LONG-TERM QUANTIZATION OF SPEECH LSF PARAMETERS

Laurent GIRIN

ICP (Speech Communication Lab.) – INPG/Univ. Stendhal/CNRS
B.P. 25 - 38040 Grenoble France
girin@icp.inpg.fr

ABSTRACT

This paper addresses the problem of coding the LSF parameters of LPC speech coders on a "long-term" basis, i.e. beyond the usual #20ms frame duration. The objective is to provide efficient LSF quantization for a speech coder with very large delay but very- to ultra-low bit-rate and good quality. To do this, a long-term model of the time-trajectory of the LSF vectors is applied on long segments of speech to capture the inter-frame correlation of the vectors over each whole segment. Using this model, it is shown that only a reduced set of LSF vectors need to be quantized to derive quantized LSF vectors at every original location. Experiments show that large gains in bit-rate over usual frame-by-frame quantization can be achieved (up to more than 50%) while preserving signal quality.

Index Terms— Very/ultra low bit-rate speech coding, LPC coder, LSF quantization, long-term model.

1. INTRODUCTION

The quantization of Line Spectral Frequencies (LSF) parameters is a major issue in LPC-based speech coders at low rates [1][2]. The LSF parameters are an appropriate representation of the LPC filter which is robust to quantization and interpolation [1]. In speech coders, the analysis and coding process is made on a short-term basis, using 20 ms-or-so signal frames. However, the LSF parameters encode the "vocal tract filter", which evolution is quite smooth and regular for many speech sequences. Therefore, *long-term* (LT) correlation between successive LSF values is expected to happen for many speech sections (in this paper *long-term* refers to considering long sections of speech, including several to many short-term frames of about 20ms). However, because of delay constraints, the inter-frame correlation of LSFs is generally considered locally, *i.e.* between two or three consecutive frames, using for example predictive [3] or matrix [4] quantization techniques.

In [5], Dusan *et al* have proposed to model the trajectories of ten consecutive LSF parameters by a fourth-order polynomial model. In addition, they implemented a very-low bit rate speech coder exploiting this idea. At the same time, we proposed in [6] to model the long-term trajectory of sinusoidal speech parameters with a cosine-based model. In [6], the size of parameter trajectories and the number of model coefficients were variable and could exhibit quite different (and often larger) combinations than the ten-to-four conversion of [5]. In the present paper, we extend the basic idea of modeling the trajectory of LSFs [5], by adapting our own approach of [6]. The objective is to provide efficient LSF quantization for a "long-term speech coder". Such a coder has a quite large delay, and can be used in applications such as half-duplex communication, speech storage, and speech synthesis. To do this, we propose a new method, based on the following process: first, speech is segmented into voiced/unvoiced sections; then a long-term model of the LSF trajectories is applied on each segment to capture the LT inter-frame correlation of these parameters. The LT model is a sum of cosine functions closely related to the well-known DCT transform. A procedure that enables switching from the long-term model coefficients to a reduced set of LSF vectors, and vice-versa, is introduced. It is directly inspired by the work in [5]. The reduced set of LSF vectors is quantized by multi-stage vector quantizers, transmitted, and used a the decoder to interpolate LSF vectors at the original locations

This paper is organized as follows. The proposed long-term model is described in Section 2. The complete long-term quantization of LSF vectors is presented in Section 3, Experiments and results are given in Section 4. Section 5 is a short conclusion.

2. LONG-TERM MODEL FOR LSF TRAJECTORIES

In this section, we first consider the problem of modeling the time-trajectory of a sequence of K consecutive LSF parameters. These LSF parameters correspond to a given (all voiced or unvoiced) section of speech signal $s(n)$, running arbitrary from $n = 1$ to N. They are obtained from $s(n)$ using a standard LPC analysis procedure applied on successive short-term analysis windows (see Section 4.1). For the following, let us denote by $\boldsymbol{K} = [n_1\ n_2\ \dots\ n_K]$ the vector containing the sample indexes of the analysis frame centers. Each LSF vector resulting from the analysis at instant n_k is denoted $\boldsymbol{\omega}_{(I),k} = [\omega_{1,k}\ \omega_{2,k}\ \dots\ \omega_{I,k}]^T$, for $k = 1$ to K (^{T} denotes the transpose operator). I is equal to 10 for telephone speech. Thus, we actually have I LSF trajectories of K values to model. For this aim, let us denote by $\boldsymbol{\omega}_{(I),(K)}$ the $I \times K$ matrix of general entry $\omega_{i,k}$: The LSF trajectories are the I row K-vectors, denoted $\boldsymbol{\omega}_{i,(K)} = [\omega_{i,1}\ \omega_{i,2}\ \dots\ \omega_{i,K}]$, for $i = 1$ to I.

