 0dB. The speech signals are composed of two sentences separated by a silent period.

![Figure 1. Variation of the noise coefficient \( \delta \).](image)

Each of the three noise reduction methods has a few parameters to adjust to optimize the results; they are as follows: for the method by Ephraim and Malah, the adjustable parameters are the forgetting factor \( \lambda \) of the a priori SNR and the speech absence probability \( q(f,m) \) which are respectively equal to \( \lambda = 0.99 \) and \( q(f,m) = 0.5 \); for the PSI method, these are the overestimation factor \( \alpha \), the spectral floor \( \beta \) and the forgetting factor \( \lambda \), respectively equal to \( \alpha = 1.5, \beta = 0.01 \) and \( \lambda = 0.7 \); for the method based on the coherence function, the adjustable parameters are the MSC power \( \alpha \), the minimal threshold \( T_{\text{min}} \), the maximal threshold \( T_{\text{max}} \) and the forgetting factor \( \lambda \). They have been set to \( \alpha = 4, T_{\text{min}} = 0.4, T_{\text{max}} = 0.99 \) and \( \lambda = 0.7 \) respectively.
We compute three objective measures, the gain $G$ on the segmental SNR, the cepstral distance $d_{cep}$ and a combined measure $M_c$ [4] including an inverse noise reduction factor and a distortion rate. We have also conducted informal listening tests to appreciate the quality of the resulting files. Fifteen listeners have been asked to judge these files in terms of distortion, residual noise and defaults brought by the processing. The different results are recorded three times at random. Each listener gives for each file a note between 1 and 5 points according to the following quality scale:

<table>
<thead>
<tr>
<th>Score</th>
<th>Quality of the Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>bad</td>
</tr>
<tr>
<td>2</td>
<td>poor</td>
</tr>
<tr>
<td>3</td>
<td>fair</td>
</tr>
<tr>
<td>4</td>
<td>good</td>
</tr>
<tr>
<td>5</td>
<td>excellent</td>
</tr>
</tbody>
</table>

Tables 1., 2. and 3. present, for each method, the three objective measures and the score issued from the informal listening tests.

<table>
<thead>
<tr>
<th>Method</th>
<th>Gain (dB)</th>
<th>$d_{cep}$</th>
<th>$M_c$</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSI</td>
<td>16.047</td>
<td>0.2572</td>
<td>0.0068</td>
<td>3.57</td>
</tr>
<tr>
<td>EM</td>
<td>18.574</td>
<td>0.3218</td>
<td>0.0084</td>
<td>3.15</td>
</tr>
<tr>
<td>Coherence</td>
<td>17.415</td>
<td>0.4687</td>
<td>0.0093</td>
<td>2.62</td>
</tr>
<tr>
<td>Composite PSI</td>
<td>16.138</td>
<td>0.2653</td>
<td>0.0039</td>
<td>3.2</td>
</tr>
<tr>
<td>Composite EM</td>
<td>18.328</td>
<td>0.2677</td>
<td>0.0080</td>
<td>3.54</td>
</tr>
</tbody>
</table>

Table 1. Stationary car noise, car moving at 130 km/h SNR = 0 dB.

<table>
<thead>
<tr>
<th>Method</th>
<th>Gain (dB)</th>
<th>$d_{cep}$</th>
<th>$M_c$</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSI</td>
<td>15.396</td>
<td>0.2879</td>
<td>0.0040</td>
<td>3.42</td>
</tr>
<tr>
<td>EM</td>
<td>17.040</td>
<td>0.3621</td>
<td>0.0149</td>
<td>3.4</td>
</tr>
<tr>
<td>Coherence</td>
<td>16.096</td>
<td>0.2652</td>
<td>0.0142</td>
<td>2.82</td>
</tr>
<tr>
<td>Composite PSI</td>
<td>13.556</td>
<td>0.2705</td>
<td>0.0044</td>
<td>2.99</td>
</tr>
<tr>
<td>Composite EM</td>
<td>16.341</td>
<td>0.2925</td>
<td>0.0144</td>
<td>3.15</td>
</tr>
</tbody>
</table>

Table 2. Stationary car noise, car moving at 90 km/h SNR = 3 dB.

<table>
<thead>
<tr>
<th>Method</th>
<th>Gain (dB)</th>
<th>$d_{cep}$</th>
<th>$M_c$</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSI</td>
<td>17.170</td>
<td>0.4901</td>
<td>0.0056</td>
<td>1.74</td>
</tr>
<tr>
<td>EM</td>
<td>16.423</td>
<td>0.5287</td>
<td>0.0203</td>
<td>1.22</td>
</tr>
<tr>
<td>Coherence</td>
<td>20.593</td>
<td>0.5978</td>
<td>0.0100</td>
<td>2.84</td>
</tr>
<tr>
<td>Composite PSI</td>
<td>17.335</td>
<td>0.8400</td>
<td>0.0081</td>
<td>1.98</td>
</tr>
<tr>
<td>Composite EM</td>
<td>16.547</td>
<td>1.0615</td>
<td>0.0251</td>
<td>1.79</td>
</tr>
</tbody>
</table>

Table 3. Non-stationary car noise.

The non-coherence between some of the objective results makes them difficult to exploit. Nevertheless, we can acknowledge that the method based on the coherence function becomes attractive in non-stationary environments. Concerning the listening tests, they appear more revealing. Even if they are not quite rigorous, they bring some information in the cases of Tables 1. and 3., the difference between the extreme scores reaching 1 point. For stationary car noise with a SNR = 0 dB, the best methods are the PSI and the composite EM methods.

For the non-stationary car noise, the method based on the coherence function is successful while the other full-band methods fail. It seems that the identification step used in the PSI method allows to improve the mono-channel approach. The non-stationary noise we simulated is very particular and it would be interesting to test these methods on real non-stationary noises.

6. Conclusion

In this paper, we proposed two noise cancelling methods using decomposition into two sub-bands: 0-1 kHz, 1-4 kHz. In the low frequencies, we perform a method derived from the STSA estimator, either a mono-channel one (Epfrain and Malah), or a bi-channel one (PSI method) while in the high frequencies, we apply a method based on the coherence function. These methods are compared to full-band techniques for three kinds of noise. The method using the coherence between observations appears attractive in non-stationary environments since it requires no learning of noise characteristics. Nevertheless, this method cannot cancel correlated noises and so, in the low frequency band, a method based on noise characteristics is used. For the situations we consider, methods using sub-band decomposition seem to be robust to the different noises. Extensive tests must be performed on real non-stationary noises.

References


The work reported here has been performed within an ESPRIT project: P6166 (REETEL : Enhancement of Hands-Free Telecommunications) sponsored by the CEC. The authors would like to thank MATRA-COMMUNICATION (Bois d’Arcy) for the availability of their database.
Adaptive Speech Enhancement with Diverse Sub-band Processing

Douglas R CAMPBELL

Department of Electrical & Electronic Engineering, University of Paisley, High Street, Paisley PA1 2BE, Scotland, UK,
Tel/Fax: +44 [41] 848 3410/3404, Email: d.r.campbell@uk.ac.paisley.cs

Abstract. An approach to speech enhancement that combines multi-sensor and sub-band adaptive methods is presented. Comparative results achieved in simulations and real acoustic experiments demonstrate that the method is capable of outperforming conventional noise cancellation.

1. Introduction

Background noise contamination of speech signals reduces the signal-to-noise ratio (SNR) of e.g. hands-free telephones, portable phones, and security screens. Speech recognition systems in particular experience problems due to levels of background noise considered quite acceptable to human listeners [1]. In addition to the noise level, the presence of multiple noise sources, reverberant environments, moving noise sources and statistically non-stationary noise sources considerably complicates the situation.

2. DSP Speech Enhancement

Methods based on noise cancellation implementations which attempt to model acoustic path transfer functions can produce excellent results in anechoic environments with localised sound radiators [2], however performance deteriorates in reverberant environments.

Multi-band processing has been found to be important when reverberation effects are significant [3, 4]. Adaptation is necessary to compensate for changing noise fields [5] e.g. statistical non-stationarity of the sources, source or sensor motion, or time varying acoustic paths. Multi-sensor methods are necessary to compensate for reverberation [6, 7] and speech/noise spectral overlap [8].

Studies on noise in automobiles [4] found that the correlation between signals from two microphones was high for frequencies below 500Hz and decreased gradually with virtually no correlation above 2kHz. The least mean square (LMS) correlated noise canceller [9] was applied to such signals by Wallace and Goubran [5]. At low frequencies noise reduction was obtained but at high frequencies, where both the correlation and noise energy were low, noise was increased. Measurements presented here (Figures 1 and 2) from office and automobile environments indicate that the peak correlation coefficient can vary significantly across frequency bands and is dependent on the location of the microphones. This supports a general approach involving diverse processing in frequency bands dependent on the correlation between the in-band signals from multiple sensors.

![Figure 1. Correlation coefficient vs frequency](image-url)
3. Multi-band Multisensor Intermittent Adaptive Scheme

Chazan et al [10] and Van Compernolle et al [11] used closely spaced microphones in an adaptive noise cancellation scheme involving the identification of a differential acoustic path transfer function during a "noise only" period in intermittent speech. Toner and Campbell [12] proposed an extension of this approach that allows noise features within sub-bands, such as the correlation between signals from multiple sensors to influence the subsequent processing.

For example, if the noise signals in a sub-band are highly correlated, an adaptive filter may be determined which models the differential acoustic path transfer function between the sensors. This can then be used in a noise cancellation format during the "speech plus noise" period to process the noisy speech signal. This scheme illustrated in Figure 3 can be described mathematically in terms of Z transformed functions as follows. The primary (P) and reference (R) signals in each sub-band are,

\[ P = F(S + H_1N) \]
\[ R = F(S + H_2N) \]

The transformed error signal is thus,

\[ E = F(1 - H_1S + (H_1 - H_1H_2)N) \]

which is a frequency domain error, weighted by the bandlimiting transfer function F. The mean-square-error (mse) function is

\[ J_\text{mse} = (2\pi)^{-1} |F|^2 |E|^2 |z|^{-1} \text{d}z \]
\[ |z| = 1 \]

Figure 3. 2-microphone processing scheme

The narrow-band noise cancellation problem is to find an \( H_3 \) such that within the sub-band defined by \( F \) the variance of \( J_\text{mse} \) is minimised. During a noise only period,

\[ J_\text{mse} = (2\pi)^{-1} |F(H_1 - H_2H_3)F^{-1}(H_1 - H_1H_3)F^{-1}E| \]
\[ |z| = 1 \]

which is minimised in the least squares sense when,

\[ H_3 = (F^H)(F^H)^{-1} \]

That is, \( H_3 \) is a bandlimited transfer function which minimises the noise power in \( E \). Using \( H_2 \) as a fixed filtering process when speech and noise are present ideally yields,

\[ E = F(1 - H_2)S \]

where the output \( E \) is a noise reduced, filtered version of the sub-band portion of the speech signal. This approach will obviously fail if \( H_1 = H_2 \), however in practical situations it is often possible to arrange the sensor placement to avoid this balance of acoustic paths.

Alternatively, if the noise signals in a band exhibit a low correlation then the processing method used by Ferrara and Widrow [13] or Zellinski [14] may be applied during the "noisy speech" period. Since in this case the primary signal noise component \( FH_2N \) is uncorrelated with the reference signal noise component \( FH_2N \), the filtered reference is an estimate of the speech signal \( S \).

If the noise power in a band is below, or the SNR above, some arbitrary threshold, then the signal in that band need not be modified.

4. Experiment - Simulated Environment

Initial testing of the method was by simulation of anechoic and echoic environments using an acoustic image method and verification of the experimental method in an anechoic chamber. Signals were synthesised with a peak
cross-correlation that varied significantly with frequency as shown in Figure 4. The data was then processed by conventional time domain LMS (CLMS), Ferrara-Widrow (FW), multi-band LMS (MBLMS) and multi-band with correlation measure (MBCM) algorithms, the latter using a mix of FW and MBLMS as dictated by the correlation of the noise signals in each sub-band. The comparative results of Figure 5 show that MBCM can significantly improve the SNR over the other methods.

![Figure 4. Correlation coefficient vs frequency: simulated data](image)

![Figure 5. SNR comparison for simulated data](image)

5. Experiment - Office Environment

Tests were then performed using data gathered in a real office environment of dimension (6 × 5 × 4)m containing desks, cabinets and computer systems. Two omni-directional microphones were placed centrally within the room spaced about 40cm apart, co-linear with and 1m distant from a loudspeaker driven by a white noise generator. The microphone signals were pre-amplified, passed through 4kHz cut-off, 5th order, elliptic anti-aliasing filters, sampled at 10kHz and converted to 12 bit word length. An adaptive filter length of 256 was found to be adequate to demonstrate the behaviour of the algorithms. In an attempt to balance computational complexity the M band MBLMS used a filter length of 256/M in each sub-band.

Toner and Campbell [12] found the convergence rate of the MBLMS method to be significantly improved over that for the well known methods of CLMS and frequency domain LMS (FDLMS) [15]. Informal listening showed the MBLMS processed speech to be both enhanced in SNR and of better perceived quality than that obtained by the other methods.

This data has since been processed using sub-band processing dependent on the correlation of the noise signals in the sub-band, i.e. MBCM. The results presented in Figure 6 and Figure 7 compare SNR achieved by the MBCM method against CLMS for different microphone spacings and orientations, and illustrate that useful SNR improvements can be obtained using the new MBCM method. Again informal listening tests confirm the SNR improvement. These results were obtained using parameters thought to be reasonable, but optimisation of the performance remains to be investigated.

![Figure 6. SNR vs MTM spacing: θ = 90°, NTM = 0.5m](image)

![Figure 7. SNR vs MTM spacing: θ = 45°, NTM = 0.5m](image)
6. Conclusion

In theory the multi-band approach using closely spaced microphones has several advantages. Operating in sub-bands allows an increased speed of adaption through the freedom to have different adaption step sizes in each band. Since the individual adaptive filters in each band will converge to narrow-band functions, faster adaption can be obtained by starting the coefficients at a suitable narrow-band set of values. As shown here, advantages can be gained by making separate decisions on the form processing should take in each sub-band. The inherent parallelism of the approach allows distribution across parallel processors of the computational load, which may be further reduced by using sub-sampling techniques and careful filterbank design to avoid aliasing problems.

Using closely spaced microphones reduces both the number of coefficients required of the adaptive filter and the sub-band computational load, both when adapting and when performing fixed filtering. It also reduces the misadjustment noise when used in the continuously adapting FW scheme.

7. Acknowledgement

The author is pleased to acknowledge the contribution of his research student E Toner and helpful discussions with D Van Compernolle of K U Leuven, Belgium and K Linhard of Daimler-Benz AG, Ulm, Germany.

References


Very Low Bit-rate Colour Video Coding Scheme Using Sub-band Motion Estimation/Compensation

Kwong H. GOH, John J. SORAGHAN and Tariq S. DURRANI

Signal Processing Division, Dept. of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, Scotland, U.K.,
Tel/Fax: +44 [0] 552 4400 / 552 2487, E-mail: goh@uk.ac.strathclyde.eee.spd

Abstract: Novel Wavelet Transform based algorithms for coding still images and image sequences are presented. The still image codec uses a new efficient adaptive bit-plane run-length coding of the Wavelet Transform coefficients of images. The main merit of this coding scheme is its simplicity requiring no training or storage of codebooks. A comparative performance between this new codec and the JPEG codec is given for the colour 'lena' image. For image sequence coding, a multi-resolution motion estimation/compensation (MRMEC) codec designed for very low bit-rate colour video transmission applications is presented. A simple buffer-rate control algorithm is designed to maintain acceptably good visual quality when transmitted at low rates. The codec is simulated to transmit at 19.2 Kbit/s and 9.6 Kbit/s respectively. Reconstruction images at 9.6 Kbit/s are shown.

1. Introduction:
The Discrete Wavelet Transform (DWT) has received considerable attention in image compression applications. In 2-d DWT, an image can be interpreted as a sum of details which appear at different resolutions. Due to the nature of its global decomposition, the 'blocking effect' coding error which would be found in conventional block-DCT coding schemes is replaced by a globally distributed error. The visual effect of the error made in WT based coding scheme depends very much on the actual coding of the WT coefficients. In this paper, new WT-based algorithms for low bit-rate still images and video coding are presented.

In still image coding, various methods of coding the Wavelet transform coefficients have been reported. Applying vector quantization to the WT coefficients was reported in [1,2]. A sophisticated embedded wavelet hierarchical image coder was developed in [3], and in [4], a tree-structured prediction algorithm was reported. In section 2, a simple and effective novel adaptive coding scheme for still images is presented. It requires no training, which is needed in the VQ approach [1,2], and no storage of codebooks which is required in both the VQ and the conventional run-length coder [5]. This codec makes use of the properties of the Human Visual System (HVS) and a simple adaptive bit-plane run-length codec (ABPRLC) to effectively code the wavelet transform coefficients. The new ABPRLC is described. Visual and compression comparisons with the JPEG standard are given for the colour 'lena' image.

For image sequence coding, various multi-resolution video coding schemes has been reported. Using the 3-d DWT to code image sequences was reported in [6,7]. However, applying 3-d WT on image sequences causes long transmission delay in real time low bit-rate video transmission applications. The multi-resolution motion estimation/compensation algorithms described in [8] were mainly applied to high bit-rate video coding. In section 3 of this paper, a novel multi-resolution motion estimation/compensation coding scheme specially designed for very low bit-rate video transmission is presented. This coding scheme utilizes the motion vectors in the lower resolution sub-bands for higher resolution sub-bands and hence reduces the information to be coded. The new coding scheme incorporates an output transmission buffer and a simple buffer fullness control algorithm to achieve constant bit-rate transmission as well as to maintain good picture quality. Coding results at various low bit-rates (9.6 and 19.2 Kbit/s) are presented.

