Multi-Mode Matrix Quantizer for Low Bit Rate LSF Quantization

Ulpu Sinervo[1], *Jani Nurminen*[2], *Ari Heikkinen*[2], *and Jukka Saarinen*[2,1]

[1]Institute of Digital and Computer Systems
Tampere University of Technology, Tampere, Finland
ulpu.sinervo@tut.fi

[2]Speech and Audio Systems Laboratory
Nokia Research Center, Tampere, Finland
{jani.k.nurminen, ari.p.heikkinen, jukka.p.saarinen}@nokia.com

Abstract

In this paper, we introduce a novel method for quantization of line spectral frequencies (LSF) converted from *m*th order linear prediction coefficients. In the proposed method, the interframe correlation of LSFs is exploited using matrix quantization where *N* consecutive frames are quantized as one *m*-by-*N* matrix. The voicing-based multi-mode operation reduces the bit rate by taking advantage of the properties of the speech signal. That is, certain parts of a signal, such as unvoiced segments, can be quantized with smaller codebooks. With this method, very low variable bit rate LSF quantization is obtained. The proposed method is suitable especially for very low bit rate speech coders in which short time delay is tolerable, and high but not necessarily transparent quality is sufficient.

1. Introduction

Quantization of linear predictive coding (LPC) parameters is an important part of a typical speech codec. Especially at low bit rates, very efficient quantization of the LPC parameters is required in terms of bit rate and quantization accuracy. Transmission of the LPC information usually consumes a significant portion of the available bits in low bit rate speech coding. At bit rates below 3 kb/s, it is common that approximately half of the bit budget is allocated for the quantization of the LPC parameters. For example in the 2.4 kb/s MELP coder more than 1.1 kb/s is reserved for the transmission of the LPC information [1].

For efficient quantization, the linear prediction coefficients are usually transformed into some other representation. The line spectral frequency (LSF) representation is one of the most popular representations due to its properties that ensure the stability of the synthesis filter and offer robust performance. Two common LSF quantization methods are split vector quantization (SVQ) [2] and multistage vector quantization (MSVQ) [3].

In the LPC parameter quantization, the objective is to avoid introducing audible distortion to coded speech. According to Paliwal and Atal, perceptually transparent split vector quantization of the LPC information is achievable at 24 bits per frame [2]. For transparent quantization their objective requirements were: an average spectral distortion (SD) of about 1 dB, less than 2% outlier frames in the range 2-4 dB, and no outliers having SD greater than 4 dB. However, Hagen *et al.* have concluded that for unvoiced spectra, the average SD of even 2 dB is tolerable and the amount of outlier frames having SD above 4 dB may be at maximum 1% [4]. According to their study, unvoiced spectra can be transparently coded using only 9 bits per frame. However, 25 bits per frame is still needed for voiced spectra in their voicing-specific quantization scheme. The saving in the total bit rate is therefore only slightly more than 10%. More bits can be saved as SVQ is extended to matrix quantization since the interframe LSF redundancy is exploited. According to Xydeas and Papanastasiou, transparent LSF quantization can be achieved at 900 b/s, whereas “high quality” is obtained at 650 b/s using split matrix quantization [5].

In this paper, we introduce an LSF quantization scheme, where a multi-frame matrix quantizer is used to exploit the correlation between the LSF parameters of the adjacent frames. In the proposed implementation, the multistage approach is employed in the quantization scheme contrary to [5] where an SVQ based technique was used. Furthermore, the quantizer has four modes based on the voicing classifications computed for each frame. The multi-mode scheme enables the usage of smaller codebooks for less sensitive modes, e.g., unvoiced segments, and thus reduces the average bit rate. With the proposed method, a very low average bit rate can be achieved.

This paper is organized as follows. Section 2 introduces the proposed multi-mode matrix quantization (MMMQ) method and an example implementation. The perceptual evaluation of the quantizer is presented in Section 3, and conclusions are drawn in Section 4.

2. Multi-mode matrix quantizer implementation

In the proposed matrix quantization scheme, the LSFs from *N* consecutive frames are gathered up to form *m*-by-*N* matrices. Each matrix is then quantized in one using some distortion criterion, e.g., *mean squared error* (MSE). The complexity of the required quantizer can be reduced using structurally constrained codebooks such as split or multistage codebooks. For multi-mode voicing-based operation, the speech frames are classified as voiced or unvoiced. The MMMQ method can thus be very easily applied to speech coders in which voicing information is already calculated for other purposes. In addition to the voiced and unvoiced classes, two transition classes are included. An LSF matrix is regarded to fall into a transition class if the speech signal is changing from unvoiced to voiced, or vice versa, during the *N* frames that form the particular LSF matrix. A block diagram of the MMMQ scheme is presented in Figure 1.