Different kinds of models can be used for representing these trajectories. As mentioned in the introduction, a fourth-order polynomial model was used in [5] for representing ten consecutive LSF values. In [7], we compared different models for long-term modeling within the sinusoidal speech framework. The Discrete Cosine Model (DCM) was the best, and because of the limitation of experimental configurations in Section 4, we consider only this model in the present paper. The DCM model is defined for each of the I LSF trajectories by:

$$\widetilde{\omega}_i(n) = \sum_{p=0}^{P} c_{i,p} \cos\left(p\pi \frac{n}{N}\right). \qquad (1)$$

where P is a positive integer defining the order of the model and the P+1 model coefficients $c_{i,p}$ are all real (note that in the present study, all I LSF trajectories are modeled with the same order P, although a specific order could be defined for each trajectory in a more general approach, at the cost of increased complexity).

Given that P is known and P+1<K, the I×(P+1) matrix gathering the model coefficients $c_{i,p}$ is obtained by minimizing the mean square error (MSE) between the model values evaluated at the analysis instants $\boldsymbol{CM}_i$ and the LSF data set $\boldsymbol{\omega}_{i,(\boldsymbol{K})}$:

$$\boldsymbol{C} = \boldsymbol{\omega}_{(\boldsymbol{I}),(\boldsymbol{K})}\boldsymbol{M}^T\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1} \tag{2}$$

where $\boldsymbol{M}$ is the (P+1)×K *model matrix* that gathers the DCM terms evaluated at the entries of $\boldsymbol{N}$, *i.e.* $\boldsymbol{M}$ is the matrix of general entry $m_{pk} = \cos(p\pi n_k/N)$, p = 0 to P, k = 1 to K. In practice, we used the regularized version of (2) proposed in [8]: a diagonal penalizing term is added to the inverted matrix in (2) to regularize possible ill-conditioning problems. In our study, setting the regularizing factor λ of [8] to 0.01 enabled to never encounter ill-conditioned matrices. Finally, please note that the modeled LSF trajectories can be expressed as the lines of the matrix:

$$\tilde{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{K})} = \boldsymbol{C}\boldsymbol{M} \tag{3}$$

3. LT QUANTIZATION OF LSF

In this section, we present the complete algorithm for quantizing each sequence of K LSF vectors. The shape of the LSF trajectories can vary widely, e.g. depending on the length of the sequence, the phonetic content, the speaker, and the prosody. Therefore, the appropriate order P of the LT model can also vary widely, and it must first be estimated based on a trade-off between LT model accuracy (for a good representation of the data) and sparseness (for bit-rate limitation). For clarity, we first present the quantization process assuming that P is known and we present next how it can be estimated for each sequence of LSFs.

The first step of the LSF coding process is to calculate the DCM coefficient matrix $\boldsymbol{C}$ of eq. (2) with the order set to P. The next step is the quantization, *i.e.* the representation of the resulting information with limited binary resource. To do this, we propose to avoid a direct quantization of $\boldsymbol{C}$, by applying an invertible transformation between $\boldsymbol{C}$ and a reduced set of LSF vectors, and then to quantize this reduced set using usual techniques. For this purpose, we first calculate the set of P+1 indexes, denoted $\boldsymbol{J} = [j_1\, j_2\, \ldots\, j_{P+1}]$, that correspond to equally-spaced time positions within the N samples of the considered section of speech (with rounding to the nearest integer if necessary). Let us then define $\boldsymbol{Q}$ a "reduced" model matrix evaluated at the instants of $\boldsymbol{J}$ (remind that P+1<K), *i.e.* $\boldsymbol{Q}$ is the square matrix of general entry $q_{pj} = \cos(p\pi j_r/N)$, p = 0 to P, r = 1 to P+1. The reduced set of LSF vectors is the set of P+1 modeled LSF vectors calculated at the instants of $\boldsymbol{J}$, *i.e.* the columns of the matrix:

$$\tilde{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{J})} = \boldsymbol{C}\boldsymbol{Q}\,. \tag{4}$$