2. WT-based Still Image Codec:
In a 2-d DWT, separable high-pass and low-pass filters are used to decompose the image in the horizontal and vertical directions respectively. The wavelet filters for the space domain decomposition use a pair of symmetrical perfect reconstruction filters[2] (lowpass: -0.05, 0.25, 0.6, 0.25, -0.05; highpass: -3/280, 3/56, 73/280, -17/28, 73/280, 3/56, -3/280). In this scheme, the image is decomposed into 5 resolution levels to produce 16 sub-bands. A uniform quantizer with a dead band around zero is used to quantize the WT coefficients, where the stepsize is varied according to the frequency level. It is known that the HVS has an imperfect response to high spatial frequency brightness transitions. Also, the HVS is more perceptible to errors in lower frequencies than those incurred in higher bands, e.g. errors in smooth areas are more disturbing to a viewer than those near edges. Hence the intensity at high spatial frequency is less important and therefore in quantizing the WT coefficients, larger quantization steps are used at higher frequency sub-bands and smaller quantization steps are used at lower frequency sub-bands. Quantizing the WT coefficients produces a large number of zeros and small magnitude coefficients. The simplest way to code this type of data will naturally be the run-length coding which is used in the JPEG codec [5]. However unlike the JPEG where the variable length code (VLC) tables used are specially customised for the DCT coefficients in fixed block size, VLC tables for the WT sub-band coefficients will have to suit both the different block sizes and the different sub-band's statistical characteristics so to accomplish optimal coding. A new lossless adaptive bit-plane run-length codec (ABPRLC) is now developed which can be used for all the sub-bands with any block size, which adapts to different ranges of entropy conditions. Each quantized sub-band is then
coded individually using this ABPRLC. The coefficients of each sub-band are considered as a set of eight 1-bit planes, each of which is run-length encoded. Ten sets of variable length code tables (VLCT, n=0,1,...9) are designed, each of which is suitable for different run-lengths of zero. VLCT 0 is designed for small run of zeros whereas VLCT 9 is for large run of zeros. The VLCT is designed in such a way that no code tables are actually required to be stored in the coder or decoder. The idea is simple and is described as follows: Only consecutive run of '0's are coded with a VLC and run of '1's will remain as it is. All the codes in VLCT 0-9 are prefixed with a start code, either '00' or '01'. The start code '00' represents run of zeros of a length less than or equal to N, whereas '01' represents run length of greater than N, where N is a predefined number for each VLCT. A run of r zeros, for r ≤ N is represented by '00' followed by (r-1) '0's and a '1'. If r > N, then the run is represented by '01' followed by INT((r-N-1)/(2^N-1)) × N + 1's where INT[,] represents the rounded down integer value, which is then followed by the N-bit binary equivalent of the remainder of the division [(r-N-1)/(2^N-1)]. N is set to 3 (suitable for coding small run lengths of zero) for VLCT 0, and is increased by 1 for each of the next VLCT to N=12 (for coding large run lengths of zero) for VLCT 9. Therefore, a VLCT with a predefined N will be suitable for coding a bit-plane that has lots of its zero-run-length in between (2^(N-1)+N-1) and (2^N+N-1) inclusive. Details of the code tables can be found in [9]. The following example illustrates the run-length coding scheme. Assume using VLCT 2, N=3 (optimal automatic selection of n will be described later), then the bit sequence:

```
00000000000000000000000000000001 000111000000000001
```

is coded as: 0111100 1 000011 0100011 1

This scheme can be used for any block size and the range of N is suited for a wide range of entropy conditions. It should be noted that using this scheme, if N or the VLCT number is known, then the VLC can be generated without any lookup of pre-stored tables. Also calculation of number of bits required for a given value of N is straightforward.

The selection of appropriate VLCT for each bit-plane is data dependent. Since the quantized coefficients are mostly small, the higher significant bit-planes will have very large run of zeros whereas lesser significant bit-planes and the sign bit-plane will have smaller run of zeros. When coding the coefficients, the best suited VLCT is chosen for each bit-plane, with a small overhead required to indicate to the receiver which VLCT (or value of N) is used for the particular bit-plane. Let NUMBIT(n) be the number of bits required to code a particular bit-plane using VLCT n, the algorithm to select n is described as follows:

For each of the 3 least significant bit-planes and the sign bit-plane,
{n=0}
while [NUMBIT(n+1) < NUMBIT(n)]
do {n=n+1}
end
For each of the 4 most significant bit-planes,
{n=9}
while [NUMBIT(n-1) < NUMBIT(n)]
do {n=n-1}
end
VLCT n will then be used for each particular bit-plane.

2.1 Coding Results:
The new coding scheme was tested on colour 'lena' image of size 512x512 pixels in all the Y, U and V components. The performance of the WT based codec using the ABPRLC is compared to the JPEG. At compression ratio of 50:1, both the WT-ABPRLC and JPEG give subjectively good quality reconstructed images as the noise is unnoticeable. At a higher compression ratio of around 100:1, the shape and edge details of the WT-ABPRLC coded image are still preserved and give quite acceptable quality. In the JPEG reconstructed image, 'blocking effect' is noticeable. When the compression ratio is increased to 178:1, the WT-ABPRLC still gives a reasonably acceptable image (fig.1a), despite the noise that is concentrated around the edges. At compression of 166:1, the JPEG reconstruction (fig.1b) gives a very blocky image. It is worth noting that the colours in the JPEG reconstruction are totally different from those in the original.

Figure 1 (a) and (b): (top and bottom respectively)
(a) WT-ABPRLC at 178:1,
Y, U and V PSNR = 27.23, 35.49 and 34.65 dB respectively
(b) JPEG at 166:1,
Y, U and V PSNR = 24.49, 26.96 and 27.54 dB respectively
3. Multi-resolution Coding Scheme:
Section 2 indicated that the WT performed favorably to the JPEG for low bit-rate still image compression. We now investigate the use of the WT for low bit-rate video coding. In [7] the 3-D DWT was used for high compression of head and shoulder image sequences. However, it is not very suitable for real-time transmission applications due to its transmission delay depending on the frame size used. Several multi-resolution motion compensation schemes were reported in [8] which utilized the motion vectors in the lower resolution sub-bands. The new multi-resolution motion estimation/compensation codec described below also utilizes the lowest resolution sub-band motion information and is designed for very low bit-rate transmission.

In very low bit-rate (4.8-19.2 Kbit/s) video transmission application such as videophones, the QCIF (176x144) format is used. The block diagram of the multi-resolution motion estimation/compensation (MRMEC) coding scheme is depicted in figure 2.

![Figure 2: Block diagram of the multi-resolution motion estimation coding scheme](image)

The image is first decomposed into 2-level resolution sub-bands using the 2-D DWT. As in section 2, the same separable high-pass and low-pass filters are used to decompose the image in the horizontal and vertical directions respectively. A 2-level DWT decomposition will give 3 sub-bands in the higher resolution level and 4 sub-bands in the lower resolution level. For the 4 sub-bands in the lower resolution level, motion estimation is performed on the lowest frequency sub-band using the full search block matching algorithm (BMA) with 4x4 sub-blocks. The minimum absolute mean motion estimation block error (MAM_MEBE) is used as the error criterion in the BMA to determine the best-match block. Two thresholds known as the error_send_threshold (EST) and the intra_block_threshold (IBT) where IBT > EST, are used to classify each sub-block into 4 different classes which determine the type of information to be transmitted to the receiver. The types of information to be transmitted are the block classification information (BCI), the motion vectors (MV) of the best-match block, the motion estimation block error (MEBE) and the intra-block (IB). The classification algorithm and the types of information to be sent for each class is described as follows:

If (MV = (0,0) and MAM_MEBE < EST)
then --> Class 1: stationary block: only the BCI is sent,
If (MV ≠ (0,0) and MAM_MEBE < EST)
then --> Class 2: fully motion compensated: the BCI and MV are sent,
If (EST ≤ MAM_MEBE ≤ IBT)
then --> Class 3: partially motion compensated: the BCI, MV and MEBE are sent,
If (MAM_MEBE > IBT)
then --> Class 4: non motion compensated: the BCI and IB are sent.

The BCI and the MV are coded with the ABPRLC described in previous section, whereas the MEBE and IB are first quantized with a uniform quantizer and then further coded using ABPRLC. The motion vectors and classifications of the lowest frequency sub-band are then used for the other 3 sub-bands in the lower resolution level since they represent the same motion structure of the image sequence at different spatial orientations. This will reduce the amount of information needed to be transmitted and also reduce the computation complexity. Statistical results show that the coefficients of the 3 sub-bands in the highest resolution level are normally small in value and have little effect on the quality of the reconstruction. Therefore for very low bit-rate applications, they are considered the least important and are normally discarded. In this coding scheme, information on the highest resolution sub-band are coded and sent only if the buffer content drops rapidly and stays very low which indicates very little motion in the sequence. No motion estimation is required for the 2 chrominance components (Cb,Cr) as they use the motion vectors and classifications of the luminance components.

A simple buffer control algorithm is used to control the buffer content fullness so as to allow constant buffer rate and to maintain acceptable picture quality. An initial error_send_threshold (EST) is set for classification of the motion estimation sub-bands into the above mentioned 4 classes. Initial quantization step-size (Q-step) for the MEBE and IB are also set. The EST and the Q-steps will be adjusted adaptively before coding the next frame depending on the buffer fullness. The buffer fullness control algorithm is described below. BF denotes the buffer fullness which ranges from 0 to 1.

If (0.9 ≤ BF < 1.0) then (EST = EST + 1; Q-step = Q-step + 2;)
If (0.75 ≤ BF < 0.9) then (EST = EST + 0.75; Q-step = Q-step + 1;)
If (0.5 ≤ BF < 0.75) then (EST = EST + 0.5; Q-step = Q-step + 1;)
If (0.25 ≤ BF < 0.5) then (EST = EST - 0.25; Q-step = Q-step - 1;)
If (0.1 ≤ BF < 0.25) then (EST = EST - 0.375; Q-step = Q-step - 1;)
If (0.0 ≤ BF < 0.1) then (EST = EST - 0.5; Q-step = Q-step - 2;)

To maintain a certain level of reconstruction quality, two picture quality control parameters (QCP) are used. These parameters which are known as QCP_T which determine the maximum limit of the EST, and QCP_Q which determine the maximum value of the Q-steps can be set by the user so as to control the picture quality. In the case of buffer overflow, one or more frames will then be skipped (not sent) for transmission which has the result of reducing the receiving frame rate. Therefore by adjusting the QCPs, the user can either choose a better picture quality but lower frame rate (small QCPs), or a higher frame rate but poorer quality (large QCPs).

3.1 Coding Results:
The coding scheme was tested on 38 frames of 176x144 colour 'miss america' image sequences at 6.25 frames/second. The chrominance components were decimated both horizontally and vertically (88x72 pixel/frame). Transmission rates of 19.2 and 9.6 Kbit/s were simulated respectively, and the QCP was selected to give an acceptable picture quality. To achieve this, the buffer size
was set reasonably large enough, which produced an increased transmission delay. In these simulations, the buffer size is set to give a maximum possible transmission delay of 0.66 seconds per frame, and the first frame is assumed to be already transmitted with an initial buffer fullness set as 0.5. Table 1 shows the performance in quality, transmission frame rate and transmission delay when transmitted at both 19.2 Kbit/s and 9.6 Kbit/s. At 19.2 Kbit/s, an average frame rate of 5.33 frame/s and an average frame transmission delay of 0.45 second/frame is achieved. The reconstruction quality is found to be acceptable with little jerkiness. Figure 3(a) shows one of the original frame figure 3(b) show the reconstruction transmitted at 9.6 Kbit/s with an average frame rate of 3.33 frame/s and an average delay increases to 0.58 second/frame. Although the transmission delay is longer and the reconstruction is jerkier, the quality is still acceptable with no blocking effect which would result if block-DCT based motion estimation algorithms were used.

Table 1: Performance of the codec simulated to transmit at 19.2 and 9.6 Kbit/s.

<table>
<thead>
<tr>
<th>Bit-rate (Kbit/s)</th>
<th>Av. frame-rate</th>
<th>Av. luminance PSNR</th>
<th>Av. frame trans. delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.2</td>
<td>5.33 frame/s</td>
<td>34.45 dB</td>
<td>0.45 sec/frame</td>
</tr>
<tr>
<td>9.6</td>
<td>3.33 frame/s</td>
<td>32.89 dB</td>
<td>0.58 sec/frame</td>
</tr>
</tbody>
</table>

Figure 3: (a) Original ‘Miss America’ (144x176 luminance)  
(b) Reconstruction simulated to transmit at 9.6 Kbit/s

4. Conclusions:
This paper has presented the performances of novel WT-based algorithms for low bit-rate image coding. The new WT-ABPRLC described used a simple and effective scheme which requires no training, no storage of codebooks and produces good quality and compression performance. Compared to the JPEG codec, results showed that the WT-ABPRLC gives better quality at high compression ratios. Compared to other hybrid WT based methods such as using VQ, it is much simpler to implement. Extending the use of WT for image sequence coding, the very low bit-rate MRMEC coding scheme incorporating a simple classification and buffer-rate control algorithm was developed. Coding results of the MRMEC codec at 9.6 Kbit/s and 19.2 Kbit/s gives acceptable results, in particular the 9.6 Kbit/s reconstructions produces visually no blocking effect. More sophisticated and accurate classification and buffer-rate control algorithm can be used to further improve the codec performance.

Acknowledgments
The authors would like to acknowledge the portable video research group (PVRG), for the use of the PVRG-JPEG 1.1 software package.

References

1217
Subband Coding of Images Using Hierarchical Quantization

Marek DOMAŃSKI, Roger ŚWIERCZYŃSKI

Politechnika Poznańska, Instytut Elektroniki i Telekomunikacji ul. Piotrowo 3a, 60-965 Poznań, Poland
Telfax: +48 (61) 782 372, E-mail domanski@ieitpp.poz.pl, roger@ieitpp.poz.pl

Abstract The paper presents a technique of efficient encoding of high-frequency image subband signals. The technique exploits some spatial correlation between subbands which is characteristic for real images. Although such a relation is clearly not valid for any image it works surprisingly well for real images. This conclusion is based on numerous experimental results showing also that more efficient recursive filters, despite of their nonlinear phase, behave not worse than their linear-phase FIR counterparts. The technique proposed is easily adaptable to color image coding.

1. Introduction
Digital image data compression techniques have been standardised for still images [1] as well as for video. The existing standards (JPEG, MPEG, H.261, etc.) are based mostly on discrete cosine transform while the later developed subband techniques seems to be potentially even more efficient. Importance of subband coding (SB) [2] stimulates interests in related encoding techniques (e.g., [3,4,13]) where encoding of high-frequency subbands is of particular interest.

The objective of the paper is to propose an efficient subband coding scheme exploiting internband correlation using the ideas somewhat similar to those from [15]. The systems considered use separable tree-structured (multi-level) filter banks with nonlinear-phase filters (first of all IIR filters).

![Tree-structured partition of the 2-D frequency band](image1)

**Figure 1.** Tree-structured partition of the 2-D frequency band (a three-level system).

In order to avoid phase problems we apply recursive systems. The basic idea of a recursive SBC system is to use the opposite directions of data processing in the coder and in the decoder, thus compensating for shifts caused by the filters. For example, an image can be processed by the coder filters from the left to the right and from the top to the bottom. In the reverse system, the decoder filters process the image from the right to the left and from the bottom to the top [5-8].

An implementation of a polyphase recursive IIR filter bank needs substantially less computational effort than for popular linear FIR quadrature-mirror filters (QMFs). For example, a polyphase implementation of a pair of the 1-D 24-tap FIR QMFs needs 12 times more multiplications than the polyphase implementation of a pair of the 1-D 9th-order elliptic filters with the equivalent attenuation characteristic [16]. Moreover, application of wave digital IIR filters guarantees stability, low sensitivity, and small roundoff noise. Such lattice polyphase filters are very appropriate for applications in recursive systems.

2. Inter-subband correlation
Usually pictures showing natural scenes exhibit some kind of cross-correlation between sub-images corresponding to different subbands of the spatial spectrum. This correlation manifests itself in similar patterns observed in various subbands (see Fig.2). It is related to some redundancy which implies an additional possibility of data compression.

![Two sub-images corresponding to two subbands of the image "Lena"](image2)

Moreover, natural scenes exhibit power spectra which decay for increasing frequencies. Therefore, the portions of the high-frequency sub-images where signal values are substantial correspond usually to the portions of the low-frequency sub-image where some changes occur. Most of the portions of the high-frequency sub-images exhibit only moderate signal values. Moreover, very small amplitudes usually have their
interpretation as noise, e.g., camera noise. Therefore the commonly used quantizer with a dead-zone and saturation produces a lot of zeros. We need not to transmit all of them and we can use the low-frequency sub-image to identify the positions of the "active" portions of the high-frequency sub-images. "Active" portions of the low-frequency sub-image correspond to details and textures in the original image and boundary effects.

We define some activity measures, e.g., for horizontal activity \[ H_x = \max \{ |I_{x,p}(n_1,n_2)-I_{x,p}(n_1,n_2-1)|, \ldots \} \]

Vertical and diagonal activities \( V_x \) and \( D_x \) are defined in a similar way.

We put forward a hypothesis that the portions of the sub-images in the high-frequency subbands that correspond to the "active" portions in the low-frequency subband well determine the reconstructed image. Other portions of the high-frequency sub-images may be skipped or coarsely quantized in order to reduce the amount of data (see Fig.3).

**Figure 3.** The subband 9 after skipping of "inactive" portions.

Of course, the hypothesis is generally not true but we show that for real images the technique proposed gives promising results.

In the technique proposed, quantization of the high-frequency sub-images is controlled according to the values of respective activity indices (horizontal, vertical or diagonal). Controlled is the quantization step, dead-zone width and saturation level. In principle, we keep the number of discrete signal values proportional to the activity indices.

There are possible some strategies of quantization control. The simplest is to skip all the pixels where activity measures do not exceed some predefined thresholds. More elaborated techniques use switching of the number of quantization levels according to value of this activity index, which correspond to the spatial spectrum in a given subband.

3. Hierarchical quantization

We use recursive filter banks, i.e., we use filters with inherently nonlinear phase characteristics. A simple and efficient implementation is for complementary filters where \( H_1(-z) = H_2(z) \). Here \( H_1 \) and \( H_2 \) denote the transfer functions of the low-pass and the high-pass filters, respectively. Therefore the group delays of both complementary filters are the same and there is no delay between the low-frequency sub-image and the high-frequency sub-images. It is why we are able to exploit inter-subband correlation in a system with nonlinear-phase filters.