To test the proposed quantization approach, a practical multi-mode matrix quantizer for 10th order LSFs was implemented. The LSF vectors from three consecutive frames were quantized simultaneously as a 10-by-3 matrix. The matrix size was chosen in such a manner that the interframe correlation can be exploited while keeping the time delay reasonable for very low bit rate speech coding applications.

In the selection of the quantization method, both split and multistage quantization were considered. While a multistage quantizer may require 10-20 times more arithmetic operations than a 5-way split quantizer proposed in [5], it is also capable of providing similar performance with fewer bits. Since the main objective in this work is to minimize the bit rate for the LSFs, the multistage matrix quantization approach was selected.

The multistage codebooks were trained with LSF vectors generated from a database that contained 110 minutes of Finnish speech from 12 different speakers (6 female and 6 male). The sampling frequency of the speech material was 8 kHz and it had both flat and modified IRS input characteristics. Roughly two-thirds of the LSF data was used for the training and the rest was reserved for the evaluation. The LSF vectors were generated from the speech data such that at first 10th order LPC analysis based on the autocorrelation method was performed for every 20 ms speech frame. A 30 ms asymmetric window, composed of Hamming and cosine windows, was used. The autocorrelations were multiplied by a lag-window containing white noise correction and a bandwidth expansion of 60 Hz, before calculating the filter coefficients using the Levinson-Durbin recursion. The resulting coefficients of the 10th order polynomial were converted into the LSF domain. A minimum spacing of 50 Hz was maintained between adjacent LSF vector elements.

The speech data was classified frame by frame as voiced or unvoiced. Voicing classification was performed using an autocorrelation-based classification algorithm introduced in [6]. The algorithm classifies each speech frame into one of four classes: voiced, jittery-voiced, plosive or unvoiced speech. Because only voiced and unvoiced classes were needed, frames classified as jittery-voiced or plosive speech were regarded as unvoiced. The mode selection for the superframes, consisting of three consecutive frames, was done as depicted in Table 1, where V stands for a frame classified as voiced and U for an unvoiced frame. The distribution of the training data was such that 47% of superframes were classified as voiced, 29% as unvoiced, 12% as transition I, and 12% as transition II.

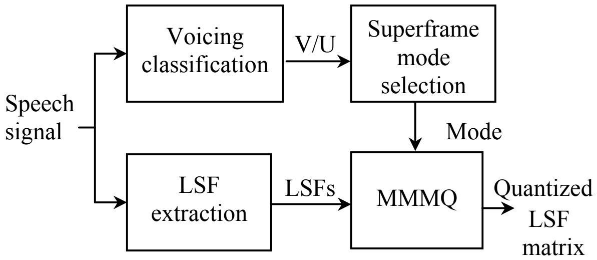

Figure 1. Block diagram of the MMMQ.

Table 1. Mode selection table.

Frame triplet	Superframe mode
VVV VUV	VOICED
UUU UVU	UNVOICED
UUV UVV	TRANSITION I
VVU VUU	TRANSITION II

To maximize the number of matrices in the training data, a sliding window was used while collecting the matrices. A three-frame window was moved one frame at a time. The codebooks for the different modes were designed using the simultaneous joint optimization procedure introduced in [7]. The *M-L* search, where the *M* best vector combinations are searched at each of the *L* stages, was used with $M = 8$. The weighted MSE was used as the distortion measure, along with the weighting proposed in [2], in which the distortion of each LSF is weighted proportionally to the value of the spectrum at the corresponding frequency *f*. The weight for the *j*th LSF of the *i*th frame is given by

$$w_{ji} = \left[P\left(f_{ji}\right)\right]^{r} c_{j} \ ,$$

where $P(f)$ is the LPC power spectrum, $r = 0.15$, and c_j is 0.8 for $j = 9$ and 0.4 for $j = 10$, otherwise c_j is 1.0.

3. Perceptual evaluation

In the context of very low bit rate speech coding (i.e., around 1 kb/s), fully transparent speech quality is probably not a realistic goal. On the other hand, accurate LPC information plays an important role in successful speech synthesis. For these reasons, our design goal was high quality but not necessarily fully transparent LSF quantization. That is, even if the coded speech might be distinguishable from the original, there are no annoying artifacts. Traditionally, objective measures such as spectral distortion have been used to evaluate the performance of an LSF quantizer. Although objective evaluation is time and cost efficient, the properties of human perception cannot be fully taken into account using only objective measures. Especially at low bit rates, the relevance of the perceptual aspects increases since often the objective measurements and the subjective quality are not correlated. Our quantizer implementation is thus evaluated subjectively.

