INCREASING QUALITY OF CELP CODERS BY SOURCE-FILTER INTERRELATION USING SELF ORGANISING MAPS

Gökhan Avkaroğulları† *and Tolga Çiloğlu*‡
†Havelsan Inc., R&D Dept., DSP and Telecom. Group, Ankara, Turkey
Tel: +90 312 2873565/161; fax: +90 312 2873568
‡Middle East Technical Univ, Dept. of EEE, 06531, Ankara, Turkey
Tel: +90 312 2102352; fax: +90 312 2101261
e-mail: · avkar@havelsan.com.tr, tolga-ciloglu@metu.edu.tr

ABSTRACT

There are various alternatives for secondary excitation formulation for CELP type speech coders. In this paper we present a secondary excitation codebook generation and search algorithm based on the information derived from the linear prediction filter. Source-filter interrelation is extracted using Kohonen Learning algorithm and filter parameters (LSFs) are clustered according to topographic neighbourhood. For each cluster a secondary excitation shape-gain codebook is generated. Using the class information that current LSFs belong, only the associated codebook is searched. Shape codebooks of size 128 and gain codebooks of size 32 lead to statistically indifferent synthetic voice quality according to listening test when compared to FS1016 CELP coder.

1 INTRODUCTION

CELP (Code Excited Linear Predictive) coder which belongs to linear predictive analysis by synthesis class of speech coders, is one of the most popular speech coding algorithms for bit rates 4800 bits/second to 9600 bits/second. In CELP type speech coders, after the short term characteristics of speech is modelled by the linear prediction filter and long term characteristics are modelled by the pitch predictor (adaptive codebook), the remaining characteristics are modelled by a fixed codebook. So the source to excite the filter to reproduce the speech synthetically, comes from two codebooks. Adaptive and fixed codebooks.

At 4-5 Kbits/second rates, CELP coders efficiently reproduce the unvoiced speech, but the reproduced synthetic voiced speech is noisy and hoarse. To overcome this problem several solutions are proposed: multi-mode excitation, multi-stage codebooks, increasing the size of fixed codebook, decreasing the frame length so increasing the convergence rate of adaptive codebook. The last three methods increase the bit rate. At 4 Kbits/second multi-mode excitation results in good performance [1]. In multi mode excitation either speech is phonetically classified, and for each class a different codebook (glottal pulse, stochastic, single pulse, regular pulse) is used, or all types of codewords are stored in a single codebook and an exhaustive search is performed.

The techniques mentioned above increase the quality, and in this paper some results for further improvement in quality and/or bit rate by using the source and filter interrelation will be reported. The clue about the interrelation comes from some speech recognition techniques where LPC information is used to extract phonetic features and satisfactory results are obtained although no emphasis is put on excitation. It is known that phonetic character of speech determines the character of excitation. This fact is an indcator of the relation between the source and the filter such that phonetic information obtained from filter parameters, in particular LSFs, may be used in determining the source (excitation vector) [2]. In the present work Kohonens Self Organising Maps are used for extracting the required information about the source filter interrelation [3].

2 CLUSTERING OF LSFs USING SELF ORGINIZING MAPS (SOMs)

Kohonens Self Organising Feature Map (SOM) [4] is an on-line vector quantisation technique, which also preserve the similarity information of input vectors. SOMs use, the organisation of brain neurones according to external stimulus being sensed, as a model. In SOMs weights are adjusted from common input nodes to M output nodes, where the output nodes are usually represented as a rectangular grid. Outputs are extremely interconnected with many local connections. Close input vectors (where closeness is a function of the distance function used) are quantised as close nodes on output grid. Similarity/closeness of input vectors is preserved. The density of the code vectors tends to approximate the density of input vectors.

SOM can be used to quantise LSFs while phonetically close LSFs will be neighbours in the map [3]. If LSFs are not only vector quantised but clustered according to phonetic similarity as well, source-filter relation can be extracted. Each cluster can be considered as a phonetic class and associated with its own codebook. This will lead to a multi-mode excitation and each mode has its own trained codebook. In such a coding scheme since

both encoder and decoder know which cluster does the current LSFs belong, the mode information need not be transmitted to decoder.