Note that the above corollary is valid only for sub-images from the same analysis/synthesis level. Therefore the coder needs a reconstructed version of the low-frequency subband. This sub-image is available only at the lowest level. Reconstructed versions of the low-frequency sub-images of higher levels can be obtained by synthesis from the sub-images at the lower level.

The decoder which receives only quantized versions of sub-images take use of the same information as the quantizer in the coder. It guarantees that no annoying displacement errors occur.

The hierarchical structure of the coder is shown in Figure 4. We denote:

- **A** Analysis filter bank (of a given level),
- **S** Synthesis filter bank (of a given level),
- **Q** Quantization controlled by the subband LL,
- **C** Control of quantization.

**Figure 4.** Quantization scheme for the two lowest levels.

The simplest technique related to hierarchical quantization consists in thresholding, i.e., a pixel in the low-frequency sub-image is considered as "active" if \( H_x > T_x \) (or \( V_x > T_x \) or \( D_x > T_x \) ), where \( T_1 \) is a predefined threshold for the level 1. We define the threshold \( T_x \) for subbands 4,5,6 (level 2) and the threshold \( T_3 \) for the subbands 1,2,3 (level 3). The technique results in ignoring a substantial part of the pixels from the high-frequency subbands.

The image has to be retrievable uniquely, and therefore in the coder all decisions can be made only using the information available in the decoder. It leads to the following algorithm \[ [12] \]:

1) Split a given image into subbands 0-9.
2) Quantize all the subbands (DPCM can be additionally applied for the subband 0).
3) Neglect the pixels in the subbands 1,2,3 which correspond to "inactive" portions of the subband 0.
4) Using the quantized and truncated versions of the subbands 0-3 reconstruct the low frequency band of the level 2.
5) Neglect the pixels in the subbands 4,5,6 which correspond to "inactive" portions of the low frequency subband retrieved in the step 4.
6) Using the quantized and truncated versions of the subbands 4,5,6 as well as the coded version of the low frequency band of the level 2, reconstruct the low frequency band of the level 1.

1219
7) Neglect the pixels in the subbands 1, 2, 3 which correspond to "inactive" portions of the low frequency subband retrieved in the step 6.

8) Encode each subband using entropy coding techniques (like Huffman coding).

The "active" portions are related to details and textures in the original image as well as to transients in subbands. The transients are particularly significant on the boundaries of the image and therefore it is very important that there are no displacement between sub-images corresponding to different subbands.

Due to this technique, filters with short impulse responses are preferable. Such a requirement is in a contradiction to minimization of width of the transient band. One can show that filters worse from the point of view of their parameters (e.g., transition-band width, stopband attenuation, etc.) might behave worse than other with shorter impulse response, i.e., their application would lead to somewhat higher compression ratios.

Here, we demonstrate the results for three well-known test pictures: "Boats", "Boy" and "Lena". First of all, we examine how the technique affects the quality of the image reconstructed. Assuming \( T_2 = T_3 = 0 \) we find that in the subbands 7-9 the fraction \( r \) of pixels ignored is very similar for all three examined test pictures. Nevertheless, the quality of the reconstructed pictures varies significantly. As a measure of the objective quality a signal-to-noise ratio is used

\[
\text{SNR}_{\text{obj}} = \frac{255^2}{\mathcal{E}_{\text{ms}}}
\]

where \( \mathcal{E}_{\text{ms}} \) is a mean square error. The subjective quality is measured as an opinion score where 5.0 denotes very good reconstruction (no degradations are visible from the normal viewing distance) and 2.0 means significant presence of annoying degradations.

The technique causes more visible degradations for pictures with many fine details than for smooth pictures although the results in SNR might be quite opposite. Similar conclusions are deducible from the results for the level 2 (e.g., see Fig.8). Obviously, the image quality deteriorates more by neglecting pixels from higher frequency subbands of higher levels. Therefore it is reasonable to choose \( T_1 > T_2 > T_3 \).

![Figure 5. Neglected pixel fraction in subbands 7-9 as a function of \( T_1 \) for the images "Boats" (solid line), "Boy" (dashed line) and "Lena" (dotted line).](image)

![Figure 6. SNR versus \( T_1 \) for the images "Boats", "Boy" and "Lena".](image)

![Figure 7. Subjective quality (opinion score) versus \( T_1 \) for the images "Boats" (solid line), "Boy" (dashed line) and "Lena" (dotted line).](image)

![Figure 8. Subjective quality (opinion score) versus \( T_2 \) for the images "Boats" (solid line), "Boy" (dashed line) and "Lena" (dotted line).](image)

![Figure 9. Comparison of SNR for the image Lena for FIR and IIR filters.](image)

1220
Table 1. Comparison of SNR for both filters for images Boats* and "Lena"

<table>
<thead>
<tr>
<th>COMPRESSION [bit/pel]</th>
<th>BOATS</th>
<th>LEN A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FIR</td>
<td>IIR</td>
</tr>
<tr>
<td>2.00</td>
<td>43.96</td>
<td>44.12</td>
</tr>
<tr>
<td>1.65</td>
<td>41.14</td>
<td>41.32</td>
</tr>
<tr>
<td>1.25</td>
<td>38.94</td>
<td>39.05</td>
</tr>
<tr>
<td>1.00</td>
<td>36.72</td>
<td>37.09</td>
</tr>
<tr>
<td>0.85</td>
<td>36.03</td>
<td>36.42</td>
</tr>
<tr>
<td>0.75</td>
<td>35.53</td>
<td>35.41</td>
</tr>
<tr>
<td>0.60</td>
<td>34.73</td>
<td>34.54</td>
</tr>
</tbody>
</table>

4. Simulation Results

The results are illustrated with experiments with real test images which prove good properties of the technique proposed. For monochrome images we got compression to 0.7-0.9 bit/pixel for good opinion score and SNR over 30 dB.

It is important to mention that the results obtained with linear-phase QMFs are not better. In our experiments we have used 5th-order elliptic filters and 24-tap FIR QMFs [11] with similar attenuation characteristic. The results for elliptic filters were even slightly better (e.g., Table 1 and Fig.9). Period extensions of lines and columns have been used.

5. Conclusions

The paper presents a concept of hierarchical quantization. The coding technique related to this concept is proved to be quite efficient for real scenes but it fails for some artificial geometrical pattern, i.e., for computer graphics.

The technique can be adopted for color images. Chrominance components are split into subbands in the frequency domain using the same scheme as applied for luminance, nevertheless we use already subsampled chrominance signals at the input and thus the level 1 of the system vanishes.

The authors consider the technique as a partial solution in a prospective method (standard) based on subband coding.

The proposed SBC system is asymmetric, i.e., the coder is much more complex than the decoder. The coder consists of an analysis filter bank, a quantizer, a synthesis filter bank, and a compressor working according to the inter-subband correlation principle. The decoder consists mostly of the synthesis filter bank which is the same as that in the coder.

The system is suitable for massively parallel implementations. The separable filter banks with simple allpass sections are pipelineable and the processing of all lines and columns can be performed fully independently. Moreover, the two all-pass sections of a filter can be implemented individually.

Similarly, the procedure of neglecting the points in the high frequency subbands is performed independently for all pixels and therefore the technique is highly parallel.

Acknowledgements

The paper has been prepared under support of National Research Committee (Poland) no. 3 0754 91 01 and S 504 002 06

References


1221
Linear Phase Wavelet Transforms for Low Bitrate Image Coding

Eduardo A. B. DA SILVA  Mohammad GHANBARI

Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, England, Tel/Fax: +44 (0206) 8793744/8792000, E-Mail: ghan@essex.ac.uk

Abstract The performance of linear phase wavelet transforms at low bitrate image coding is investigated. The influence of certain characteristics of these transforms such as: regularity, number of vanishing moments, filter length, coding gain and the shape of the wavelets on the picture quality is analysed. The assessment was carried out by subjective tests on images coded with different wavelet transforms. The results show that features like regularity, number of vanishing moments and coding gain do not have any important impact on the final image quality. On the other hand, the shape of the synthesis wavelet, which determines the visibility of the coding errors on the reconstructed images, is very important. Analysis of the obtained data strongly suggests that design of good wavelet transforms for low bitrate image coding should take into account chiefly the shape of the synthesis wavelet, and the coding gain to a smaller extent.

1 Introduction

A Wavelet Transform is the decomposition of a signal into a set of basis functions consisting of contractions, expansions and translations of a mother function $\phi(t)$, called the wavelet [1]. From a coding point of view it is equivalent to the analysis of a signal into several frequency bands, each one representing a trade off between time and frequency resolution. The wavelet transform represents a new approach towards subband coding [1], that has been attracting the signal coding community with growing interest. In this approach, other features of the filter banks, such as regularity and number of vanishing moments may have to be considered [2]. One question that naturally arises is: to what extent do these features affect the performance of wavelet transforms in low bitrate image coding? The aim of this paper is to determine the influence of wavelet characteristics such as regularity, number of vanishing moments, filter length, coding gain and the shape of the wavelets on the picture quality.

2 Wavelets and Coder Used

Since for image coding applications it is convenient to have linear phase wavelets, orthogonal wavelets, which do not have linear phases [1], are not used in the assessment. Thus, the wavelets assessed have to be biorthogonal. We have designed 21 such wavelets, which have a reasonably high coding gain [3], with the aim that the product of the analysis and synthesis low pass filters have the maximum number of zeros at $\omega = \pi$ [4]. The order of this product varies from 6 to 26. For the sake of comparison, the wavelet approximated by the QMF filter bank 16A proposed in [5] is also assessed, for this filter represents a good approximation to an ideal filter, with the advantage of having linear phase.

The two-dimensional wavelet transforms employed here were of separable type, where one-dimensional transformations were applied in both horizontal and vertical directions of the images. A three stage decomposition was done in each direction, ending up with 9 high pass bands and 1 low pass band. The low pass band was coded with PCM at 8 bits/pixel, and the other coefficients were coded with a multistage product code lattice based vector quantiser [6]. The assessment was carried out by subjective tests where 7 expert viewers judged the quality of coded pictures on a 5 grade impairment scale. The LENA test image of 256x256 resolution was coded by all 22 wavelet transforms at a target bitrate of 0.5 bit/pixel.

3 Wavelet Characteristics

The influence of several features of the wavelets in their subjective performance was evaluated. Among these features are the regularity and the number of vanishing moments of the analysis and synthesis wavelets [2], and the generalised coding gain defined by Katto and Yasuda [3]. Normally in low bitrate image coding, coding distortions are visible but to some extent can be tolerated. Of course, in a good low bitrate image coding scheme errors should have the least visibility. From the wavelet theory, it can be easily seen that an error introduced by a coefficient into a reconstructed signal will be spread out inside the signal by the synthesis wavelet [2]. The degree and the nature of the spread of error depends on the shape of the synthesis wavelet. Figures 1.a-d show some examples of these wavelets. The wavelet in figure 1.d, which has larger oscillations than the one in figure 1.c, tends to spread the coding errors to a larger extent, and hence producing more visible noise. This suggests that the performance of a biorthogonal wavelet transform in low bitrate image coding is influenced by the shape of its synthesis wavelet. Thus, in analysing the correlation among the
wavelet features and their performance in low bitrate image coding, it is important that a measure of the visibility of the coding errors based on the shape of the synthesis wavelets be included in the evaluation.

As a measurement of this visibility, the lengths of the analysis and synthesis filters and the order of the product of these filters were considered. Also, in order to take into account the behaviour associated with the shape of the synthesis wavelets, such as how fast its amplitude decays towards its tails, we have devised a measure called here the Peak to Peak Ratio (PPR). It is the ratio of the maximum peak to peak value of the wavelet to the arithmetic mean of the second and third maximum peak to peak values. For figures 1.a to 1.d they are, respectively, 1.6, 1.4, 1.6 and 1.4. The larger the PPR is, the more "concentrated" the wavelet becomes, hence the spread of the error becomes smaller and less visible.

4 Results

Figures 2 and 3 show the results of the subjective rating in terms of the regularity and the number of vanishing moments of both analysis and synthesis wavelets, the analysis and synthesis filter lengths, the coding gain and the peak to peak ratio (PPR). The optimum least squares fit to the data of each figure is shown by a straight line. From these figures the only features which can be significantly correlated to the subjective grade are the ones which include the PPR. The PPR alone, for example, has a fitness correlation coefficient of 0.75, which is reasonably high. Table 1 lists the correlation coefficients of the least squares fits of figure 2. Surprisingly, even the coding gain is poorly correlated to the subjective rating, which has a low correlation coefficient of 0.33, as shown in table 1. The regularity and the number of vanishing moments of both the analysis and synthesis wavelets also seem to have little influence on the performance of the wavelet transforms. Figures 3.a, 3.b and 3.c also show that there is no relation between the lengths of the filters and the subjective performance. Interestingly, it has also been verified that the product of the PPR and the coding gain has an even greater correlation to the subjective grade than the PPR alone. This is demonstrated in figure 3.f, as a plot of the subjective grade versus the product of the coding gain and the PPR. The graph shows a very strong correlation between the two. From table 1, its value is 0.87, which is quite high.

5 Conclusions

It can be concluded that the performance of linear phase wavelet transforms in low bitrate image coding is only slightly influenced by factors such as regularity, number of vanishing moments, filter length or coding gain. However, the visibility of the synthesis wavelet, measured by its peak to peak ratio (PPR), has a very strong correlation with the wavelet transform performance. In addition, the product of the PPR by the coding gain shows even higher correlation to the subjective performance. These results strongly suggest that design of wavelet transforms for low bitrate image coding should be primarily aimed at maximising the product of
Figure 2: Subjective rating versus: (a) regularity of the analysis; (b) regularity of the synthesis; (c) vanishing moments of the analysis; (d) vanishing moments of the synthesis.

<table>
<thead>
<tr>
<th>feature</th>
<th>grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>reg-A</td>
<td>-0.59</td>
</tr>
<tr>
<td>reg-S</td>
<td>0.52</td>
</tr>
<tr>
<td>vm-A</td>
<td>0.26</td>
</tr>
<tr>
<td>vm-S</td>
<td>-0.24</td>
</tr>
<tr>
<td>L0</td>
<td>0.08</td>
</tr>
<tr>
<td>L1</td>
<td>-0.14</td>
</tr>
<tr>
<td>order</td>
<td>-0.03</td>
</tr>
<tr>
<td>c.gain</td>
<td>0.33</td>
</tr>
<tr>
<td>PPR</td>
<td>0.75</td>
</tr>
<tr>
<td>PPRxc.gain</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 1: Correlation coefficients between several wavelet features and the subjective grade.

REFERENCES


Figure 3: Subjective rating versus: (a) L0; (b) L1; (c) order; (d) coding gain; (e) PPR; (f) PPR x coding gain
Adaptive Subband VQ for Image Coding

Stathis P. VOUKELATOS, John J. SORAGHAN and Tariq S. DURRANI

Signal Processing Division, Dept. of Electronic and Electrical Eng., University of Strathclyde, Glasgow G1 1XW, Scotland, UK, Tel: +44 [0141] 552 4400 x2499, E-Mail: stathis@spd.eee.strathclyde.ac.uk

Abstract. In this paper an adaptive VQ scheme for still image compression is investigated. Adaptive VQ overcomes the problems present in trained VQ systems, where a training set of images is employed to design an optimal fixed codebook, by deriving the set of training vectors from the input image to be encoded. Pruned tree searched VQ (PTSVQ) is investigated as an alternative to the optimal full search VQ, in order to reduce the computational complexity of the encoding process. An adaptive hybrid VQ system is also developed, where multiresolution (wavelet) decomposition of the image is used. The simulation results demonstrate the substantial decrease in the encoding computation times obtained with PTSVQ, for only a small performance penalty, and the increased coding efficiency achieved when adaptive VQ is applied to the wavelet transform domain.

1. Introduction

Vector quantization (VQ) [1] is a process where the data to be quantized is broken into small blocks or vectors, which are then sequentially encoded using a codebook of representative vectors. According to Shannon’s rate-distortion theory, better performance is always achievable, in theory, by coding vectors instead of scalars, even when the data source is memoryless. The application of VQ in image coding has been extensively studied in recent years [2] and a variety of VQ based coding systems have been developed. The main advantages of vector quantization are the high coding efficiency and the very simple structure of the decoder. One of the key issues in the design of a VQ system is to identify a good codebook of representative vectors. The approach most widely used in practical VQ schemes is to derive a sequence of training vectors from a set of training images and apply a codebook design algorithm, such as the LBG, on the training vectors. Therefore, a suitable training set must be found and a sufficiently large codebook must be designed so that acceptable performance can be maintained over all possible input images. This results in an increase in the bit rate and in the amount of memory required for codebook storage.

In this paper we investigate and extend an adaptive vector quantization scheme [3], where, instead of using a training set of images, a set of local codebooks are generated using the vectors of the input image. These adaptively designed codebooks are better able to track the variations in local image statistics. One of the drawbacks of adaptive VQ is the computational complexity of the encoding process due to the fact that a set of codebooks have to be generated on-line for each input image. In order to reduce the codebook design complexity, pruned tree searched VQ (PTSVQ) [4] is investigated. The performance of variable rate PTSVQ is superior to fixed rate TSVQ and comparable to the optimal full search VQ. The simulation results demonstrate that codebook design time is significantly reduced when PTSVQ instead of full search VQ is used.

Vector quantization has also been successfully used in hybrid subband coding schemes for the quantization of the subband coefficients [5-7]. Subband coding systems are free from the blocking artifacts, present in other transform coding techniques, and are better able to match the characteristics of the human visual system. Here we present an adaptive subband VQ system for the encoding of still images in the wavelet transform domain. The simulation results confirm the improved coding efficiency that the wavelet transform provides.

2. Adaptive Vector Quantization

Nonadaptive VQ systems employ a single large codebook, generated from a training set of images. The difficulties involved in designing and storing a large fixed codebook are overcome with the adaptive VQ scheme, where a set of small codebooks is designed from the vectors of the input image. The input image is first partitioned into a set of nonoverlapping rectangular subimages. A set of local codebooks is then generated using the vectors of each subimage as a training set. The vectors of each subimage are quantized with the corresponding codebook. The output of each vector quantizer is entropy coded using a Huffman code. The Huffman codes are designed adaptively using the distribution of the vectors in the codewords of each codebook.