The proposed method uses the fact that the matrix $\boldsymbol{C}$ of (2) can be exactly retrieved from the reduced set of LSF vectors, by:

$$\boldsymbol{C} = \tilde{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{J})}\boldsymbol{Q}^T\left(\boldsymbol{Q}\boldsymbol{Q}^T\right)^{-1}. \tag{5}$$

Therefore, the quantization strategy is the following. Only the reduced set of P+1 LSF vectors is quantized, instead of the overall set of K original vectors, as would be the case in a usual coding schema. This is done by using Multi-Stage Vector Quantization (MS-VQ) techniques [9] presented in Section 4. The indexes of the P+1 codewords are then transmitted. At the decoder, the resulting quantized vectors, denoted $\hat{\boldsymbol{\omega}}_{(\boldsymbol{I}),p}$, for p = 1 to P+1, are retrieved from the codewords. They are then gathered in a I×(P+1) matrix denoted $\hat{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{J})}$, and the DCM coefficient matrix is estimated by applying (5) with the quantized reduced set of LSF vectors instead of the unquantized reduced set:

$$\hat{\boldsymbol{C}} = \hat{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{J})}\boldsymbol{Q}^T\left(\boldsymbol{Q}\boldsymbol{Q}^T\right)^{-1} \tag{6}$$

Eventually, the resulting DCM coefficients are used to recalculate the "quantized" LSF vectors at the original K time instants n_k by applying the following variant of (3):

$$\hat{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{K})} = \hat{\boldsymbol{C}}\boldsymbol{M} \tag{7}$$

Note that the resulting LSF vectors, which are the columns of the above matrix, are abusively called the "quantized" LSF vectors, although they are not directly generated by MS-VQ. This is because they actually are the LSF vectors used at the decoder for signal reconstruction. Note also that the {K, P} values must be transmitted to the decoder as additional information. However, since the average number of sections per second is low (e.g., about 3 for the voiced sections), the {K, P} pair can be coded with very few bits, say a few 10s of bits/s, using e.g., Huffman coding. Thus, this additional bit-rate remains significantly lower than the gain provided by the method (see next section).

Let now consider the problem of estimating P. For this aim, a performance criterion for the overall process is first defined. This criterion is the usual Average Spectral Distortion (ASD) measure, which is a standard in the LSF quantization literature [2] :

$$ASD = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{2\pi}\int_0^{2\pi}\left[10\log_{10}P_k\left(e^{j\omega}\right) - 10\log_{10}\hat{P}_k\left(e^{j\omega}\right)\right]^2 d\omega} \tag{8}$$

where $P_k(e^{j\omega})$ and $\hat{P}_k(e^{j\omega})$ are the LPC power spectra corresponding to the original and quantized k-th LSF vectors of the considered sequence, respectively. In practice, ASD is calculated using a 512-bins FFT. We then fix a maximal value for ASD, denoted ASD_{max}, and we apply the following iterative algorithm.

Algorithm for LT quantization of a *K*-sequence of LSF vectors

1. Choose a value for ASD_{max}. Set P = 1;
2. Apply the LT LSF quantization process, *i.e.*:
   - calculate $\boldsymbol{J}$ (vector of regularly spaced breakpoints),
   - calculate $\tilde{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{J})} = \boldsymbol{\omega}_{(\boldsymbol{I}),(\boldsymbol{K})}\boldsymbol{M}^T\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{Q}$ ((2) and (4)),
   - quantize $\tilde{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{J})}$ to obtain $\hat{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{J})}$,
   - calculate $\hat{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{K})} = \hat{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{J})}\boldsymbol{Q}^T\left(\boldsymbol{Q}\boldsymbol{Q}^T\right)^{-1}\boldsymbol{M}$ ((6) and (7));
3. Calculate ASD between $\boldsymbol{\omega}_{(\boldsymbol{I}),(\boldsymbol{K})}$ and $\hat{\boldsymbol{\omega}}_{(\boldsymbol{I}),(\boldsymbol{K})}$ with (8);
4. If ASD>ASD_{max} and P<K-2, set P←P+1, and go to step 2, else (*i.e.* if ASD<ASD_{max} or P=K-2), terminate the algorithm.