In our work, we did not put emphasis on transparent quantisation of LSFs using SOMs. We used SOMs just to cluster the LSFs. Training of SOMs is usually as follows:

- Initialize weights to small random numbers
- Present new input
- Compute distance to all output nodes
- Select the node with minimum distance
- Select the neighbourhood of the output node
- Update weights from input to selected output nodes
- Repeat by going to step 2

Definition of neighbourhood is critical in the performance of the SOM. Neighbourhood is a function of time. Best results are obtained when the neighbourhood is selected fairly wide in the beginning and then let to shrink with time. If the radius defined for neighbourhood is selected using the distance function, a good minimum of VQ distortion is achieved. The distance function we used is the one in [5].

We have tried three different SOM approaches to achieve the minimum spectral distortion output map which is not folded. The output map was 16 by 8 (7 bit) with dimension of 10 (no splitting of LSFs). In Classical Kohonen Self Organising Map (CSOM) approach [4] with conscious term, the neighbourhood was determined using the Cartesian distance of nodes on map. The second algorithm which is very similar to CSOM uses the distance function on output nodes to determine the neighbourhood. The last technique we used was the Statistical Kohonen Self Organising Map [6] (SSOM). The worst results are obtained with CSOM both for spectral distortion and folding as shown in Figure 1. Folding of the map makes us unable to extract the neighbourhood relation of LSFs. The second approach resulted better for spectral distortion, but the map was still very folded as can be seen in Figure 2. Using SSOM approach we achieved both the best spectral distortion measure and nearly unfolded map as in Figure 3. Spectral distortion is 3.5 dB for training and 4.5 dB for test sets. These are 3.8 dB for both training and test sets for the first stage (7 bit) of a multi stage VQ that quantise LSFs with transparent quantisation.

In most of the speech coders where phonetic labelling (voiced/unvoiced etc.) is done, acoustic parameters are used for phonetic classification. We only used the short-term spectral information, and need not label the classes. The map is clustered according to distance function used. Plotting the distances of neighbouring nodes enable us inspection of "bubbles" in Figure 4. Bubbles will show us that member nodes of the bubbles are phonetically close nodes. So LSFs with different phonetic character will be in different bubbles. Each bubble is considered as a cluster and we have seven of them.

The performance of the SSOM as a vector quantiser is not as good as the one which is trained using LBG algorithm. Although it is reported that this can be achieved with SOMs[3], we couldn't end up with such a quantiser. Another approach may be the use of "soft competition scheme" proposed by Yair et.al. [7]. We found it hard to implement, due to numerical problems. Even a small change in initial temperature resulted the codebook to merge to same point or a training time that tends to be infinite. We will discuss the use of first stage of a multi stage VQ with transparent quantisation for clustering of LSFs in the conclusion part.

3 SECONDARY EXCITATION CODEBOOK TRAINING

After the LSFs are clustered looking at the bubbles on the map, for each bubble a different gain-shape codebook is trained. The number of clusters does not change the computational complexity of the coder, since clustering is done via a simple look-up table, if the first stage of vector quantiser is used for clustering (which is currently not the situation). The excitation vector is the sum of the outcome of the adaptive codebook and outcome of the fixed codebook, which is the codebook of the phonetic class the current LSFs belong.

We have trained both 64 codeword and 128 codeword shape codebooks with 16 codeword (becomes 32 when sign information is added) gain codebooks for each of seven classes of LSFs. The training sequence is obtained using inverse filtering the speech training set with linear prediction filter and removing the adaptive codebook (self-excitation) contribution. We did the training using initial codebooks which are composed of the codeword that were randomly selected from training sequence. We run the training algorithm many times to achieve small distortion. It is believed that better codebooks would be obtained if splitting technique was used.