The effective bit rate is determined by the amount of data required to transmit both the labels of the quantized vectors and the codewords of all the codebooks. The contribution of each of these two components in the overall bit rate depends on the selection of the subimage and vectors sizes, in which the initial image is partitioned.
3. Pruned Tree Searched VQ (PTSVQ)

One of the drawbacks of adaptive VQ is the large computational effort required to design the individual codebooks of each subimage. Each codebook is generated by applying the well known LBG clustering algorithm to the vectors of the corresponding subimage. This algorithm is guaranteed to converge to a locally optimum codebook after a variable number of iterations. In order to reduce the number of computations a tree structured codebook design approach is used.

Fixed length tree structured codebooks are widely used in vector quantization systems. However, because of the constraints imposed on the training sequence during the codebook design process, tree structured VQ is suboptimal compared to full search VQ. Improved performance can be achieved with variable length trees, which result in quantizers with variable output rate. The approach used here is to design a large initial fixed length tree and then selectively prune off some of the branches of the tree, so that the best trade-off between rate and performance is achieved. The pruning algorithm used is the generalized BFOS algorithm [4]. It begins with an initial large fixed length tree and prunes it back until it reaches the subtree with the smallest average distortion for a given average length.

If \( T \) is the initial tree, there corresponds to every pruned subtree \( S \) of \( T \) (denoted \( S \subseteq T \)) a variable rate tree structured VQ with average rate \( h(S) \) and distortion \( \delta(S) \). The operational distortion-rate function

\[
\hat{D}_F(R) = \min \{ \delta(S) \mid h(S) \leq R \text{ and } S \subseteq T \}
\]

specifies the optimal trade-off between rate and distortion, where the quantizer is constrained to be some pruned subtree of \( T \). The pruning algorithm is capable of listing the vertices of the lower boundary of the convex hull of the distortion-rate function, along with their associated pruned subtrees.

The vectors of each subimage are coded with the corresponding pruned tree and then the resulting codeword labels are entropy coded, using a Huffman coder. Since entropy coding is used at the output of the vector quantizer, it is desirable to obtain a pruned tree with the minimum average distortion for a given entropy or vice versa. Thus, the pruning algorithm is applied to trade-off average distortion with entropy instead of average length.

4. Adaptive Subband VQ

Improved coding efficiency and image quality is obtained when the adaptive VQ approach presented earlier is applied in the wavelet transform domain. In the hybrid adaptive VQ scheme developed here, a wavelet transform is applied to split the input image into a number of multiresolution subbands. A biorthogonal wavelet basis, reported in [5], is used to implement the wavelet transform. The two biorthogonal filters are both very close to an orthonormal wavelet filter.

The lowest resolution subband, which contains most of the energy, is encoded losslessly because of its importance in the quality of the reconstructed image. The remaining subbands have low energy and mainly contain edge information of the original image. A large percentage of the coefficients of these subbands has very small magnitude. By setting these coefficients to zero, significant savings in bit rate and compression times can be achieved and, moreover, the error introduced may be smaller than if the coefficients were vector quantized. Moreover, compression time is substantially reduced since only a subset of the vectors of each subband participates in the adaptive codebook design and quantization process.

A regular decomposition, using a quadtree structure, is applied to each of these subbands in order to identify coefficient blocks of low activity. The subband coefficients of the blocks whose variance is below the preset variance threshold, used in the quadtree decomposition process, are set to zero. The remaining small blocks, that have a higher variance, are collected and used as training vectors to design adaptively a codebook for each subband. These blocks are then vector quantized with the corresponding codebook. Both full search and pruned tree searched codebooks are used to compare their performance.

The bit allocation among the subbands is an important design issue of the adaptive subband VQ system. The bit allocation is determined by the selection of the variance threshold and minimum block size allowed in the quadtree decomposition as well as the codebook size for each subband. Generally, the subbands corresponding to higher spatial frequencies are allocated a smaller number of bits. In our simulations a good bit allocation has been determined experimentally for the test image used.

5. Simulation results

The test image used in the simulations is the popular 512 x 512 image “Lenna”. The intensity of each pixel is coded with 256 grey levels (8 bpp). The original test image is depicted in Figure 1.

In adaptive vector quantization the input image is divided into rectangular subimages and vectors. The performance of adaptive VQ for different subimage and vector sizes is shown in Figure 2 and Figure 3, respectively. The two graphs have been produced using pruned tree searched codebooks. The best performance is obtained when subimages of size \( 64 \times 64 \) and vectors of size \( 2 \times 2 \) are used. In the results of Figure 4, the test image has been compressed to a bit rate of 1.02 bpp using full search VQ and PTSVQ. The simulations were performed on a SUN SPARC-2 workstation and the CPU time for compression is also given in

Figure 1. The original image “Lenna”.

1227
Figure 2. Rate-distortion curves for different subimage sizes. Vector size = 2 × 2

Figure 3. Rate-distortion curves for different vector sizes. Subimage size = 64 × 64

each case. It can be observed that the performance of PTSVQ is comparable to that of full search VQ, while the execution time is substantially reduced.

In order to evaluate the performance of adaptive subband VQ, the test image was decomposed into 10 subbands. In the results shown in Figure 5, the test image has been coded at a bit rate of 0.46 bpp using full search VQ and PTSVQ for the design of the subband codebooks. The lowest resolution subband was coded at 8 bpp. A variance threshold of 2.5 is used for the quadtree decomposition in the remaining subbands. Codebook and vector sizes have been allocated experimentally among the subbands. Pruned tree searched VQ achieves again a significant reduction in compression time with hardly any performance penalty.

A comparison of adaptive VQ and adaptive subband VQ demonstrates the improved performance that is achieved when vector quantization is applied to the subband coefficients instead of the original image. Apart of the savings in bit rate, also the perceptual quality of the reproduced images is better, with very small visible degradation.

6. Conclusions

This paper examined the enhanced performance of an adaptive subband VQ system over adaptive VQ. When adaptive vector quantization is applied in the subband domain, higher compression ratios and improved perceptual quality of the reproduced images can be achieved. The cost for that improved performance is an increase in the complexity of the system. Pruned tree searched VQ has also been investigated as an alternative to the optimal full search VQ, in order to reduce the computational complexity of the encoding process. It is shown that PTSVQ can achieve a significant reduction in compression time with only a

Figure 4. Adaptive Vector Quantization.
(a) Full search VQ. Bit rate: 1.02 bpp, PSNR: 32.64 db, CPU time: 898 sec.
(b) Pruned tree VQ. Bit rate: 1.02 bpp, PSNR: 32.42 db, CPU time: 209 sec.

slight degradation in performance. Further speed up can be obtained if the inherent parallelism of the adaptive VQ scheme is exploited. Adaptive vector quantization, either in the spatial or the wavelet transform domain, can be a strong alternative for applications where the robustness that nonadaptive systems usually lack, is required.

References


Figure 5. Adaptive Subband Vector Quantization.
(a) Full search VQ. Bit rate: 0.46 bpp, PSNR: 32.94 db, CPU time: 27.5 sec.
(b) Pruned tree VQ. Bit rate: 0.46 bpp, PSNR: 32.93 db, CPU time: 66 sec.


Image Subband Coding with Spatially Adaptive IIR Filter Banks - Automatic Filter Selection

Rune SØRHUS and John Håkon HUSØY

Rogaland University Center, Dept. of Electrical and Computer Engineering,
P.O.Box 2557 Ullandhaug, N-4004 Stavanger, Norway

Abstract. In subband coding at low bit rates the most annoying quantization distortion appears either as blocking effects or as ringing noise. The exact nature of the noise is dependent on the filter bank used and local characteristics of the image. Use of a filter bank with short channel impulse response, results in blocking effects in smooth image areas. Long impulse responses results in ringing noise in the vicinity of edges. To reduce these effects a solution is to apply spatially adaptive IIR filter banks as introduced independently in [1] and [2]. In this paper we present a technique for automatic selection of filter bank in an image subband coder. For a 32 x 32 pixel image block a separate filter bank is chosen for each filtering direction, among a set of given filter banks. This is done by classifying a feature vector extracted from a DCT of the block. The classification is done by a trained LVQ network. The filter selection for the edge blocks is taken special care of in a postprocessing step. The visual apperance of ringing and blocking effects is clearly reduced in most of the coded images.

1. Introduction

During the last years much attention has been given to the issue of constructing proper filter banks for use in image subband coders. A common approach has been to design filter banks with "good" frequency responses in terms of stop band attenuation, etc. This way of designing filter banks generally leads to long impulse responses for the filter banks' channels. Long impulse responses results in ringing noise in the vicinity of edges (discontinuities) when the coding is done at low bit rates.

On the other hand, use of filter banks having short impulse responses, results in blocking effects in smooth areas of the decoded image.

To reduce these effects a solution is to apply spatially adaptive IIR\(^1\) filter banks with perfect reconstruction as introduced independently in [1] and [2]. This means that we select a filter bank with short impulse response in areas containing edges in the filtering direction, and filter banks with longer impulse responses in smooth areas. Since the filter banks used in this work are separable, an independent selection has to be done for both the horizontal and the vertical filtering direction. The problem is how to do this selection automatically. A control system for adapting the filters by examination of the first backward signal difference, is proposed in [3]. In this paper we present a technique for the automatic selection of filter bank for local image areas of fixed size.

The basic principle of our automatic selection scheme is to partition the image into uniform square blocks of 32 x 32 pixels. For one image block a separate filter bank is chosen for each filtering direction, from a set of predefined filter banks. The idea is to select the filter bank for each block that results in minimal quantization noise in the decoded image with respect to the properties of the human visual system (HVS).

Because we use a set of predefined filter banks, this approach results in a problem of classification: Each image block has to be classified according to its characteristics, into one of several filter bank classes (in both directions). This classification is done by means of Kohonen's Learning Vector Quantization (LVQ) network [4, 5].

An advantage of our approach is that the amount of side information will always be known before coding, and in practice it is almost negligible. On the other hand, using our adaptive coder to get less visible quantization noise, we have to pay through an increased complexity in the encoder. In the decoder stage, the complexity is the same as without the adaptive scheme.

In the next section we show how we can make the IIR filter banks adaptive and at the same time retain perfect reconstruction. Section 3 deals with the problem of automatic selection of filter banks for the image blocks. In section 4 some experimental results are discussed. The paper is summarized and concluded in section 5.

2. Perfect Reconstruction with Adaptive Filter Banks

The principle of analysis/synthesis systems in subband coding is well described in many publications, for instance in [6, 7, 8]. In this work we use a tree-structured, separable filter bank with 8 x 8 uniform subbands. The filter bank is based on 2 band IIR quadrature mirror filters, realized using allpass polyphase structures as illustrated in Figure 1 [9].

The allpass filters \(A(z)\) are of first order with time-variant coefficients \(a(n)\) (filters of first order is sufficient to obtain satisfactory filter characteristics). Setting \(x(n) = y_1(n)\), the crossover terms in Figure 1 cancel out, and one channel of the adaptive 2 band filter bank prototype can be illustrated as in Figure 2. If the signal is of finite length, we can, as shown in [1], ensure perfect reconstruction by setting the modified allpass coefficients, denoted \(a_2^*(n)\), in the synthesis stage to

\[
a_2^*(n) = a_1(-n+1).
\]
Since the allpass filters used in this work are of first order the coefficients $a_f(n)$ may be altered at any point in time.

3. Automatic Filter Selection

Automatic filter bank selection is done by classifying a feature vector, $e$, corresponding to the image block. The classification is done by Kohonen’s LVQ network [4, 5]. The network was trained and tested using a set of feature vectors classified by manual inspection.

The amount of edge pixels, and sharpness and direction of edges, are the main characteristics that determine the choice of filter bank for the particular image blocks. Before the extraction of the feature vectors a preprocessing step is included to emphasize these basic characteristics and to simplify the feature extraction, see Figure 3.

In Figure 3 $x(n)$ indicates the image to be processed, $P(u,v)$ indicates the preprocessed image and finally $E$, $F$ and $G$ indicates $M_B \times N_B$ matrices where $M_B$ and $N_B$ are the number of image blocks in vertical and horizontal direction, respectively. The elements of $E$ are the feature vectors $e_{mn}$ with dimension $N_E$. In $F$ and $G$ the elements are in the form $f_{mn} = (f_{2mn}^h, f_{2mn}^v)$, where $f_{2mn}^h$ and $f_{2mn}^v$ indicate the selected filter banks in horizontal and vertical direction, respectively, for image block number $m,n$.

3.1 Preprocessing

The preprocessing step consists of a Discrete Cosine Transform (DCT) [10] of each image block

$$C(u, v) = \alpha \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} B(x, y) \cos \frac{(2x+1)u\pi}{2N} \cos \frac{(2y+1)v\pi}{2N}$$

where $B(x, y)$ is the image block signal, and

$$f_{(x,y)} \rightarrow \text{Preprocessing} \rightarrow P(u,v) \rightarrow \text{Feature extraction} \rightarrow E \rightarrow \text{Classification by LVQ} \rightarrow F \rightarrow \text{Postprocessing} \rightarrow G$$

Figure 3. Automatic filter selection.

$$\alpha(u) = \begin{cases} \frac{1}{\sqrt{N}} & \text{for } u = 0 \\ \frac{2}{\sqrt{N}} & \text{for } u = 1, \ldots, N - 1 \end{cases}$$

followed by the nonlinear transform

$$P(u,v) = \frac{1-e^{\alpha C(u,v)}}{1+e^{\alpha C(u,v)}} = \left\lfloor \frac{\alpha C(u,v)}{2} \right\rfloor$$

where $C(u,v)$ denote the coefficients of the DCT. The coefficient $\alpha$ was experimentally estimated to 0.07.

The amplitudes of the DCT high frequency coefficients increase with the degree of sharpness of the edges (discontinuities) in an image block. Since most of the signal energy in an image lies in the lower frequency coefficients, these coefficients will always be of significantly higher amplitude than the high frequency coefficients, even with very sharp edges in the image block. Therefore, we found it necessary to emphasize the high frequency coefficients according to eq.(4).

3.2 Feature Extraction

Extraction of the image block’s features is done by measuring simple local statistics (mean value) of the preprocessed image block. Edges in the image block with different orientation causes high amplitudes in different areas of the “DCT-plane”. By partitioning the block into sectors as shown in Figure 4, we are able to separate image blocks that contain edges with different orientation. Sharp edges cause high amplitudes in areas containing both low and high frequency coefficients, while smoother edges result in smaller amplitudes in the “high frequency area”. By partitioning the block into divisions, image blocks containing edges of different sharpness may be separated. This results in a block partitioned into local zones as shown in Figure 4, each zone determined by its sector and division. The divisions used in this work are of uniform size. The sector borders are symmetrical about the antidiagonal resulting in an odd number of sectors.

To measure the activity (or the energy) inside each zone we compute the mean value of each zone. These values constitute
separate coefficients in the feature vector. The feature vector's dimension is then equal to the number of zones. Because of its lack of relevant information, the DC-coefficient is not taken into account as part of any zone. The exact number of zones and the placing of the zone borders (sector and division borders) was found experimentally and depends on the number and the characteristics of the predefined filter banks.

3.3 Classification Using LVQ

For one image block the filter bank used in both of the filtering directions was selected (classified) in one single operation. Use of 2 different filter banks in each direction gives rise to a total of 4 classes, with 3 different filter banks the number of classes is 9. 25 reference vectors were used to represent each class in the initial codebook. Then, with 4 classes, we get a total of 100 codebook vectors. During the training cycle, some of the codebook vectors may be assigned to another class, depending on the distribution of the training vectors.

The codebook vectors were trained using a set of feature vectors corresponding to image blocks classified by manual inspection. Additionally, some of the manually classified images were transposed, leading to an increased number of training/test vectors without further manual classification.

Training took place in two separate steps: The first step consists of several iterations using an optimized version of the LVQ1 algorithm (OLVQ1) [11], which places the reference vectors near their optimal positions in the hyperspace in a small number of iterations. The number of iterations was 40 times the number of codebook vectors $N_C$. This is followed by a step in which the goal is to adjust the decision borders between codebook vectors corresponding to different classes. The learning algorithm used here is the LVQ3 algorithm, with 500 x $N_C$ iterations.

After the training of the codebook vectors, these vectors were used to classify the feature vector of each image block using the principle of nearest neighbor classification.

3.4 Postprocessing

Due to the fact that the filter banks are realized by circular convolution [12], the choice of filter banks for the image edge blocks must be taken special care of in a postprocessing step, see Figure 3. The reason for this is the possibility that the image edge pixels give rise to discontinuities that cause severe ringing effects near the image edges. A comparison of the results from classification by LVQ and classification through a algorithm based only on the pixels in the block that constitutes the image edge pixels, is done, and an appropriate filter bank is chosen based on a set of rules [13].

4. The Coding Scheme

8 x 8 filter banks were used, giving a total of 64 spatial frequency bands. To represent these subbands in a bit efficient manner, the LP-LP band was coded separately using DPCM [14] with adaptive prediction coefficients as described in [6]. Quantization of the LP-LP band was done using a Laplacian quantizer with 5 bits per pixel.

Higher subbands were coded according to [6]. First they were scanned in a zig-zag fashion, one scan consists of corresponding samples from the 63 higher subbands. The scans were then thresholded and quantized using a uniform quantizer before being runlength and entropy coded.

<table>
<thead>
<tr>
<th>$ij$</th>
<th>00</th>
<th>02</th>
<th>20</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>61%</td>
<td>20%</td>
<td>17%</td>
<td>2%</td>
</tr>
<tr>
<td>02</td>
<td>19%</td>
<td>72%</td>
<td>4%</td>
<td>5%</td>
</tr>
<tr>
<td>20</td>
<td>16%</td>
<td>7%</td>
<td>70%</td>
<td>7%</td>
</tr>
<tr>
<td>22</td>
<td>5%</td>
<td>14%</td>
<td>13%</td>
<td>68%</td>
</tr>
</tbody>
</table>

Table 1. Confusion matrix that shows the percentage results after classifying using 2 different predefined filter banks, type 0 and type 2. The indexes $ij$ denote the filter bank chosen in horizontal and vertical directions, respectively. Number of zones was 15, using 3 divisions and 5 sectors.