The perceptual quality of the quantizer implementation was evaluated using the *comparison category rating* (CCR) method [8]. In the CCR procedure, the participants listen to pairs of speech samples and for each pair judge the quality of the second sample compared to the first. The quality is evaluated using a seven-point scale ranging from –3 (*much worse*) to +3 (*much better*). The sample pairs as well as the sentences in each pair are presented in random order.

The perceptual evaluation process was conducted with 24 naive and 11 expert listeners. The test material consisted of Finnish sentences spoken by three female and three male speakers, one sentence from each speaker. The material was divided into three parts such that each participant listened 42 sample pairs including four reference pairs. The sentences were filtered to have a modified IRS characteristic. The material was listened with high-quality headphones.

Each pair of speech samples contained a *reference sentence* obtained without any quantization, and a *test sentence* with quantized LSFs. The test sentences were generated using the arrangement shown in Figure 2. The residual signal was obtained by filtering the original speech signal with an LPC analysis filter with the unquantized coefficients. The residual signal was then used to excite an LPC synthesis filter with the quantized coefficients.

Part of the test sentences contained fully quantized LSFs, whereas other sentences had the spectrum quantized only for frames belonging to one of the four classes. In the former case, six different bit allocations were tested to find the combination that satisfies the requirement of high-quality quantization. In the latter case, the objective was to find for each mode the minimum codebook size that yields perceptually transparent LSF quantization. Percentages of different frame types in the test sentences are given in Table 2.

The traditional MNRU references were excluded from the test to avoid confusing the listeners since distortions caused by the poor LSF quantization are fundamentally different than those caused by the added noise. However, to ensure that listeners make use of the whole rating scale (from –3 to +3), the test material contained also four reference sample pairs, in which the quality of the other sentence was degraded. These signals were processed such that a part of the frequency components (i.e., FFT coefficients) was randomly removed or attenuated. From samples Ref. 1 and Ref. 3 (see Table 5) 30% and 50% (respectively) of coefficients were removed. A part of frequency components were attenuated in samples Ref. 2 and Ref. 4 by zeroing 60% and 50% of coefficients, respectively, and by smoothing the spectra. Two reference samples were spoken by female and two by male speaker, as shown in Table 5.

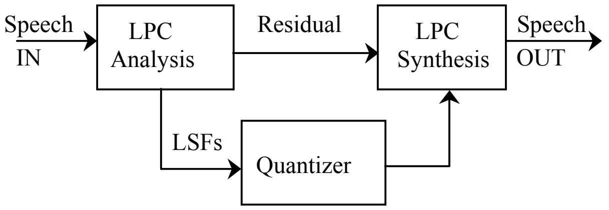

Figure 2. System for generating test sentences for perceptual evaluation.

The data gathered from the listeners was treated so that at first the average score for each sample pair was obtained by calculating the mean of all marks given to that pair. The overall CMOS (comparison mean opinion score) for the separate quantization conditions, i.e., a certain codebook combination, was then calculated by averaging the scores of the sample pairs relating to that condition. The CMOS results, divided into naive and expert listener groups (the latter marked with the subscript E), are presented in Tables 3 and 4. In Table 4 the codebook sizes are given in the following order: voiced – unvoiced – transition I (unvoiced to voiced) – transition II (voiced to unvoiced). The confidence intervals for the means of scores are presented in Figures 3 and 4. The scores for the reference sample pairs are shown in Table 5.

From the results shown in Table 3 it can be seen that nearly transparent quantization of voiced spectra can be achieved with 40 bits per superframe. For other modes, more than 16 bits are apparently needed for the transparent quality. The most important result is, however, that for good-quality quantization the average bit rate can be reduced to around 10 bits per frame, as the results presented in Table 4 suggest.

Table 2. Distribution of different frame types in the training data and in the test sentences.

Data	Voiced	Unvoiced	Trans. I	Trans. II
Training	47 %	29 %	12 %	12 %
Test	43.5 %	33.5 %	10 %	13 %

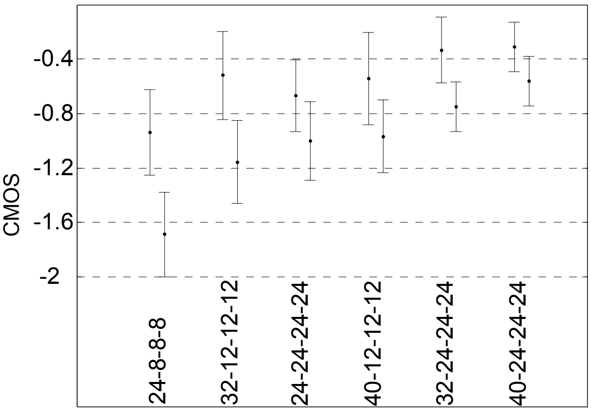

Figure 3. 95% confidence intervals for CMOS (left bar) and $CMOS_E$ (right bar) of the different quantization modes.