4 RESULTS

We used an FS1016 CELP coder for comparison with our scheme. In both configuration a tenth order linear prediction filter is used. The filter parameters are scalar quantised using a 34 bit quantiser. Both configurations use an adaptive codebook with fractional resolution and delta coding as the pitch predictor. In FS1016 coder the secondary excitation is extracted using a 9-bit, stochastic, ternary valued, overlapped codebook. In our scheme after LSFs are calculated and interpolated, the class information of the current frame's LSFs is extracted. The search for secondary excitation is done only in the codebook that is generated for that class. We have compared

the outputs of the coders for both 6-bit and 7-bit shape codebooks in our configuration. We conducted listening tests with 8 listeners, using test speech both from male and female speakers. As an objective comparison basis segmental SNR is used. Using 6-bit shape codebooks resulted 1.5 dB decrease in segmental SNR compared to FS1016 coder. Contrary to high difference in segmental SNR, listening tests show that for male speakers the outcome of coders were statistically indifferent, but for female speakers FS1016 performed better. Using 7-bit shape codebooks resulted 0.8 dB decrease in segmental SNR compared to FS1016 coder. But listening tests show that both for male and female speakers the outcome of coders where statistically indifferent.

5 CONCLUDING REMARKS

Our present work on Code Excited Linear Predictive Coders demonstrated that the filter information can be used in determining the secondary excitation. This information can be used to construct many small secondary excitation codebooks which will result not only decrease in bit rate but also decrease in search complexity. We have shown that for 128 codeword codebooks the performance achieved with FS1016 can be achieved. This means a decrease in bit rate of 267 bits/second and also a decrease in search complexity of nearly 75autocorrelation technique is used which unfortunately requires increase in the need of storage capacity.

We could not achieve the performance of a multi stage vector quantiser (MSVQ) for LSFs as the one in [8] using SSOM. We believe that the first stage of a multi stage VQ can be used for LSFs clustering. Three methods can be used. If splitting technique is used for training of the first stage of the coder, taking care of topographical neighbourhood relations while splitting, will result a coder that is not folded and with good spectral distortion. The second approach may be grouping the codewords according to their distances to each other using appropriate threshold levels. The third approach may be training a SSOM and placing the codewords of the MSVQ on a two dimensional map using the closeness to the nodes of the map of SSOM. Using a MSVQ will result further decrease in bit rate and better performance.

References

[1] R.L. Zinser, S.R. Koch "CELP Coding at 4.0 Kbit/sec and Below: Improvements to FS1016", *Proc. ICASSP*, vol.1, pp. 313-316,1992.

[2] A.N. Suen, J.F. Wang and T.C. Yao, "Dynamic Partial Search Scheme for Stochastic Codebook of FS1016 CELp Coder", *Proc. IEEE Vis. Image Signal Process.*, vol.142, no.1, pp. 52-58, Feb. 1995.

[3] L.A. Hernandez-Gomez and E. Lopez- Gonzales,"Phonetically Driven CELP Coding Using Self Organizing Maps",*Proc. ICASSP*, vol.2, pp.628-631, 1993.

[4] T. Kohonen, "The Self Organizing Map", *Proc. IEEE*, vol.78, pp. 1464-1480, 1990.

[5] K.K. Paliwal and B.S. Atal, "Efficient Vector Quantisation of LPC Parameters at 24 bits/frame", *Proc. ICASSP*, 1991.

[6] E. Germen, S. Bilgen, "A Statistical Approach to Determine the Neighbourhood Function and Learning Rate in Self Organizing Maps", *Proc. Int. Con. on Neural Information Processing and Inteligent Information Systems*, vol.1., pp.334-337, 1997.

[7] E. Yair, K. Zeger, A. Gersho, "Competitive Learning and Soft Competition Scheme for Vector Quantizer Design" *IEEE Trans. Signal Proccessing*, vol.40, no.2, February 1992.

[8] W. P. LeBlanc, B. Bhattacharya, S. A. Mahmoud and V. Cuperman, "Efficient Search and Design Procedures for Robust Multi-Stage VQ of LPC Parameters for 4 Kb/s Speech Coding", *IEEE Trans. Speech and Audio Proccessing*, vol.1, pp.373-385, Oct. 1993.