5. Coding Experiments: Results and Discussion

In order to evaluate the performance of our automatic filter selection scheme, we first manually classified a test set of image blocks. Then, the same test set was classified using automatic filter selection. Using a set of 2 filter banks for each direction, for a total of 4 classes, the automatic classification results in an average total accuracy of 67.3% relative to the manual classification. Considering the fact that the manual classification may vary from person to person, this must be regarded as a good result. The difference between the reconstructed images filtered by manually and automatically selected filter banks, is almost imperceptible. One reason is that the feature vectors of the incorrectly classified blocks lie near the class borders.

In Table 1 a confusion matrix is shown that indicates the nature of these misclassifications. The indexes $ij$ denote the filter bank chosen in horizontal and vertical directions, respectively. Given one class in the first column, the corresponding rows indicate how many of the blocks manually classified to this class, that were automatically classified to the classes given in the top row. This classification example was made using 3 divisions and 5 sectors, the sector borders were placed at 12°, 30°, 60° and 78° in proportion to the x-axis.

The results above do not necessarily prove that the performance of our automatic adaptive subband coder is superior to a subband coder using fixed filter banks. To see that it actually leads to visibly better reconstruction, several images were coded and reconstructed using the automatic adaptive coder. The same images were also coded and reconstructed using nonadaptive coders with different filter banks. Then the resulting images were compared with respect to the visible appearance of ringing and blocking effects. The experiments show that these effects are clearly reduced in most of the images coded by the adaptive coder.

Figure 5 shows an example of an image coded using the different subband coders, all at 0.35 bpc. At the top, the original image is shown. After coding with a fixed filter bank with short impulse response, the reconstructed image looks like the one below the original. Distinct blocking effects can be observed at the sloping edges. Below this image, the image coded with a fixed filter bank with long impulse response, is shown. The ringing effects are visible around the edges and close to the top border of the image. This phenomenon is due to the fact that we are using circular convolution, and that there is a discontinuity at the opposite image borders. At the bottom we see the image coded with the adaptive scheme using the 2 filter banks above. The zone structure used, is the same as previously described. From this we can confirm the assumption that use of our adaptive coding scheme leads to a joint reduction of ringing and blocking effects in the reconstructed image.

1232
6. Summary and Conclusion

In conclusion, we have devised a coding system that provides fully automatic filter bank selection for fixed local image areas. Compared to subband coders using fixed filter banks, coders using this adaptive scheme provide coding results with superior quality.

References


Appendix A

<table>
<thead>
<tr>
<th>Filter bank 0</th>
<th>Filter bank 1</th>
<th>Filter bank 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_0 )</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>0.100</td>
<td>0.250</td>
</tr>
</tbody>
</table>

Some additional coding experiments were done using a set of 3 different predefined filter banks. The result was a slight visible improvement from using 2 filter banks.

For all images, the amount of side information is constant and almost negligible. Using 2 different filter banks for each filtering direction, only 0.002 bits per pixel is used for the side information.

Figure 5. Detail from the picture house coded at 0.35 bpp. Top (1st): Original image. 2nd: Reconstructed image after filtering/coding using filter bank no. 0. 3rd: After using filter bank no. 2. Bottom (4th): After using the adaptive scheme.
A Region-Based Discrete Wavelet Transform

Herjan J. BARNARD, Jos H. WEBER, and Jan BIEMOND

Information Theory Group, Department of Electrical Engineering, Delft University of Technology,
P.O. Box 5031, 2600 GA Delft, The Netherlands.
E-Mail correspondence address: weber@et.tudelft.nl

Abstract. Region-based image coding methods divide the image into regions and code the segmentation information and the texture separately. We introduce a region-based discrete wavelet transform (RBDWT) to code the texture. This RBDWT has the advantage that it enables to distribute the available bit rate over subbands as well as regions. First, we show the principles of the RBDWT and explain the details for implementation. Then, the performance of the new method is evaluated for two test images. The subjective and objective coding performance of the RBDWT are compared with the standard discrete wavelet transform (DWT). The signal-to-noise ratio is slightly lower for the new method, but the ringing artefacts are smaller and the edges are sharper, yielding a good visual quality. Therefore, the new image coding technique is at least competitive to the standard DWT coding scheme. Finally, suggestions are given for further research in this new direction.

1. Introduction

The discrete wavelet transform (DWT) is used in many areas of signal processing, in particular for the compression of digital images (e.g. [1]). The DWT divides the information in several time-frequency regions, called subbands here. Thus, the available bit rate can be distributed over the subbands according to their relative importance for the human visual system.

Region-based, or contour-texture, coding techniques have a different approach to coding images than the DWT. The idea is to segment the image into regions which correspond, as much as possible, to the objects in the image [2]. Now, the available bit rate is distributed over the segmentation information and the various region contents according to their relative importance. The segmentation information is coded with a chain-code. The texture information can be transmitted by the coefficients of a 2-D polynomial that is fitted to the segment data. A drawback is that the regions have to be rather small to be represented with reasonable accuracy by a polynomial. This leads to high transmission costs for the segmentation information. Another method for texture coding is the use of generalized orthogonal transforms, introduced in [3]. Then, the basis images of the transform are calculated by orthogonalizing standard basis images with respect to the shape of the region. Also for this method the regions are usually rather small, because of the computational burden of orthogonalization.

In this paper we introduce a region-based discrete wavelet transform, and consider its application to image coding. It combines the two coding principles mentioned above in that its bit allocation algorithm is able to distribute bits over subband regions, such that important features in the spatial as well as the frequency domain can be represented accurately. In order to do this, the image is divided into regions, which ideally correspond to the main objects in the scene. Then, the texture is coded by the DWT, which has low complexity and is perfectly able to represent the remaining, less important, edges in the scene. Therefore, the regions may be rather large, thus reducing the coding overhead of transmission of the segmentation information.

The standard DWT as described in literature cannot be used to code the texture. The reason for this is that a d-level decomposition requires the regions to consist of square blocks of size $2^d \times 2^d$. In general, such regions will not correspond to the shape of the objects in a scene. However, the efficient signal extension method described in [4] enables us to apply a DWT on arbitrarily sized regions. To distinguish this implementation of the DWT from the standard DWT, we use the name region-based DWT (RBDWT).

2. The Implementation of the RBDWT

We assume that the segmentation into regions has already been performed by a proper split-and-merge method (e.g. [5]), taking into account that the texture is coded by the RBDWT, that is, a division into regions according to the main edges. The standard (dyadic) 2-D discrete wavelet transform for images decomposes an image into four subbands (called LL, LH, HL, and HH), after which this splitting can be repeated on the LL subband. In Figure 1a such a decomposition is shown.

Now we wish to decompose the image into four subbands, where each subband is subdivided into regions as well. Four ‘child’ regions of decomposition level 1 correspond to one ‘parent’ region of the original image. The division of the subbands into regions has the restriction that each parent region in the original image must be perfectly reconstructible from the four corresponding child regions in the subbands. This means that the total number of subband coefficients in the four child regions must be equal to the size of the parent region. This can be achieved by the method as demonstrated in Figure 1bcd. The segmentation mask of the original image is subdivided into 2x2 blocks and for each block the upper-left part is put into the LL band. In the same way, the upper-right part of each block falls in the LH band, and so on. The procedure can be
viewed as an algorithm to downsample in four ways a segmentation by a factor 2 per dimension. Note that the two congruent regions in the centre are divided over the subbands in a different way, because of the different parity of the pixel coordinates.

For the filtering and downsampling, we apply the efficient signal extension method as explained in [4], which preserves the perfect reconstruction property. We filter per segment line, where a segment line is defined to be a horizontal (or vertical in the case of column filtering) sequence of connected pixels in the same region. The segment lines can be divided into four classes that each require their own implementation of the filtering and downsampling. The four classes are illustrated in Figure 2. They are distinguished by the parity of the column number at which the segment line starts and by the parity of the segment line length. The pixel and coefficient values are denoted by a letter in a small square and downsampling is illustrated by using empty squares. In each class only the analysis filtering is shown, where H and G are the low-pass and high-pass filter, respectively. The dark rectangles in the upper and lower line of every block represent the signal and the subband, respectively, whereas outside the dark rectangles the extensions are shown. The circles indicate the special samples defined in [4]. The coefficients in a grey background do not have to be transmitted, i.e., the total number of subband coefficients is equal to the number of pixels in the original signal.

A segment line of length one requires a different processing. If this sample is situated at an even-numbered position, then it is filtered by the low-pass filter to produce a coefficient in the low-pass subband. This amounts to just copying the sample value into the low-pass subband. However, if the segment line of length one is located at an odd-numbered position, it will always give a zero after the extension, filtering, and downsampling, because of the zero mean of the high-pass filter. Thus, we will have lost some information. Our solution to this problem is to keep the sample value and copy it directly into the high-pass subband. If such a sample contains low-pass information, we call the subband coefficient a single. In other words, singles are low-pass coefficients that are positioned in a high-pass subband. If the row filtering is followed by the column filtering, then singles appearing in the HH subband originate from row filtering only. This means that at the decoder the column filtering must precede the row filtering to preserve the perfect reconstruction property.

The singles are quantized separately and the variance calculation and bit allocation of the remaining subband coefficients is performed without the singles. This guarantees that the singles do not disturb any statistics of the remaining regions. A more detailed description of the implementation can be found in [6].

3. The Coding Scheme

The image coding scheme we used in the experiments is shown in Figure 3. At the encoder, the image is segmented into several regions of similar variance and grey-value. The number of regions depends on the image content and is a compromise between accurate segmentation and transmission costs for the contours.

For the filter bank we used the biorthogonal wavelet filter pair of lengths 9 and 7 [7], and we applied a symmetric boundary extension.

The bit allocation was performed with the algorithm of [8], which minimizes the distortion in mean square error sense for a specified rate. The variance of every subband region is used to control the bit allocation process. The vari-

Figure 3 The region-based discrete wavelet transform image coding scheme. The segmentation controls the transform (DWT), the variance calculation (V), and for the low-pass subband also the quantization (Q). The transmitted data consists of the segmentation information, the variances, and the quantizer output. The bit allocation (BA) is repeated at the decoder.
ances are transmitted to enable the decoder to reproduce the bit allocation process of the encoder.

The quantization method was PCM for the coefficients in all high-pass subbands. The coefficients of the low-pass subband were DPCM quantized per region: the first subband coefficient of a region was PCM quantized with 6 bpp and all other coefficients were predicted with 3, 2, or 1 neighbouring coefficients depending on the number of neighbours in the upper left area of the coefficient. The prediction coefficients were fixed. The singles were coded by the average coefficient value of their corresponding region in the low-pass subband.

We applied uniform threshold quantizers which were designed for generalized Gaussian distributions, with optimized shape parameters. The quantizer output was entropy-coded in our implementation.

Two test images were used: Camera-man and Vectra. Both images are sized 256*256 pixels and contain sharply defined objects, for which region-based coding methods are most suited. The segmentation of the Camera-man image uses 11 regions, whereas the Vectra image has been segmented into 52 regions. These segmentations were selected manually from several different segmentations using the split-and-merge algorithm of [5]. For a complete coding scheme, the segmentation process should be performed automatically, taking into account the texture coding by the RBDWT. However, here we are mainly interested in the evaluation of the texture coding by the RBDWT. Figure 4 shows the Camera-man image, its segmentation, and the 10-band decomposition of the segmentation mask. There are 199 singles. Also the Vectra image was decomposed into 10 bands, which yielded 481 singles.

For bit rates of 1.0 bpp and higher the differences between the two methods will not be visible, because then the transmission costs of the segmentation information become relatively less important and the subband coefficients will be quantized with similar accuracy. For low bit rates that are only slightly above the number of bits needed for transmitting the segmentation information, the costs of the chain code will completely consume the channel capacity, such that for the texture only one quantized low-pass value per region remains. Therefore, we choose a bit rate of between ca. 0.20 and 0.40 bpp.

4. Comparison To the Standard DWT

For the comparison of the RBDWT to the standard DWT, we expect four advantages of the RBDWT. Firstly, after the segment boundaries have been determined, the filtering takes place inside relatively homogeneous regions. Consequently, the high-pass subbands are expected to contain less energy than for the standard DWT, since filtering across sharp edges is avoided. Secondly, the DPCM coding of the low-pass subband will have a better prediction, because the spatial correlation inside each region will be higher for the low-pass subband as a whole. In the third place, quantization errors do not influence neighbouring regions and the ringing artefacts near sharp edges will be much less than for standard DWT. Finally, the bit allocation should be able to favour important regions in the image, as bits can be assigned to each subband region separately. Disadvantages of the RBDWT in comparison to the standard
DWT are the time-consuming generation of the segmentation and the lower bit rate available for the transform coefficients because of the segmentation information.

We compared the average variances of the subbands of the RBDWT to the subband variances of the standard DWT, where we assumed zero mean. For the low-pass subband we calculated the prediction error variance, which decreased enormously, as expected. However, most of the high-pass subbands have an unexpected higher variance. The explanation for this effect is most probably that the fourth type of extension is not preserving the statistics, but is increasing the variance instead.

Next, we show in Figure 5a/b the image quality obtained by the RBDWT and the standard DWT for comparable bit rates, when the contour coding is assumed to require 1.5 bits per link. Note that the face is reconstructed very well by the RBDWT method. This accords with our expected advantage of the RBDWT in that it is able to divide its bits not only over the subbands, but also over the spatial domain. Further, the edges are much sharper for the RBDWT and the ringing artefacts are smaller, as had been expected. In general, the standard DWT yields a more natural texture inside the regions. The same holds for the Vectra image.

Further, the peak signal-to-noise ratios were compared. For the Camera-man image the RBDWT and standard DWT yielded, 25.2 dB and 26.4 dB, respectively, and for the Vectra image 28.2 dB and 29.3 dB, respectively. That is, in both cases the standard DWT has a slightly better PSNR.

Our experiments do not allow us to reach firm conclusions yet. However, we may say that the RBDWT seems to be a competitive alternative to the standard DWT in terms of subjective visual quality. The sharpness of the edges and the small ringing artefacts are often considered more important than the lack of texture in some segments.

5. Conclusions and Further Research

In this paper, we have presented a new texture coding method for contour-texture coding techniques which is called the region-based discrete wavelet transform (RBDWT). It is able to distribute the available bit rate over subbands as well as regions. Although the standard DWT has a higher PSNR than the RBDWT, the subjective coding performance of the RBDWT has emerged to be quite competitive. In particular, the edges are sharper and ringing artefacts are smaller. Further research should be directed to the optimization of the segmentation and quantization processes. Finally, the possible application of the RBDWT in a video coding scheme deserves further research.

References

A Study on FIR Filters for Subband Coding of Images

Hugues BENOIT-CATTIN, Atilla BASKURT, Françoise PEYRIN, Robert GOUTTE

Ultrasound and Signal Processing Lab - CREATIS URA CNRS 1216
INSA Lyon, 69621 Villeurbanne, France

Abstract. This paper presents a study on FIR filters mostly used in classical subband coding scheme (CQF,QMF, Wavelets). Filters will be first compared using their inherent characteristics, and then different parameters are computed to evaluate the filters performance. Regarding the results obtained, it appears that, in a compression approach, all the filters studied whatever their regularity, phase characteristic and selectivity, would lead to the same kind of performance.

1. Introduction

Subband coding of images introduced by Vetterli [1] and Woods [2] is a well known approach in image data compression. This approach uses a tree-structured analysis/synthesis scheme. The basic idea of such scheme is to split the frequency band of an image into subbands, to code each subband according to the associated subimage statistics, and then to perform reconstruction by interpolating the decoded subimages.

In a subband coding application, one could choose the coding scheme to be applied on each subband, the number of decomposition to proceed, the convolution method, and the kind of filters used to perform the decomposition and the reconstruction. From an image compression point of view, this paper proposes a study which compares some of the mostly used filters, in order to underline possible advantages and disadvantages of these filters.

In section 2, a typical subband coding scheme will be detailed; in section 3, the different kind of filters and their inherent characteristics are presented; section 4 contains the definition of the different parameters chosen to compare the filters; the main results obtained are given and discussed in section 5.

2. Subband Coding Scheme

A classical subband coding scheme is represented Figure 1. The original image is first decomposed into two bands by low and high filtering and subsampling by two rows, and then doing the same on columns to get four subimages (LL₁, LH₁, HL₁, HH₁). For example HL₁ is the original image decomposed by a High pass filter on rows and then by a Low pass filter on columns.

After coding/decoding, the subimages are interpolated by a symmetrical process in order to reconstruct an image which is in general an approximated version of the original one. Indeed, either the analysis/reconstruction process or the coding/decoding one could introduce a controlled distortion.

3. Studied Filters

In the development that follows, we have deliberately limited our study to orthonormal filter banks: orthonormal means that the reconstruction filters are designed by just conjugating and translating the decomposition filters. In fact, only the low pass filter Bs is needed to get the three others. Different ways of determining Ha, Br, Hr have been proposed: Quadrature Mirror Filters [3], Conjugate Quadrature Filters [4], Wavelet filters [5],[6].

The coding/decoding process introduces constraints on the choice of Bs filter:
- minimization of magnitude distortion;
- linearity or minimization of phase distortion;
- high selectivity in the frequency domain;
- regularity around the Nyquist frequency [8];
- short length to minimize the bordering effect in the analysis/reconstruction process.

We have compared four different groups of filters for the analysis/reconstruction process:
- Johnston QMF[3]: These filters have been designed for achieve perfect reconstruction in subband coding. But, a small magnitude distortion is introduced in the real implementation of these filters;
- Smith/Barnwell CQF [4]: They have been developed for speech subband coding. Very selective but not "smooth" in the frequency

![Figure 1. Subband coding scheme.](image-url)
domain, they are designed by using classical filter design algorithm like Parks Mc Clellan one [7].
- Daubechies Orthogonal Wavelets Bases Filters [5]: They have been introduced by associating orthonormal wavelets bases with multiresolution analysis. Filters designed this way are CQF with high regularity (smoothness) and low selectivity.
- CQF designed by Rioul [8]: These filters are "between" Daubechies and Smith/Barnwell filters in terms of regularity and selectivity. They are designed using an algorithm [9] very closed to Parks Mc Clellan one in which a new constraint concerning the regularity (number of zeroes at z=1) is introduced.