4. EXPERIMENTS

4.1. Database

We used sentences from the TIMIT database [10], filtered to the 300-3400 Hz telephone band, and resampled at 8 kHz. The LSF vectors were calculated using the autocorrelation method, with a 25 ms Hamming window, high-frequency pre-emphasis with the filter $H(z)=1-0.9375z^{-1}$, and 10 Hz-bandwidth expansion. A total of 176 speakers (half male and half female) of the eight different dialect regions of TIMIT were used for building the training corpus used to design the MS-VQ quantizers (see next subsection), leading to a total of 223,501 voiced LSF vectors and 74,533 unvoiced LSF vectors. The voiced/unvoiced segmentation was based on the TIMIT label files. In parallel, 88 other speakers (also half male, half female, and from eight dialect regions) were used for the test corpus. This test corpus was used to test the LT coding process, and it contains 67,080 voiced vectors from 4,656 sections, and 22,101 unvoiced vectors from 4,427 sections.

4.2. MS-VQ Codebook Design

For the quantization of LSF vectors, we designed two-stages MS-VQ: the quantization error at the output of the first VQ block is quantized by a second block. The quantized vectors are reconstructed by adding the outputs of the two blocks. In such structure, the global complexity is highly reduced compared to single-stage VQ. In this study, different quantizers are used for voiced or unvoiced LSF vectors (as in [11]). We used a resolution ranging from 15 to 24 bits/vector, which generally corresponds to poor-quality to transparent or "close to transparent" quantization (depending on the quantizer structure) [2][9].

The design of the quantizers was made by applying the LBG algorithm [12] on the (voiced or unvoiced) training corpus described in the previous subsection, using the weighted Euclidian distance of [2]. The LBG algorithm was first used to design the first-stage codebook. Then, the difference between each LSF vector of the training corpus and its associated codeword was used for the design of the second-stage codebook, again with the LBG algorithm. When the total number of bits was even, the two blocks were allocated the same number of bits. Otherwise, the first block was allocated one bit more than the second block.

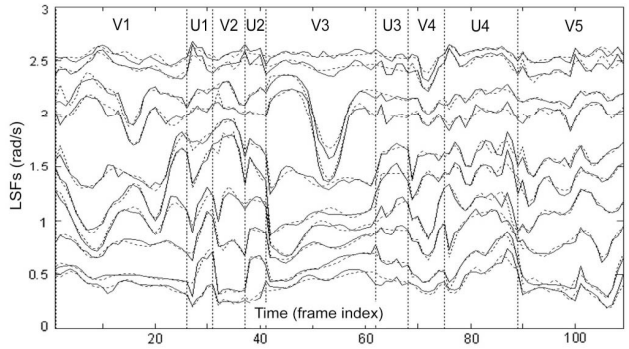

Figure 1: LSF trajectories corresponding to the sentence "*Elderly people are often excluded*" from the TIMIT database, pronounced by a female speaker. The vertical lines define the long term frame boundaries. The total number of LSF vectors is 108. Solid line: original LSF vectors; dotted line: LT coded LSF vectors with ASD_{max} = 2dB for the voiced sections and ASD_{max} = 3dB for the unvoiced sections, and with r = 20 bits/vectors.

4.2. Results

We present on Fig. 1 the original LSF trajectories of a peculiar sentence of the test corpus ("*Elderly people are often excluded*" pronounced by a female speaker), together with the corresponding LT-coded LSF trajectories, after termination of the algorithm. This sentence contains five voiced (V) sections and four unvoiced (U) sections. In this experiment, the ASD target (ASD_{max}) was fixed to 2 dB for the voiced sections and 3 dB for the unvoiced sections. The resolution r was fixed to 20 bits/vector for both voiced and unvoiced quantizers. The total number of DCM coefficients for all the sentence was 50, for $K = 108$ original LSF vectors. Fig. 1 illustrates the ability of the LT model of LSF trajectories to globally fit the original LSF trajectories, even if the model coefficients are calculated from the quantized reduced set of 50 LSF vectors (as is the case at the decoder).

Now, we present global results obtained on the entire test database (4,656 voiced sections or 4,427 unvoiced sections) in terms of ASD (8) and average bit-rate. Since the LT coding scheme is an intrinsic variable-rate technique, this latter is defined by:

$$b = \left(r \times \sum_{m=1}^{M} \left(P_m + 1 \right) \right) \Big/ \left(h \times \sum_{m=1}^{M} K_m \right) \qquad (9)$$

where, m indexes each sequence of LSF vectors of the database, M is the number of sequences, r is the resolution of the quantizer (in bits/vector), and h is the hop size of the LSF analysis window (we set h = 20 ms, to fairly compare our method with the frame-by-frame approach, since 20 ms is a usual frame spacing for LSF coding). Note that, in the LT coding process, increasing the resolution does not necessarily increase the bit-rate, as opposed to usual coding methods, since it may lead to decrease the number of LT model coefficients.