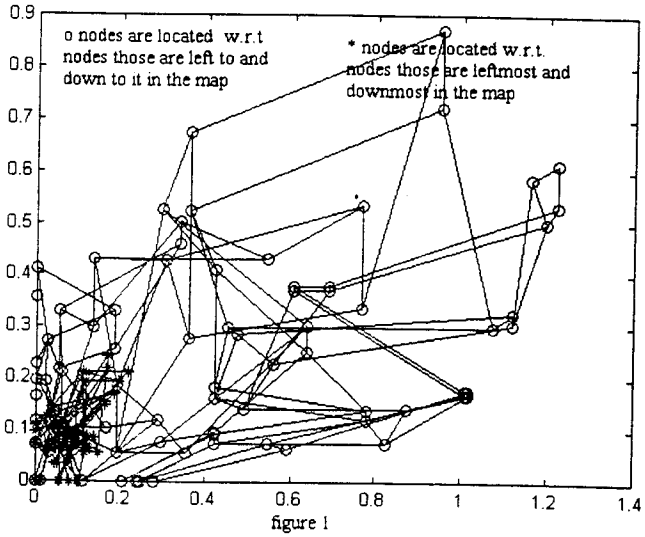

figure 1

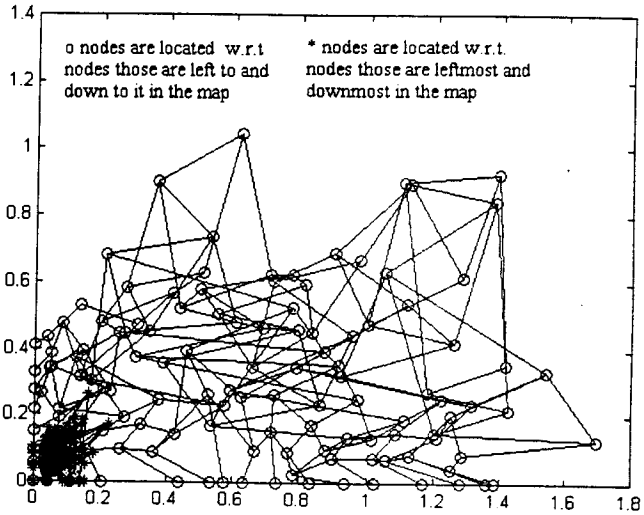

figure 2

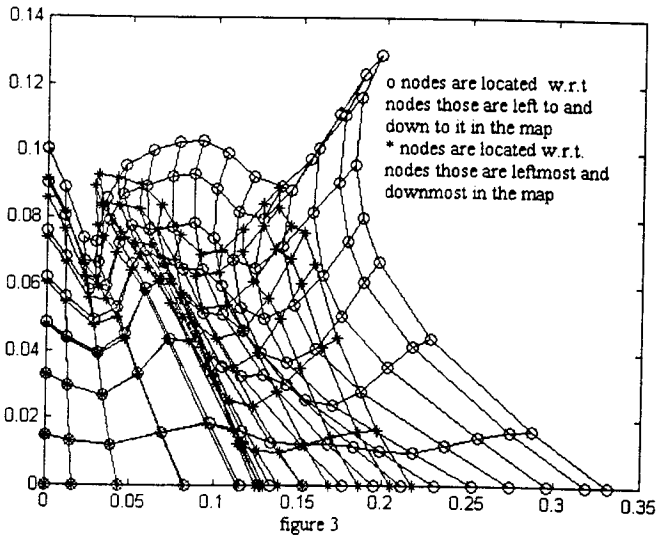

figure 3

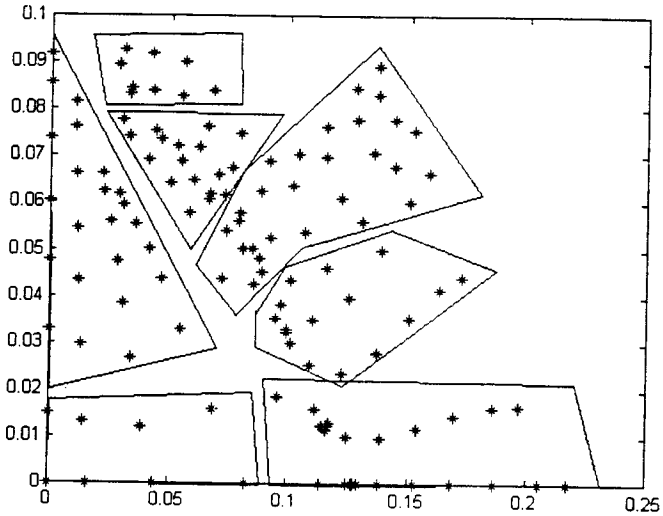
figure 4