Filters with approximate linear phase and minimum phase have been studied in the case of CQF and Daubechies filters.

<table>
<thead>
<tr>
<th>Filter Abbreviation</th>
<th>Transition Band</th>
<th>Regularity (zeros at z=1)</th>
<th>Phase Characteristic</th>
<th>Perfect Reconstruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sm min</td>
<td>0.446</td>
<td>0</td>
<td>minimal</td>
<td>yes</td>
</tr>
<tr>
<td>Sm lin</td>
<td>0.446</td>
<td>0</td>
<td>linear</td>
<td>yes</td>
</tr>
<tr>
<td>Db min</td>
<td>0.595</td>
<td>4</td>
<td>minimal</td>
<td>yes</td>
</tr>
<tr>
<td>Db lin</td>
<td>0.595</td>
<td>4</td>
<td>linear</td>
<td>yes</td>
</tr>
<tr>
<td>Rl min</td>
<td>0.460</td>
<td>2</td>
<td>minimal</td>
<td>yes</td>
</tr>
<tr>
<td>Rl lin</td>
<td>0.460</td>
<td>2</td>
<td>linear</td>
<td>yes</td>
</tr>
<tr>
<td>Jn</td>
<td>0.480</td>
<td>0</td>
<td>linear</td>
<td>nearly</td>
</tr>
</tbody>
</table>

Table 1. Filters inherent characteristics.

We reduced our study to eight tap filters which allows to limit bordering effects and computational time and leads to reasonable selectivity. From these taps, we extract the main characteristics of the filters as their normalised transition bandwidth, their regularity given in number of zeroes at z=1, their phase characteristics. More, we have tested their capability of perfect reconstruction. All these features are given in Table 1. The filters abbreviations are the following: Sm, Db, Rl, Jn, for Smith/Barnwell, Daubechies, Rioul, Johnston respectively, _min for minimum phase characteristic, and _lin for approximate linear phase characteristic. The transition bandwidth corresponds to the band between the cut-off frequency and the beginning of the stop band at -40 dB.

![Figure 2. Attenuation in dB of the studied linear low pass filters.](image)

On Figure 2, we have represented the attenuation in dB of four studied kind of filters (Sm lin, Db lin, Rl lin, Jn). In the pass band (which is not represented) the filters have the same response. In the stop band, the Daubechies filters which is the less selective has the greatest attenuation with no oscillation. On the contrary, the Smith/Barnwell filter which is the most selective one, presents many oscillations, and at the Nyquist frequency the attenuation is not infinite (-40 dB). Between this two extremal filters, there are the Rioul filter and the Jonshon QMF.

![Figure 3. Zoom of the frequency response near the Nyquist frequency of the studied linear low pass filters.](image)

The regularity is equivalent to the number of time that the successive derivatives of the frequency response are zeroes at the Nyquist frequency (number of zeroes at z=1). On Figure 3, one can see that the more regular the filter is, the smoother the frequency response in the stop band is, and the less oscillations there are. Daubechies filters have the maximum regularity (4), whereas the Smith/Barnwell and the Jonshon filters have a zero regularity (therefore their frequency response is different from zero at the Nyquist frequency). The frequency responses of these two filters mainly differs by the number of oscillations, the amplitude of the oscillations, and the Nyquist frequency slope.

4. Parameters of Comparison

The subband coding scheme (Figure 1) was applied on "lena" of size 512x512 pixels coded on 8 bits. (Filters taps have been normalised to have their sum equal to "1"). We used a classical convolution process to perform the decomposition/reconstruction; the bordering effect of (L-1) pixels is kept: this way, no filter could take advantage of their phase characteristic. After one decomposition, all the subimages (L1, L1, H1, H1) have been quantized to the nearest integer value.

On these subimages, many parameters are computed in order to evaluate the different filters:
- the subimage dynamic important for quantization. Results are given Table 2.
- the entropy (H) has been computed from the quantization levels using the formula (1):

\[
H = \sum_{i} (-P_i \log(P_i))
\]  

(1)

where \( P_i \) (ni/N) is the probability of the quantization level i, ni the number of pixels quantized at level i, N the total number of pixels, and m the number of quantization levels. An average of the entropy \( H \) in the subbands is given by:

\[
H_m = (H(\mathrm{L}_1) + H(\mathrm{L}_1) + H(\mathrm{H}_1) + H(\mathrm{H}_1))/4
\]  

(2)

\( H_m \) informs us on the residual redundancy between subbands. Results are given Table 3.

1239
the Correlation length computed the same way on rows and on columns of each subimage gives an indication on the block size to choose for quantization, i.e. vector quantization of the subimages. Formula (3) gives the correlation on the row \(i\) (\(CL_i(k)\)) for a distance of \(k\) pixels:

\[
CL_i(k) = \frac{\sum_{j=1}^{N-4} X(i,j) \times X(i,j+k)}{\sum_{j=1}^{N} X(i,j)^2}
\]

where \(X(i,j)\) is the grey level of the pixel \((i,j)\), and \(N\) the number of columns. \(CL_i(k)\) is computed from \(k=1\) to \(k=K\) with \(CL_i(k) < T\), where \(K\) is the correlation length on the row \(i\) and \(T\) is a given threshold. Here we chose \(T\) equal to 0.95; this threshold is high but is a good guarantee to have a stationary image along \(k\) pixels. Results are given Table 4.

Three others parameters based on SNR are computed:
- reconstruction with all the subimages without quantization, which shows the magnitude distortion only due to the filtering process;
- reconstruction with \(LL_1\) and \(HL_1\), which are the most entropic subimages;
- reconstruction with only \(LL_1\) without any quantization, which characterizes the capability of the filter to concentrate the energy in the low frequency subimage \(LL_1\).

Results are given Table 5.

5. Results

The dispersion of the dynamic of the subimages (see Table 2) obtained after one decomposition is very low. The only differences appear on the subimage \(HL_1\), where the dynamic differs from minimal phase filters and external phase filters. But this difference does not affect the grey level histograms: these histograms have the same shape than the original image one for \(LL_1\) subimage, and a typical laplacian shape [6] for the three other subimages (\(LH_1\), \(HL_1\), \(HH_1\)).

<table>
<thead>
<tr>
<th>Sm_min</th>
<th>Sm_lin</th>
<th>Db_min</th>
<th>Db_lin</th>
<th>RI_min</th>
<th>RI_lin</th>
<th>Jn</th>
</tr>
</thead>
<tbody>
<tr>
<td>(LL_1)</td>
<td>267</td>
<td>259</td>
<td>263</td>
<td>250</td>
<td>265</td>
<td>258</td>
</tr>
<tr>
<td>(LH_1)</td>
<td>82</td>
<td>129</td>
<td>82</td>
<td>134</td>
<td>83</td>
<td>131</td>
</tr>
<tr>
<td>(HL_1)</td>
<td>194</td>
<td>112</td>
<td>109</td>
<td>115</td>
<td>104</td>
<td>113</td>
</tr>
<tr>
<td>(HH_1)</td>
<td>49</td>
<td>43</td>
<td>51</td>
<td>42</td>
<td>49</td>
<td>44</td>
</tr>
</tbody>
</table>

Table 2. Subimages dynamic (Lena dynamic:245).

With all the filters, the subimages obtained are similar by their dynamic and their histogram, thus it is not surprising to find no difference in the subimages entropy (Table 3). All the \(LL_1\) subimages have an entropy very close from the original image (around 7.6 bits), whereas the other subimages entropy is lower: around 3.5 bits for \(LH_1\), around 3.7 for \(HL_1\), and around 3.0 for \(HH_1\). The average of the subimages entropy which is nearly constant whatever the filter, is of course lower than the original image entropy (4.5 bits).

In Table 4, we have only represented the average and the standard deviation of the correlation length average obtained with all the filters. Again one can see that the deviation is very small. For the \(LL_1\) subimage the correlation length corresponds to the original one divided by two, consequence of the subsampling by 2. There is no correlation between pixels for the subimages \(HH_1\) on rows and columns, \(HL_1\) on the rows and \(LH_1\) on the columns, and the correlation is very small (around 1.2) for \(LH_1\) on the rows, and \(HL_1\) on the columns.

A block quantization could be applied on \(LL_1\) subimages, but to be efficient the block size must be adapted to the correlation length. At the opposite on the high frequency subimages (\(LH_1\), \(HL_1\), \(HH_1\)), where the correlation length is close or equal to one, the best quantization will be scalar (block size equal to one).

<table>
<thead>
<tr>
<th></th>
<th>(LL_1)</th>
<th>(LH_1)</th>
<th>(HL_1)</th>
<th>(HH_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg</td>
<td>5.293</td>
<td>1.159</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>StDev</td>
<td>0.046</td>
<td>0.009</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4. Subimages Correlation Length Average (Lena: 10.335/rows, 17.623/columns).

Regarding Table 5, it appears that QMF which does not afford perfect reconstruction, seems to give good results when we cancel the high frequency subimages. Also, it appears that the filters with minimum phase characteristic give results 0.5 dB better than those with approximate linear phase. It is clear that SNR is not a visual criterion; however this mathematical parameter yields to a good measure of the reconstruction quality which seems necessary to compare filters.

<table>
<thead>
<tr>
<th>Filter Abbreviation</th>
<th>SNR with all the subimages</th>
<th>SNR with (LL_1) and (HL_1)</th>
<th>SNR with (LL_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sm min</td>
<td>36.199</td>
<td>32.716</td>
<td></td>
</tr>
<tr>
<td>Sm lin</td>
<td>∞</td>
<td>35.576</td>
<td>32.000</td>
</tr>
<tr>
<td>Db min</td>
<td>36.482</td>
<td>32.878</td>
<td></td>
</tr>
<tr>
<td>Db lin</td>
<td>∞</td>
<td>35.626</td>
<td>32.285</td>
</tr>
<tr>
<td>RI min</td>
<td>36.514</td>
<td>32.986</td>
<td></td>
</tr>
<tr>
<td>RI lin</td>
<td>∞</td>
<td>35.842</td>
<td>32.228</td>
</tr>
<tr>
<td>Jn</td>
<td>57.658</td>
<td>36.259</td>
<td>32.659</td>
</tr>
</tbody>
</table>

Table 5. Reconstructed Image SNR.

The regularity property of the filters does not affect the global quality of the reconstructed image. This could be explain by the fact that the regularity effects take place in the stop band for attenuation greater than 40 dB. However, to show the local artefacts due to the regularity, we present on Figure 4, a local zoom of two reconstructed images obtained with Sm_lin and Db_lin respectively.
In the region of the eye, Db_min introduces more distortion along the edges than Sm lin. This is due to the lower cut-off frequency of the Db_min filter. On the other hand, Gibbs effects appear in homogeneous regions like on the nose when using Sm lin. This is explained by the oscillations of the filter near the Nyquist frequency (see Figure 2 and Figure 3) due to a regularity zero.

The same kind of results have been obtained on different modalities of images, and also with 16 taps filters.

6. Conclusion
This study has shown that in a subband coding scheme using orthonormal filters, the main filters family studied gave the same kind of performance. The influence of the phase characteristic, the regularity, the selectivity are not enough significant to justify an adaptation of the coding process to the analysis/decomposition filters. But, in the case of very specific image modalities (high frequency or very homogeneous images) an adapted choice of the filter will improve the subband coder performances.

The choice of linear phase characteristic filters will be mainly justified by specific convolution process, and specific coding algorithm.

Effective differences, mainly due to the phase characteristic, could be observed between the filters; by convoluting small images with long filters (more than 16 taps), but this is not realistic in a compression approach.

References
Design of Recursive Filter Banks for Subband Coding of Images

Marek DOMAŃSKI, Roger ŚWIERCZYŃSKI

Politechnika Poznańska, Instytut Elektroniki i Telekomunikacji ul. Żelazna 2a, 60-965 Poznań, Poland
Tel/fax: +48 (61) 782 572, E-mail domanski@ietpp.poz.pl, roger@ietpp.poz.pl

Abstract The paper deals with design of recursive filter banks for applications in subband image coding. Considered is optimisation of the design made from the point of view of efficiency of subband coding of images. The results are applicable to still images and intraframe coding of video. Considered are separable lattice filters in polyphase implementations and reversion arrangements. The paper explains why more selective filters can be related to worse coding results. The importance of short impulse responses is proved. Two ways of improving design of recursive filters are suggested. The first approach is based on limiting the maximum pole modulus and minimising the stop-band energy. The second, more promising, approach minimises estimated amount of information related to transients and aliasing.

1. Introduction

Subband coding (SBC) is a well established technique for image data compression used in a variety of applications [1]. The basic idea of SBC is to split up the two-dimensional (2-D) frequency band of an image into subsampled channels which are encoded using techniques accurately matched to the individual signal statistics and possibly to the properties of the visual system. The human visual system has the property of separability, so the filter bank is a very important part of such a system. Design of analysis and synthesis filter banks strongly influences efficiency of the overall SBC system. Therefore it is important to define properly the requirements for the filters. We are going to show that classic design optimised from the point of view of minimum ripple and transient-bandwidth width is insufficient.

The objective of the paper is to select the parameters of IIR filter banks for applications in subband coding (SBC) of images. Of course, there exists an excellent book [5] and a lot of very good papers on filter banks. Nevertheless very few [8,9] of them have considered the problem from the point of view of coding efficiency. We show that the classical quality measures are not proper for the applications considered. Therefore we establish some other criteria useful in image coding applications and we consider efficient IIR lattice filters for polyphase implementations.

2. Separable IIR Filter Banks

We deal only with separable filter banks which are commonly used because of their efficient implementations. As opposed to most works in subband coding of images which use the linear phase finite impulse response (FIR) filters we concern the infinite impulse response (IIR) filters which are preferable because of their lower computational costs. For example, a polyphase implementation of a pair of the 1-D 24-tap FIR QMF needs 12 times more multiplications than the polyphase implementation of a pair of the 1-D 5th-order elliptic filters with the equivalent attenuation characteristic. Moreover, application of wave digital IIR filters [6] guarantees stability, low sensitivity, and small roundoff noise. In order to avoid phase problems, we use recursive systems. The basic idea of a recursive SBC system is to use the opposite directions of data processing in the coder and in the decoder, thus compensating for shifts caused by the filters (e.g., [2,4,7,10]). We consider only separable tree-structured (multi-level) filter banks with half-band filters.

Consider an one-dimensional (1-D) two-band system. The output signal is

\[ Y(z) = 0.5U(z)[F_1(z^2)F_1(z^{-2}) + F_2(z^2)F_2(z^{-2})] + 0.5U(z)[F_1(z^2)F_1(z^{-2}) - F_2(z^2)F_2(z^{-2})], \]

where \( F_1(z^2) \) and \( F_2(z^2) \) are all-pass transfer functions. The all-pass sections create a pair of complementary filters which are called as bireciprocal or power symmetric.

![Figure 2. The two-phase recursive system with all-pass sections](Image)

Substituting \( z = e^{j\omega} \) we get by inspection that under absence of quantization the signal is perfectly reconstructed (i.e., \( Y = U \)) for all \( F_1 \) and \( F_2 \). From this point of view, the design of the filter banks does not seem to be important. Let us verify this unexpected conclusion by an experiment with subband coding of real images.

3. Experimental Results with Elliptic Filters

We use classic elliptic filters in lattice structures which are appropriate for polyphase implementations. Existence of a simple algorithm to design them [12] makes their synthesis easy.

In order to ensure repeatability of the results the encoding technique has been fixed on the method described in
This technique ensures efficient encoding of high-frequency subbands by exploiting the inter-subband correlation.

The results are given in Table 1. They do not support the expectation that a "better" filter (higher stopband attenuation, more narrow transition band) will lead to higher compression.

**Table 1. Coding using elliptic filters**

<table>
<thead>
<tr>
<th>Elliptic filter order</th>
<th>3</th>
<th>5</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum passband attenuation [dB]</td>
<td>29.4</td>
<td>42.9</td>
<td>51.3</td>
</tr>
<tr>
<td>Cut-off frequency</td>
<td>0.150</td>
<td>0.180</td>
<td>0.228</td>
</tr>
<tr>
<td>Compression ratio [%]</td>
<td>21.5</td>
<td>21.7</td>
<td>23.7</td>
</tr>
<tr>
<td>SNR [dB]</td>
<td>42.5</td>
<td>42.2</td>
<td>41.07</td>
</tr>
</tbody>
</table>

We have designed a 9th-order filter that is expected to be better than 3rd- and 5th-order filters. The transition-band width has decreased and the maximum modulus \( r_m \) of the transfer function roots has increased (Fig.3). Increasing of \( r_m \) increases also length of transients [5]. Therefore the transients are longer and it causes an increase in the number of pixels with higher values in the high-frequency subbands. This increase is relatively rapid as length of transients grows exponentially with \( r_m \). The conclusion is that a filter which is better in the terms of the ripple values and the transition-band width may be worse from the point of view of the efficiency of subband coding.

The histograms of high-frequency subbands exhibit high peaks for low values and very moderate values for higher pixel values. Long transients change this histogram increasing the entropy in high-frequency subbands. Therefore transients in the outputs of the filter banks increase the amount of information that is to be coded. In order to avoid this disadvantage we need fast decay of transients in sub-images is a very important issue.