Table 3. Listening test results for the different modes obtained by quantizing only the LSF-matrices that belong to the particular class.

Mode	Bits	CMOS	$CMOS_E$
Voiced	24	-0.44	-0.64
	32	-0.25	-0.62
	40	+0.08	-0.06
Unvoiced	8	-0.23	-0.46
	12	+0.06	-0.31
	16	-0.27	-0.20
Transition I	12	-0.10	-0.13
	16	-0.06	-0.32
Transition II	12	-0.10	-0.19
	16	-0.02	-0.03
Transition I + II	12	-0.08	-0.57
	16	-0.08	-0.24

Table 4. Listening test results for the different bit allocations.

Bit allocation V–U–UV–VU	Bits/frame (average)	CMOS	$CMOS_E$
24–8–8–8	5.2	-0.94	-1.66
32–12–12–12	7.1	-0.52	-1.16
24–24–24–24	8.0	-0.67	-1.00
40–12–12–12	8.4	-0.54	-0.97
32–24–24–24	9.3	-0.33	-0.74
40–24–24–24	10.5	-0.31	-0.55

4. Conclusions

A novel multi-mode matrix quantization method for quantizing the LPC information at low bit rates was proposed. A practical implementation employing the multistage quantization approach was described. The results of the perceptual evaluation with Finnish speech indicated that high-quality quantization can be obtained using, on average, approximately 10 bits per frame.

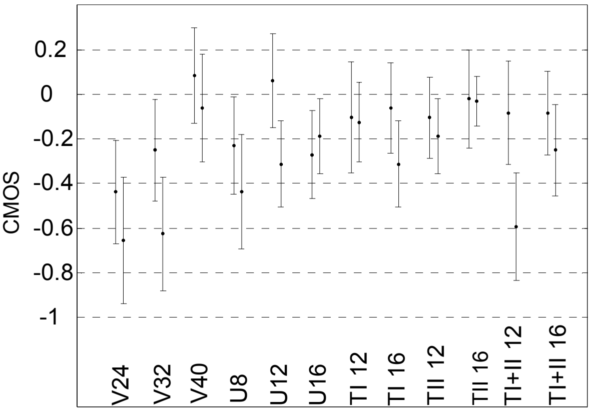

Figure 4. 95% confidence intervals for CMOS (left bar) and $CMOS_E$ (right bar) of the different bit allocations.

Table 5. CMOS results for the reference samples.

Reference sample	CMOS	$CMOS_E$
Ref. 1 female 30% removed	-2.50	-3.00
Ref. 2 female 60% attenuated	-2.50	-2.94
Ref. 3 male 50% removed	-1.92	-2.69
Ref. 4 male 50% attenuated	-1.75	-1.50

5. References

[1] Supplee, L. M., Cohn, R.P., Collura, J. S., and McCree, A. V., "MELP: The New Federal Standard at 2400 bps", in *Proc. 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Apr. 1997, pp. 1591-1594.

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

[3] Gersho, A. and Gray, R. M., *Vector Quantization and Signal Compression,* Kluwer Academic Publishers, Boston, 1992.

[4] Hagen, R., Paksoy, E., and Gersho, A., "Voicing-Specific LPC Quantization for Variable-Rate Speech Coding", *IEEE Trans. Speech and Audio Proc.*, Vol. 7, pp. 485-494, Sep. 1999.

[5] Xydeas, C. S. and Papanastasiou, C., "Split Matrix Quantization of LPC parameters", *IEEE Trans. Speech and Audio Proc.*, Vol. 7, pp. 113-125, Mar. 1999.

[6] Heikkinen, A., *Development of a 4 kbps Hybrid Sinusoidal/CELP Speech Coder*, Doctoral Dissertation, Tampere University of Technology, Tampere, Finland, June 2002.

[7] LeBlanc, W. P., Bhattacharya, B., Mahmoud, S. A., and Cuperman, V., "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, pp. 373-385, Oct. 1993.

[8] *Methods for subjective determination of transmission quality*, ITU-T Recommend. P.800, 1996.