The results are presented in terms of distortion-rate curves in Fig 3 for the voiced sections, and in Fig. 4 for the unvoiced sections. Each one of the curves on the left corresponds to a fixed resolution (which value is plotted), the ASD target ASD_{max} being varied with a 0.1dB step. The curve on the right corresponds to the frame-by-frame quantization, for which the performances were also calculated for comparison, for different resolutions.

It can be seen that the curves corresponding to the LT coding are all situated on the left of the curve of the frame-by-frame quantization. They thus correspond to smaller bit-rates. Moreover, by taking the leftmost point, the gain in bit-rate for approximately the same ASD can be very large, depending on the considered region and the chosen LT coding configuration. For instance, for voiced speech and for an ASD approximately below 1.8 dB, the use of a 24 bits/vector quantizer is an optimal choice (but untested greater resolutions are likely to provide even better results). Also, in this region, the bit-rate difference for the two methods increases as the ASD increases. For example, with r = 17 bits/vector, the ASD obtained with the frame-by-frame quantizer is 1.82 dB, for a bit-rate of 850 kbits/s = 17×50 bits/s. At the same time, the LT coder with r = 24 bits/vectors provides 1.78 dB of ASD at 540.3 bits/s. Thus, the bit saving is about 36.5% (310 out of 850). For ASD values above 1.8 dB, the plot is a little more intricate, since the different LT coding curves are crossing each other. This crossing effect illustrates the trade-off between quantization accuracy and modeling accuracy that has already been mentioned. However, very large gains in bit-rate can again be obtained. Moreover, the (optimal) bit-rate *difference* between the leftmost

point from all LT coding curves and the point of the frame-by-frame coding is quite stable across ASD values: it remains close to 300 bits/s. Therefore, the *relative gain* in bit-rate between the two methods is increasing with the ASD. For example, the ASD obtained with the frame-by-frame quantizer at $r = 15$ ($b = 750$ bits/s) is equal to 2.05 dB. The same ASD value is obtained with the LT coding with $r = 21$, with a bit-rate of 420.5 bits/s. Thus, the bit saving is here 329.5 out of 750, *i.e*. 43.9%. For $r = 11$ bits/vector ($b = 550$ bits/s) the frame-by-frame quantization provides 2.68 dB of ASD, while the LT coding provides 2.67 dB of ASD at a bit-rate of 256.2 bits/vector (with $r = 16$ bits/vector). Thus, the bit-rate difference is here 293.8 bits/vector and the relative gain of the LT coding over frame-by-frame coding reaches 53.4% in this low resolution region.

For unvoiced sections, the general trends discussed in the voiced case can be retrieved in Fig 4. However, the bit-rate gains are generally lower than in the voiced case, although they remain significant. For example, the LT coder with $r = 17$ bits/vectors provides 2.22 dB of ASD, with a bit-rate of 394.8 bits/s. The same ASD is obtained with the frame-by-frame quantizer with $r = 11$ bits/vector (bit-rate = 550 bits/s). The bit-rate gain is thus 155 out of 550 (28.2%). Again, larger relative gains are likely to be obtained for lower untested resolutions and lower coding quality.

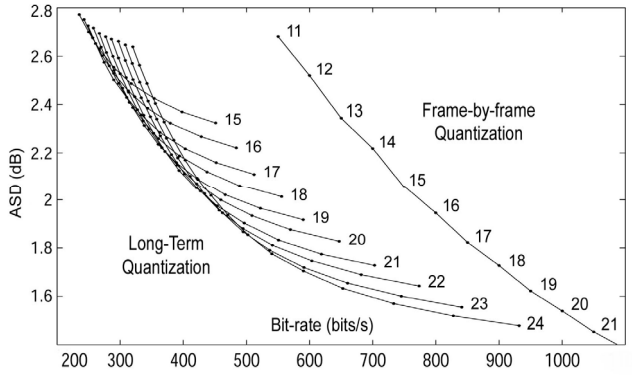

Figure 3: Average spectral distortion (ASD) as a function of the average bit-rate, calculated on the 4,656 voiced sections of the test database (67,080 vectors), and for both the LSF LT coding (series of curves on the left) and frame-by-frame LSF quantization (curve on the right). The plotted numbers are the resolutions (in bits/vector). For each resolution, the different points of each curve on the left cover the range of the ASD target.