4. **Elliptic Filters**

The use of classic elliptic bireciprocal [6] (power-complementary [5]) pairs of odd-order filters is a very practical solution. The two-phase implementation of such a 5th-order filter needs in average only 1 multiplication and 4 additions per pixel of the input image, while these figures for 7th-order filters are 1.5 and 5.5, respectively. The elliptic bireciprocal filters exhibit only one degree of freedom by their design, i.e., stopband and passband ripples are directly tied to the transition-band width which is symmetric with respect to the frequency 0.25 (when the sampling frequency is normalised to unity). Decreasing the width of the transition bands decreases signal energy in the stopband. It results in more efficient subband coding. Nevertheless this positive effect is partly compensated by decrease in the minimum stopband attenuation \( a_s \). Much worse effects has inevitable growth of the maximum modulus \( r_m \) of the transfer function roots (which is highly nonlinear as discussed in the previous section). It results in longer transients with their negative influence on coding efficiency. For a given \( r_m \), \( a_s \) is growing as the filter order grows. Nevertheless small dynamic range of image and video signals does not need high attenuation. 54dB is the upper reasonable value for 8-bit representations usually used for images. In practice, for highpass filters with transmission zero at \( z=1 \) the values of \( a_s = 40-45 \) dB are quite enough. Therefore the filters orders exceeding 5 are not necessary. Thus we recommend a 5th order filter with only two free coefficients in its transfer function

\[
H(z) = \frac{a_0 + z^{-2}}{1 + a_0 z^{-2}} \frac{a_1 + z^{-2}}{1 + a_1 z^{-2}}
\]

5. **Optimised Filters**

The choice of the coefficients values influences coding efficiency which is related to transient length in a very complicated way. It influences also inter-subband cross-talk and shape of histograms of individual sub-images. Therefore the best values of the coefficients \( a_0 \) and \( a_1 \) can be found only experimentally using test images (Table 2).

Resigning from equiripple design the bireciprocal filters can be optimized according to the minimum bitrate criterion. Assuming a given \( r_m \) we have only one degree of freedom and we are able to find the optimal values of \( a_0 \) and \( a_1 \) experimentally, i.e., simulating the coding/decoding process for each value of the second root pair. Although the obtained filters seem to be worse then elliptic the performance in subband coding is slightly improved (see Table 2).
Table 2. Experimental results for optimised filters

<table>
<thead>
<tr>
<th>x (cutoff frequency)</th>
<th>a_0</th>
<th>a_1</th>
<th>e_r (stopband energy)</th>
<th>SNR[db]</th>
<th>b/pixel</th>
<th>SNR[db]</th>
<th>b/pixel</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.22</td>
<td>0.2855</td>
<td>0.7821</td>
<td>0.0057</td>
<td>41.65</td>
<td>1.783</td>
<td>35.48</td>
<td>0.977</td>
</tr>
<tr>
<td>0.20</td>
<td>0.2102</td>
<td>0.7145</td>
<td>0.0086</td>
<td>42.17</td>
<td>1.750</td>
<td>35.69</td>
<td>0.962</td>
</tr>
<tr>
<td>0.18</td>
<td>0.1602</td>
<td>0.6662</td>
<td>0.0109</td>
<td>42.21</td>
<td>1.734</td>
<td>35.71</td>
<td>0.956</td>
</tr>
<tr>
<td>0.16</td>
<td>0.1240</td>
<td>0.6297</td>
<td>0.0127</td>
<td>42.25</td>
<td>1.724</td>
<td>35.69</td>
<td>0.950</td>
</tr>
</tbody>
</table>

Figure 4. Attenuation of an elliptic filter (dashed line) and its optimized version (continuous line).

6. Redundant Order Filters

We modify the search process by increasing the degree of freedom, i.e., increasing the number of the pole pairs. On the other hand we propose objective criteria which should measure the amount of information added by transients as well as by nonideal separation of subbands. Therefore we avoid simulation experiments with test images which would be very laborious in the case of more degree of freedom.

From the point of view of such filter parameters like transition band width, stopband and passband ripple the filter order is too high. This redundancy gives us the necessary degree of freedom. Of course, here we face the obvious limitation due to the well-known relation between the passband width and the rise time.

Nevertheless we are able to improve the overall system quality by optimisation of the filter bank. Let define the performance index as

\[ E = E_b + E_h + E_a \]

where:
- \( E_b \) - information related to the transients at the image (sub-image) boundaries,
- \( E_h \) - information related to the transients caused by the image content,
- \( E_a \) - information related to the aliasing effect.

Each of the above components is calculated in terms of entropy with respect to the modelled image statistics. For the sake of simplicity some basic properties are assumed:
- first order Markov process,
- equalised histogram,
- fixed correlation between two adjacent pixels.

The numerical data for optimisation are gathered from the basic test images.

6. Conclusions

Efficient subband coding of video needs special design of IIR elliptic filters. The classic design where ripple and transition width are maximised is inadequate to the actual requirements of image and video coding. Much better results can be obtained by minimising entropy of the sub-images related to the subbands.

Acknowledgements

The paper has been prepared under support of National Research Committee (Poland) no. 3 0754 91 01 and 8 S504 002 06

References

On The Choice of Multirate Filter Banks for Image Coding

Patrice ONNO and Christine GUILLEMOT

C.C.E.T.T
4, rue du Clos Courtel
35 512 Cesson-Sévigné FRANCE
Phone: 33 99 12 47 96 33 99 12 45 04
E-mail: onno@ccett.fr guillemo@ccett.fr

Abstract. This paper focuses on the tradeoffs in the choice of both characteristics of the filters used in the multirate filter banks and of the structure of frequency decomposition for image compression purposes. Wavelet and wavelet packet transforms and M-channel uniform filter banks are considered. Among the various filter design criteria investigated, the maximum coding gain criterion proves to be the most relevant for the three kinds of transformations considered. These three transformations are used in a coding scheme where the quantization is optimized in the rate-distortion sense. Both scalar and $D_4$ lattice vector quantization are considered. Comparative results obtained on still pictures are discussed.

1 Introduction

The field of multirate filter banks has progressed considerably over the last decade and has shown promising results in image compression. The underlying theory progressed from two-channel filter banks, and wavelet transforms to multichannel filter banks including modulated filter banks and wavelet packet transforms. These transforms rely on the definition of bases for the expansion of nonstationary signals, and provide a multiscale analysis of signals useful for compression purposes. In the choice of these bases for a given application, there are a number of questions that need to be addressed. In this paper, some of these questions including the choice of the optimization criteria in the filter design but also the tradeoffs between structures of frequency decomposition respectively provided by wavelet transforms, wavelet packets and uniform M-channel filter banks are addressed. Another question addressed here is whether in the case of iterated structures of decomposition, such as wavelet transforms or wavelet packet transforms, the optimization criterion should be applied on the filters resulting from the iterated structure.

In an attempt to answer these questions, a large number of filters are synthesized by using different design criteria such as frequency selectivity, maximum coding gain, aliasing energy, regularity. For each of the considered structure of frequency decomposition, wavelet transform, wavelet packet transform or uniform M-channel filter banks, it turns out that the design criterion optimum for compression both in the rate-distortion sense and in the visual sense is the maximum coding gain. Also, it is shown here that for the maximum coding gain criterion, filters with very close responses are obtained when the design optimization is carried out on the iterated structure of the filter bank.

These various transformations are then integrated in a picture coding scheme. The procedure used for the coding of the subband signals is based on an algorithm minimizing the overall distortion obtained with a specific type of quantizer for a given bit rate. Both uniform scalar and $D_4$ lattice vector quantizers are considered. Picture coding results in terms of SNR and subjective quality obtained with the various optimized transformations are discussed.

2 Wavelet and Wavelet Packet Transforms

The first question addressed here is the best criterion to be used in the design of wavelet and wavelet packet filters, to be applied for video compression. For wavelet filters, it is shown in [1] that trading frequency selectivity for regularity improves the compression efficiency. However, in [2] it is observed that an appropriate choice of transition bandwidth provides, when using the maximum frequency selectivity criterion, filter banks optimum for compression. In addition, it was found in [2] that optimum filter banks were obtained with a faster convergence by using a maximum coding gain criterion expressed as

$$G_c = \frac{1/2(\sigma_G^2 + \sigma_G^2)}{\sqrt{\sigma_G^4 + \sigma_G^4}},$$

(1)

where $\sigma_G^2 = G_0 R_{xx} G_0$ is the variance of the low frequency subband and $R_{xx}$ is the covariance matrix of an order 1 autoregressive signal with a correlation factor $\rho = 0.95$. The resulting filters turn out to have very poor regularity, although their performance is comparable to the performance obtained with maximally regular filters such as the filters introduced in [3]. The questions that arise naturally are 1) whether similar conclusions can
be drawn in the case of wavelet packets, and 2) whether the same results would be obtained if the optimization was carried out on the resulting iterated filters.

Wavelet and wavelet packet transforms implemented with iterated two-channel orthogonal CQF filters are considered here. Methods for synthesizing such filter banks are now well known. One of these methods relies on the design of the product filter \( P(z) = G_0(z)G_0(z^{-1}) \) where \( G_0(z) \) is the analysis lowpass filter. The design constraints can be imposed on the product filter and the respective filters are then obtained by spectral factorization of the product filter \( P(z) \). This method is widely used for synthesizing filters with varying degrees of regularity [1, 4]. The regularity conditions turn out to be conditions of flatness of the filter response for \( \omega = \pi \), and can be reformulated as linear conditions on the coefficients of the product filter.

The method based on spectral factorization is well suited when the design problem can be formulated as linear equations of the product filter coefficients. However, the use of design criteria such as frequency selectivity, maximum coding gain, minimum aliasing energy, degree of phase linearity, ..., involves nonlinear constraints directly on the coefficients of the paraunitary filters [2], and the spectral factorization method cannot be used.

The procedure that we have here is based on the lattice structure representation of the paraunitary filter bank [5] as shown in figure (1) and on nonlinear optimization of objective functions expressed as

\[
\Phi = \int_{\omega_s}^{\pi} |G_0(e^{j\omega})|^2 d\omega
\]  

(2)

for the frequency selectivity criterion, where \( \omega_s \) is a parameter controlling the transition bandwidth and the frequency selectivity. When varying degrees of regularity are added to the frequency selectivity or to the maximum coding gain, the objective functions to be minimized are given by respectively

\[
\Phi = \int_{\omega_s}^{\pi} |G_0(e^{j\omega})|^2 d\omega + \lambda \left[ \sum_{l=0}^{r} \frac{\partial^l |G_0(e^{j\omega})|}{\partial \omega^l} \bigg|_{\omega = \pi} \right],
\]

(3)

and

\[
\Phi = -\sigma_0^2 + \lambda \left[ \sum_{l=0}^{r} \frac{\partial^l |G_0(e^{j\omega})|}{\partial \omega^l} \bigg|_{\omega = \pi} \right],
\]

(4)

with \( r = 0, ..., L/2 - 1 \), where \( L \) is the length of the filter. However to have a better control of the characteristics of the filters generating the final subbands to be coded, the optimization criterion can be applied on the filters resulting from the iterated structure. When using the maximum coding gain criterion on the iterated structure, the objective function to be maximized is then expressed as

\[
G_c = \frac{1}{M} \sum_{i=0}^{M-1} \sigma_i \prod_{i=0}^{M-1} (\sigma_i)^{1/M},
\]

(5)

where \( \sigma_i \) is the variance in the \( i \)th frequency subband computed by

\[
\sigma_i^2 = h_i^T R_{xx} h_i,
\]

(6)

with \( h_i \) is the equivalent filter for the \( i \)th branch. The filter \( (h_i)_{i=0} \), for the lowpass frequency subband, obtained after \( J \) iterations can be expressed as

\[
H_0(z) = G_0(z^{2^J}) \ast G_0(z^{2^J-1}) \ast \ldots \ast G_0(z^2) \ast G_0(z)
\]

(7)

The filters obtained by the two methods of optimization have very close frequency characteristics. The similarity increases with the length of the filters. An example for filters of duration 12 is shown in figure (2). A large number of filters with different characteristics of frequency selectivity, regularity, coding gain, aliasing energy have been designed and used in both wavelet and wavelet packet structures.

The wavelet packet concept uses an adaptive tree structure and allows to best match the decomposition to the signal nonstationary characteristics. The principle of adaptive tree structures consists in first finding all possible wavelet packet bases and then in finding the best basis according to a minimum distortion criterion by a pruning algorithm [6]. This is carried out as one step of two nested optimization processes involving also the evaluation of the best quantizer in each subband described in section 5.

3 M-Channel Filter Banks

The iterated structure inherent to wavelet transforms and to wavelet packets leads to filters of large duration which may bring ringing artifacts after quantization and add increased delays. On the other hand, the wavelet packet structure allows to tailor the structure of the decomposition to the signal content. It is thus interesting to determine whether fixed transformations can provide
comparable results in terms of SNR and subjective quality. Here, we use cosine modulated exact reconstruction filter banks [7, 8] synthesized following a procedure described in [9] allowing to derive solutions of arbitrary length and for an arbitrary number of channels.

Considering the modulation function given by \( c_{k,n} = 2 \cos((2k + 1)\pi(n - \frac{\alpha}{2})/2M - (2k + 1)\frac{\alpha}{2}) \), \( M \) is a prototype filter of arbitrary length of the form \( m \cdot M - R \), where \( M \) represents the number of channels and \( m \) and \( R \) are arbitrary integers, can be decomposed into \( 2M - 20 + 1 \) polyphase components as

\[
H(z) = \sum_{q=0}^{2M-20} z^{-(N+\theta)}z^{-\theta} G_q(z^{2M})
\]

(8)

where \( \theta \) takes the values:

<table>
<thead>
<tr>
<th>( M ) even</th>
<th>( \theta = 0 )</th>
<th>( \theta = 0.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M ) odd</td>
<td>( \theta = 0.5 )</td>
<td>( \theta = 0 )</td>
</tr>
</tbody>
</table>

and where the polyphase components \( G_q(z) \) and \( G_{2M-20-q}(z) \) are given by the closed forms

\[
\begin{cases}
G_q(z) = \sum_{p=0}^{(m+1)/2-1} (-1)^{p+1} h(N+q + \theta - 2(p+1)M) z^{-p} + (-1)^{p} h(N+q + \theta + 2pM) z^{-p} \\
G_{2M-20-q}(z) = \sum_{p=0}^{(m+1)/2-1} (-1)^{p+1} h(N-q-\theta-2pM) z^{-p} + (-1)^{p} h(N-q-\theta+2(p+1)M) z^{-p}
\end{cases}
\]

The parameter \( \alpha \) depends on the parameter \( \theta \) and takes the values \( \alpha = 1 \) if \( \theta = 0 \) and \( \alpha = 0 \) otherwise. In the case \( \alpha = 1 \), the polyphase components, for \( q = 0 \) verify

\[
\begin{align*}
G_0(z) &= \sum_{p=0}^{(m+1)/2-1} (-1)^{p+1} h(N+q + \theta - 2(p+1)M) z^{-p} \\
G_{2M}(z) &= \sum_{p=0}^{(m+1)/2-1} (-1)^{p+1} h(N+q + \theta + 2pM) z^{-p}
\end{align*}
\]

The index \( p \) varies in an interval such as

\[
0 \leq \begin{pmatrix}
N+q + \theta - 2(p+1)M \\
N-\theta + 2pM \\
N-q - \theta - 2pM \\
N-q - \theta + 2(p+1)M
\end{pmatrix} < m \cdot M - R.
\]

(9)

It can be shown [9] that in the general case of arbitrary length prototype filters the perfect reconstruction property is guaranteed if the polyphase components verify the relations:

\[
\begin{cases}
\text{for } 0 \leq q \leq M - 1 \\
G_q(z^{-1})G_0(z) + G_{2M-20-q}(z^{-1})G_{2M-20-q}(z) = \frac{1}{2M} \\
G_q(z^{-1})GM(z) = \frac{1}{2M} \text{ if } \alpha \neq 0
\end{cases}
\]

The \( M \) pairs of polyphase components \( \{G_q, G_{2M-20-q}\} \) must be power-complementary pairs and can be designed using two-channel lattice structures as shown in figure (1). The optimization of these filter banks relies on the optimization of the prototype filter, and various design criteria similarly to section 2 have been considered.

4 Quantization

Both scalar and \( D_4 \) lattice vector quantization are considered here. The \( D_4 \) lattice is defined by:

\[
D_4 : \{(x_1, x_2, x_3, x_4) \in \mathbb{Z}^4; x_1 + x_2 + x_3 + x_4 \text{ : even}\}
\]

Both the \( L_1 \) and \( L_2 \) norms for the definition of the vectors are tested. The choice of one norm versus the other may be influenced by the statistics of the subband signals. The \( L_1 \) norm may be more adapted to encode a signal with a Laplacian distribution whereas the \( L_2 \) norm is better adapted for encoding signals with a Gaussian distribution [10]. However no significant difference has been observed. The data is quantized on this lattice using a procedure inspired from the fast algorithm developed in [11] for determining the closest lattice point to an arbitrary point. However, here adjacent hyperspheres or hyperpyramids are also considered in the quantization process in order to have a good tradeoff between the distortion and the data rate needed for the overhead information. Two entropy coders are used. One is used to encode the norm of the vectors, and the other one is used to encode the position of the vectors, or vector labels on a given hypersphere or hyperpyramid. The corresponding distortion in all the subbands is also computed. Scaling factors are applied on the subband vectors in order to vary rate and distortion. This produces a convex hull of points in the rate-distortion plane.

5 Bit Rate Allocation

Step 1: Optimum Data Rate Allocation The first step (and unique step in the case of M-channel uniform filter banks) consists in searching the optimal point of the convex hull of points. The problem is then defined in the best quantizers and in allocating bits from a fixed given budget \( R_d \) so as to minimize the overall distortion \( D \). This is a constrained optimization problem that can be formulated as the minimization of [6]

\[
\min_{q_i} \lambda R_d + \lambda \Delta \text{ with } \Delta = \sum_i d_i^2 \text{ and } R = \sum_i r_i^2 \leq R_d
\]

(10)

where \( d_i^2 \) and \( r_i^2 \) represent respectively the mean quadratic error and the rate allocated in the \( i^{th} \) subband quantized with the \( i^{th} \) quantizer \( q_i^2 \) and where \( \lambda \) is a Lagrange multiplier. As the filter banks used are orthogonal, the minimization of the distortion in the reconstructed signal after synthesis is equivalent to the minimization of the summation of the distortion in all the subbands. The search for the optimum quantizers is carried out in two successive steps.

- For a given \( \lambda = \lambda_i \) in each subband \( i \), the quantizer \( j \) minimizing \( d_i^2 + \lambda_j r_j^2 \) is determined. The corresponding point (distortion, overall bit rate) is determined by using the relations \( D_{\lambda_i} = \sum_i d_i^2 \) and \( R_{\lambda_i} = \sum_i r_i^2 \).
- The value \( \lambda_i \) of \( \lambda \) allowing to reach the desired bit rate \( R_d \) is then determined by the evaluation of \( \max_{\lambda_i} D_{\lambda_i} + \lambda_i (R_{\lambda_i} - R_d) \).
Step 2: Optimum structure of decomposition
In the case of adaptive transformations, namely wavelet-packet transforms, a second step consists in optimizing the number of decomposition levels, in order to minimize the overall distortion, following a tree pruning approach [6].

6 Results and Discussion
The source material consists of still pictures taken from "Mobile and Calendar" and "Flower Garden" sequences. Figure 3 shows comparative results between the three types of transformations. Although all the simulation results cannot appear here, a large number of filters with different length and characteristics have been used. It has been observed that for the three types of transformations considered, the best results in compression, both in SNR and visual quality, are obtained with filters optimized with the maximum coding gain criterion. Modulated filter banks of different length of the form \( mM - R \) give comparable results, for \( R = 0, 1, 2 \). The number of frequency bands is also an important factor. Obviously, the wavelet packet transform gives the best results in terms of SNR, since the transformation is adapted to the signal characteristics. At bit rates in the range 1.3-2 bpp, better results are obtained also visually with the wavelet packet structure. However, at low bit rates (below 1 bpp), the long duration of the equivalent filters resulting from the iteration in the wavelet packet structure brings very annoying ringing artefacts. In our experiment and at a bit rate \( < 1 \) bpp, though the 4x4 modulated filter banks give the poorest performance in terms of SNR, they provide the best visual quality. For example the filter banks \( M=4, L=8, 16 \) outperforms significantly in terms of visual quality the filter banks \( M=8, L=16, 32 \). The last observation concerns the significant improvement both in SNR (around 2-3dB) and in visual quality brought by the lattice vector quantization, and the entropy coding methods of the vectors. A psychovisual weighting of the distortions to improve the subjective quality is under investigation.

References
Image Coding by Block Prediction of Wavelet Coefficients

Roberto RINALDO and Giancarlo CALVAGNO

Dipartimento di Elettronica e Informatica, Università di Padova
Via Gradenigo 6/a, 35131 Padova, Italy
Tel/Fax: +39 [49] 8287735 / 8287699, E-Mail: rinaldo@dei.unipd.it

Abstract. In this paper we present a modified version of a coding scheme for still images that exploits similarities among detail signals in a pyramid multiresolution decomposition of the image. In a pyramid (wavelet) decomposition, the image is transformed into a set of subimages, each corresponding to a certain scale and orientation. In the proposed scheme, blocks of a subimage are predicted from blocks of the same dimension and located in the lowest resolution subimage with the same orientation. When the prediction of a block fails, the subband coefficients of that block are uniformly quantized and entropy coded. The experimental results show that the performance of our scheme is superior, for both visual quality and mean square error, to that obtainable with other popular image coders like JPEG.

1. Introduction

Digital storage and transmission of images calls for sophisticated data compression techniques. Subband coding [1] is a very interesting coding scheme since it accommodates a transmission quality that can flexibly adapt to the available channel capacity or storage requirements. Pyramid (wavelet) decomposition provides a multiresolution representation of signals that has been used for image analysis and coding [2]. The redundancy of the multiresolution representation has been recognized by several authors [3, 4] and it can be exploited to provide a more efficient representation of data. Shapiro in [3] suggests a coding scheme that aims to predict negligible energy regions across scales of a pyramid decomposition by using the inherent similarity among subbands. Pentland and Horowitz [4] employ a pyramid decomposition in connection with vector quantization of subbands coefficients. The conditional probabilities between vector codes in adjacent scales is used to provide an efficient predictive coding algorithm. Although from a different perspective, the idea of exploiting similarities between blocks in images inspires fractal block coders that have been recently proposed in the literature [5, 6, 7].

In a recent work [8] we propose a novel coding procedure that exploits similarities between blocks in detail signals of a multiresolution decomposition of the image. The coding scheme consists of dividing each subimage of the pyramid decomposition into a set of nonoverlapping range blocks. The range blocks are matched against domain blocks of the same dimension taken from lower resolution subimages. The scheme efficiently exploits the redundancy of the multiresolution decomposition since range blocks are predicted from domain blocks. Furthermore, following the scheme for fractal block coders, blocks are multiplied by a constant and possibly isometrically transformed (i.e., reshuffling of the block coefficients is carried out) before matching. Whenever the block matching procedure is not satisfactory with respect to a target coding quality, the block is uniformly quantized and entropy coded.

In this work we modify the scheme of [8] and obtain improved results in terms of coding time and achieved compression ratio for a given reconstruction quality. In particular, we consider a reduced search region for the domain blocks and we use a different quantization strategy for nonpredicted blocks. Entropy constrained uniform quantizers are designed to provide the minimum output entropy for a fixed overall quantization error power. Reconstruction values are optimally computed from the quantized output symbols. Also, a recursive scheme for block type identification is used in the present scheme. The proposed technique performs well with respect to the most recently proposed coding schemes for still images in terms of compression ratio, visual quality and computation time.

2. Pyramid Predictive Coding

In a pyramid subband decomposition [9], the input image \( y \) is first decomposed into four subimages (or subbands) \( y^L, y^{Lh}, y^{Lh}, y^{hh} \), where the pair of superscript letters denotes the row-column filtering operations performed to obtain the subimage. For instance, subimage \( y^{Lh} \) is obtained by lowpass filtering the rows and high-pass filtering the columns of \( y \), followed by a factor two subsampling in each direction. This procedure is then iterated to obtain a multilevel pyramidal decomposition of the image \( y \). Denoting with \( \hat{y}_n \) the input image \( y \), at each level \( n \) the image \( \hat{y}_n \) is decomposed into the four subimages \( \hat{y}_n^L, \hat{y}_n^{Lh}, \hat{y}_n^{Lh}, \hat{y}_n^{hh} \), for \( n = 1, \ldots, n_M \). The result of such a decomposition is a set of subimages which are localized in scale, orientation and space.

In our coding procedure, we predict blocks in the subimage \( \hat{y}_n^h \) (or \( \hat{y}_n^{Lh}, \hat{y}_n^{hh} \)), \( n = 1, \ldots, n_M - 1 \), from blocks of the
subimage $y_{nM}^b$, (or $y_{nM}^h, y_{nM}^b$). The subimages at the lowest resolution $n_M$ are coded independently using PCM.

Each subimage $y_n^b$ (or $y_n^h, y_n$), $n = 1, \ldots, n_{M-1}$, is divided into a set of non-overlapping range blocks $\{r_i\}$, whose size may vary depending on the frequency band: the pool of domain blocks for image $y_{nM}^b$ (or $y_{nM}^h, y_{nM}$) consists of the (possibly overlapping) blocks of the same dimension found in the image $y_{nM}^b$ (or $y_{nM}^h, y_{nM}$). For each range block $r_i$ in the subimage $y_n^b$ (or $y_n^h, y_n$), we find a domain block $d_j$ in the subimage $y_{nM}^b$ (or $y_{nM}^h, y_{nM}$) such that the map $\tau_i$ defined as

$$\hat{r}_i = r_i(d_j) = \alpha_i \left( \mathcal{I}_i(d_j) \right)$$

minimizes the mean square error distance $D(r_i, \hat{r}_i)$ between $r_i$ and $\hat{r}_i$ with respect to the contrast scaling $\alpha_i$ and the four possible rotations $\mathcal{I}_i$ of the block coefficients. A code for the range block $r_i$ is therefore given by the relative coordinates of the domain block $d_j$, an identifier for the symmetry $\mathcal{I}_i$, and the scaling parameter $\alpha_i$.

After all the subband images are predicted from the blocks in the resolution subimages, an approximation to the original image is reconstructed from the subband coefficients.

Unlike other block-based coders, the coding scheme presented here does not give any blocking effect even at low bit rates, as expected from a coding scheme that operates in the subband coefficient domain. Furthermore, for low-pass images, the subimages corresponding to the same orientation have similar spectral content. As a consequence, domain blocks can be searched in regions of small dimension (namely, the subimages at level $n_M$ of the wavelet decomposition) and still provide a good match in terms of the mean squared error distance.

3. Description of the Coder

In this section we will describe in some detail the proposed pyramid predictive coder. In the following, the coder organization is described for $512 \times 512$ images, but it can be easily adapted to images with different dimensions. We consider a pyramid subband decomposition with five levels. The filters used to compute the subband decomposition are 24-tap linear phase nearly orthogonal and quasi-perfect reconstruction FIR quadrature mirror filters [10]. The image is symmetrically extended before filtering [11].

The coding procedure can be conceptually divided into two steps: a block prediction (BP) step and a residual block coding (RBC) step. In the first step we try to predict blocks from the lowest-resolution subimages, as described below. If the prediction for a range block $r_i$ is not satisfactory at this stage, the actual coding of $r_i$ is deferred to the RBC step.

3.1 Block Prediction

To provide initial conditions to the BP step, subband $y_n^b$ is DPCM coded. Differences between adjacent coefficients are coded using an 8-bit uniform quantizer, followed by entropy coding. Subbands $y_n^h, y_n^b, y_n^h$ are coded using a 7-bit uniform quantizer, followed by Huffman coding. For the subbands at level 5, we calculate symbol probabilities by assuming a laplacian distribution for the difference and subband coefficients, in order not to transmit the Huffman tree.

### Table 1. Summary of the block range dimensions.

<table>
<thead>
<tr>
<th>Subband</th>
<th>$4 \times 4$ (RBC)</th>
<th>$4 \times 4$ (BP)</th>
<th>$8 \times 8$</th>
<th>$16 \times 16$</th>
<th>$32 \times 32$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_{4}^{b} + y_{4}^{h}, y_{4}^{b}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$y_{2}^{b} + y_{2}^{h}, y_{2}^{b}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$y_{1}^{b} + y_{1}^{h}, y_{1}^{b}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

As already explained in section 2, subbands from level four to one are partitioned into square range blocks. The block dimensions are tailored to the spectral energy of the subbands: thus, smaller blocks are used for high energy subbands, as shown in Table 1.

For each range block of a given dimension in $y_n^b$ ($y_n^h, y_n^b$), a domain block of the same dimension is searched in $y_{nM}^b$ ($y_{nM}^h, y_{nM}$). Thus, there are 169 possible $4 \times 4$ domain blocks in subband $y_n^b$, 81 possible 8 x 8 blocks and only one $16 \times 16$ block, which covers $y_n^b$ entirely. The $32 \times 32$ range blocks considered for $y_n^b$ ($y_n^h, y_n^b$) are possibly predicted from subband $y_{nM}^b$ ($y_{nM}^h, y_{nM}$), which has dimension $32 \times 32$. Fifteen bits are used to specify couples of $4 \times 4$ domain block addresses, while 32 bits are necessary to specify 5-tuples of $8 \times 8$ domain block addresses.

For each block $r_i$ in $y_n^b$ ($y_n^h, y_n^b$), $n = 1, \ldots, 4$, a block $d_j$ is searched in $y_{nM}^b$ ($y_{nM}^h, y_{nM}$) in order to minimize $D(r_i, \tau_i(d_j))$. We only consider 4 rotations for the symmetry $\mathcal{I}_i$; this has been shown to be sufficient, since nearly no improvement is obtained by considering reflections as well [7]. For each of the four rotations, the optimal scaling parameter $\alpha_i^\circ$ is readily obtained by setting to zero the derivative of $D(r_i, \tau_i(d_j))$ with respect to $\alpha_i$, namely

$$\alpha_i^\circ = \frac{\sum_{l,m} r_i(l,m) d_j(l,m)}{\sum_{l,m} d_j^2(l,m)}$$

Once the domain block $d_j$ and the optimal transformation are found, the mean squared error $D(r_i, \tau_i(d_j))$ is compared to a target value $P$ of the mean reconstruction error for the entire image. If $D(r_i, \tau_i(d_j))$ exceeds $P$, the range block is divided into four smaller blocks, like in the quadtree scheme of [6], and domain blocks of smaller dimension are considered. The smallest allowable dimension is $4 \times 4$.

If the prediction scheme of a $4 \times 4$ block is still not satisfactory, i.e., it gives a distortion $D > P$, the range block will be coded during the RBC step, using a pixel-based coding scheme.

Let us respectively denote with $r_i$ and $d_j$ the range and the (isometrically transformed) domain block. If the optimal value of the scaling parameter $\alpha_i^\circ$ is quantized with an error $\Delta \alpha_i = g_0(\alpha_i) - \alpha_i$, the block mean squared error (mse) increases to

$$P_i(\alpha_i^\circ + \Delta \alpha_i) = P_i(\alpha_i^\circ) + (\Delta \alpha_i)^2 S(\alpha_i),$$

$$S(\alpha_i) = \sum_{l,m} d_j^2(l,m),$$

where $l,m$ denotes the coefficient position inside block $d_j$. Thus, a quantization error $\Delta \alpha$ determines a contribution
to the total mse that depends on the domain block energy. A histogram of $\alpha_i$, where each occurrence is counted $S(\alpha_i)$ times, shows that the best fitting distribution is laplacian, with variance given by

$$\sigma^2 = \frac{\sum_i \alpha_i^2 S(\alpha_i)}{\sum_i S(\alpha_i)}. \tag{2}$$

We use a 6-bit uniform quantizer for $\alpha_i$, followed by entropy coding of the output symbols. At the receiver, the reconstruction value corresponding to the output symbol $i$ is computed as the centroid [12]

$$\alpha_{r,i} = \frac{\int_{a_{i-\Delta/2}}^{a_{i+\Delta/2}} a f_\sigma(a) da}{\int_{a_{i-\Delta/2}}^{a_{i+\Delta/2}} f_\sigma(a) da }, \tag{3}$$

where $\Delta$ is the quantization step and $f_\sigma(a)$ is the Laplacian distribution with variance (2). The scaling parameter $\alpha_i$ is set to zero whenever the variance of the range block is less than 0.5$P$. The blocks corresponding to the same spatial location at finer scales with the same orientation are also tested to determine if their variance is negligible. In such a case, these blocks are not coded at all. The first range block with $\alpha_i = 0$ is encoded with a special symbol indicating the insignificance of all the corresponding blocks in finer scales. The technique is similar to that described in [3, 13].

When $D(r_i, r_d(i)) < P$, the range block $r_i$ is coded with the bits specifying the address of the domain block, 2 bits to specify the isometry, and 6 bits to code the scaling parameter $\alpha_i$.

Additional bits per block are needed to specify the block type, in order to correctly decode the stream of bits relative to each block. For the subbands at level 4, one bit per block is sufficient to distinguish between the BP and RBC alternatives. Each 8 x 8 block at level three can be split into four 4 x 4 blocks or predicted as it is. One bit is sufficient to distinguish between these two possibilities. Only in the case the block is split, one bit for each of the resulting 4 x 4 blocks is needed to distinguish between the BP and RBC alternatives. Thus, additional bits are used only when necessary. The same strategy is applied to classify blocks of subimages at levels two and one of the pyramid decomposition. Starting from the largest allowed block dimension, one bit specifies if the block is predicted as it is or split. In case it is split, the four originated blocks are recursively classified with one bit indicating prediction or further splitting, for the intermediate block sizes, and prediction or RBC, for the smallest block size.

3.2 Residual Block Coding

The block prediction procedure described above can give unsatisfactory results for certain 4 x 4 blocks. In that case, one possible strategy could be to propagate the block splitting and consider 2 x 2 range blocks. We found more convenient to code the individual coefficients inside a block using a uniform quantizer and entropy coding.

It is possible to show that the asymptotic performance of uniform quantizers is optimal for entropy coding, with a quantized output symbol entropy given by [14]

$$H_b = h_s - \frac{1}{2} \log_2 12D, \tag{4}$$

where $h_s$ is the differential entropy of the source and $D$ is the distortion. At high bit rates, the distortion of the uniform quantizer is well approximated by $\Delta^2/12$, where $\Delta$ is the quantization step. We denote by $h_{sk}$ the differential entropy, by $f_k$ the total number of coefficients in the residual blocks, and by $\Delta_\alpha$ the quantization step in subband $k$, $k = 1, \ldots, K$. Using (4) and $D_k = \Delta_k^2/12$ for the distortion in subband $k$, the optimal uniform quantizers are determined by solving the problem

$$\min \quad R = \frac{1}{N} \sum_{k=1}^{K} f_k (h_{sk} - \log_2 \Delta_k), \quad \text{s.t.} \quad E = \frac{1}{N} \sum_{k=1}^{K} f_k \Delta_k^2/12, \tag{5}$$

where $N = \sum_{k=1}^{K} f_k$ and $E = 0.5 P$ is the imposed average distortion. By modeling the coefficients in subband $k$ as Laplacian random variables with variance $\sigma_k^2$, problem (5) can be solved by means of Lagrange multipliers and has the solution

$$\Delta_k = \Delta_0 = \sqrt{12E}, \quad k = 1, \ldots, K.$$ 

The coefficients in the residual blocks are therefore coded by using a uniform quantizer with quantization step $\Delta_0$, followed by Huffman coding of the output symbols. At the receiver, the coefficients are optimally reconstructed as the centroids of the quantization intervals, like in (3), where we suppose a Laplacian distribution for the non-predicted subband coefficients.

4. Results

In this section, we present some experimental results to evaluate the performance of the proposed predictive coder. Original images are 512 x 512 gray-level images, coded with 8 bits per pixel (bpp).

Figure 1 shows the peak signal to noise ratio (PSNR) versus the bit rate used to code the test image "Lenna"
with the proposed coder. For comparison, we reported on the same plot the coding results obtained with the coder presented in [8] (PPC) and with the JPEG coding system. As seen, the new coding method performs better over the entire range of bit rates, with an improvement in PSNR that is almost independent of the bit rate.

Figures 2.a and 2.b show the original and reconstructed image "Lenna" at 0.25 bpp. The visual quality of the reconstructed image is fairly good, even if some artifacts and ringing effects are noticeable near the edges. The total coding time for "Lenna" was about 2 minutes on a Sun SPARC station IPC.

Similar coding time figures were obtained for different bit rates and other images, with a great improvement over the scheme of [8]. Our results confirm the improvement over standard block coders and JPEG in terms of achieved PSNR at a fixed bit rate and visual quality. However, the coding time of our scheme is slightly greater than that required by JPEG.

References