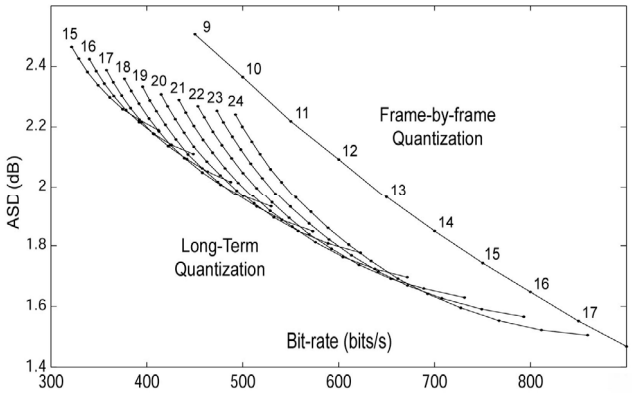

Figure 4: Same as Fig. 3, but for the unvoiced database (22,101 vectors / 4,427 sections).

5. CONCLUSION

As a conclusion it can be noted that a preliminary series of listening tests were conducted to confirm the efficiency of the long-term quantization of LSF parameters from a subjective point of view. Coded signals were generated by filtering the residual signal through a synthesis filter derived with LSF parameters coded by the two methods, LT coding and frame-by-frame quantization, with a similar ASD. These listening tests showed that, globally, the preference score was close to 50%-50%, indicating that the two methods perform equally on the average. Since the signals coded with the LT coding require much less bit-rates for the same ASD (up to more than 50%), these tests confirm the efficiency of the proposed method. More detailed results will be reported in further publication.

Future work will mainly focus on the elaboration of several complete speech coders functioning at very- to ultra-low bit-rates and exploiting the long-term approach. For such an application, the model orders could be further decreased, compared to the results presented in the present paper, while preserving acceptable quality.

6. REFERENCES

[1] J. Pan and T.R. Fischer, "Vector quantization of speech line spectrum pair parameters and reflection coefficients," *IEEE Trans. Speech and Audio Proc*., vol. 6, No. 2, 1998; pp 106-115.

[2] K.K. Paliwal and B.S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech and Audio Proc*., vol. 1, Jan. 1993, pp. 3-14.

[3] M. Yong, G. Davidson and A. Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction," in *Proc. IEEE-ICASSP*, 1988, pp. 402-405.

[4] C. Tsao and R.M. Gray, "Matrix quantizer design for LPC speech using the generalized Lloyd algorithm," *IEEE Transaction Acoust., Speech, and Signal Proc*., Vol. ASSP-33, No. 3, 1985, pp 537-545.

[5] S. Dusan, J. Flanagan, A. Karve & M. Balaraman, "Speech coding using trajectory compression and multiple sensors," *Proc. Int. Conf. on Speech & Language Proc.,* Jeju, 2004.

[6] L. Girin, M. Firouzmand & S. Marchand, "Long-term modeling of phase trajectories within the speech sinusoidal model framework," *Proc. Int. Conf. on Speech & Language Proc.,* Jeju, 2004.

[7] L. Girin, M. Firouzmand, and S. Marchand, "Comparing several models for perceptual long-term modeling of amplitude and phase trajectories of sinusoidal speech," *Proc. Interspeech Conf.,* Lisboa, Portugal, 2005.

[8] O. Cappé, J. Laroche & E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points," *Proc. IEEE Workshop Applications Signal Proc. Audio Acoustics*, 1995.

[9] W. P. LeBlanc *et al*., "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding," *IEEE Trans. Speech and Audio Proc*., vol. 1, No. 4, Oct. 1993, pp 373-385.

[10] J. S. Garofolo *et al*., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, 1993.

[11] R. Hagen, E. Paksoy and A. Gersho, "Voicing-specific LPC quantization for variable-rate speech coding," *IEEE Trans. Speech and Audio Proc*., vol. 7, No. 5, 1999, pp 485-494.

[12] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. COM-28, 1980, pp. 84-94.