CELP SPEECH CODING: COMPARISONS IN TERMS OF QUANTIZATION TECHNIQUES FOR THE SYNTHESIS FILTER PARAMETERS

R.S. Maia, R.J.R. Cirigliano, D. Rojtenberg, F.G.V.Resende Jr.

Programa de Engenharia Elétrica/COPPE, DEL/EE
Universidade Federal do Rio de Janeiro
C.P. 68504, 21945-970, Rio de Janeiro, RJ, Brasil
Emails: {maia, rjcirig, rojtenberg, gil} @lps.ufrj.br

Abstract - *Vector quantization of the synthesis filter parameters in Code-Excited Linear Prediction (CELP) speech coders is a common procedure nowadays. This paper describes a CELP coder implementation and makes a comparison in terms of quality and bit rate when vector and scalar quantization of the synthesis filter parameters are employed.Usage of vector quantization in comparison to scalar quantization allows for a bit rate reduction of 340 bps, giving rise to a 4,06 kbps CELP coder, keeping a similar subjective evaluation.*

1. INTRODUCTION

The growing necessity of telephony systems to offer better products and the progress in digital technologies have increased digital speech processing research. Today's internet applications and the recent improvement of telephony systems demand the development of innumerous speech coding techniques. The goal of most of these techniques is to obtain good speech quality at low bit rates.

A currently quite employed technique for speech coding at low bit rates is the Code-Excited Linear Prediction (CELP) [1]. To work at 4 kbps with a good quality, systems often use vector quantization (VQ) for linear prediction coefficients. Even though this technique is computationaly more complex than scalar quantization (SQ) [2, 3, 4], it has been the preferred solution to code 10 coefficients using less than 30 bits.

In this article, a CELP coder based on the implementation described in [5] will be used to compare three different quantization techniques for the synthesis filter: the scalar quantization, the multi-stage vector quantization and the split vector quantization.

This work is organized as follows. In Section 2 the standard CELP coder is presented. Section 3 describes the linear prediction coefficients (LPC) analysis used to obtain the synthesis filter parameters. Section 4 details the VQ techniques implemented. Section 5 deals with the excitation parameters for the CELP coder. Section 6 shows the results and Section 7 presents the conclusions.

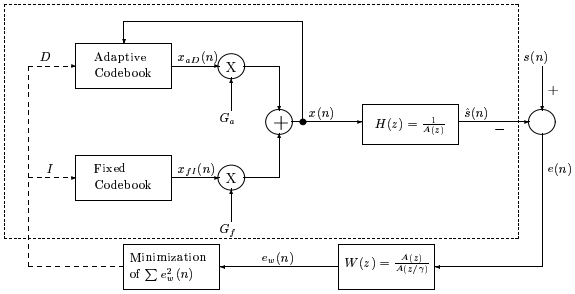

Fig. 1. Block diagram of the CELP coder.

2. CELP CODER

Figure 1 shows the block diagram for the CELP coder. The reconstructed signal is obtained passing the excitation signal $x(n)$ through the synthesis filter

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}. \qquad (1)$$

The LPC, $\{a_i : i = 1, \ldots, p\}$, are obtained by the LPC analysis. The excitation signal $x(n)$ is composed of a signal from the adaptive codebook, $x_{aD}(n)$, and a signal from the fixed codebook, $x_{fI}(n)$, weighted by their respective gains G_a and G_f. The adaptive codebook brings information related to the periodicity of voiced speech and the fixed codebook is composed of codevectors that represent the residual signals without short and long term correlations.

The information sent to the decoder includes the synthesis filter $H(z)$ coefficients, the indexes D and I and the fixed and adaptive codebooks gains G_a and G_f. These last four parameters are obtained through a technique named analysis-by-synthesis [5], i.e., some codevectors combinations are tested, being chosen the one that minimizes the energy of the weighted error $e_w(n)$, resulted from filtering the error signal $e(n)$ (given by the difference between the target signal and the reconstructed signal) through the error weighted filter $W(z) = \frac{A(z)}{A(z/\gamma)}$, where $\gamma = 0.8$ [5] is the adjustment factor.

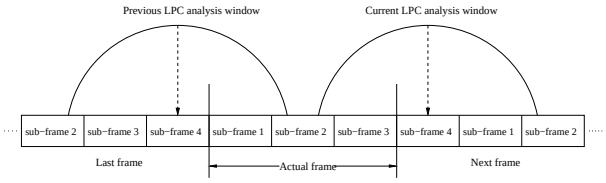

Fig. 2. Position of the window for the LPC analysis.

The CELP coder processes speech signals in frames. In each frame the LPC $\{a_i : i = 1, \ldots, p\}$ for the synthesis filter are calculated. Each frame is divided into subframes, and the best excitation $x(n)$ is found for each subframe.

3. THE LPC ANALYSIS

The CELP coder implemented uses frames of 30 ms divided into four subframes of 7,5 ms each. The LPC analysis is undertaken using the 10^{th} order ($p = 10$) autocorrelation method, and Hamming windows of 25 ms centered in the last subframe of each frame as shown in Figure 2. The LPC are transformed in line spectral frequencies (LSF). To obtain a soft transition between two consecutive groups of LSF, they are interpolated. For the interpolation, the subframes are numbered from 1 to 4 as shown in Figure 2. The i-th coefficient from the subframe n is obtained by

$$w_i^n = (1 - q_n)w_i^a + q_n w_i^c, \tag{2}$$

where w_i^c refers to the coefficient w_i from the current frame and w_i^a refers to the coefficient w_i of the previous frame. The weighting vector used is $\mathbf{q} = [0,25 \quad 0,5 \quad 0,75 \quad 1]^T$. It can be easily observed that $w_i^4 = w_i^c$, $\forall i$, once that the Hamming window is centered on the 4^{th} subframe.

4. THE LSF VECTOR QUANTIZATION

In this article the SQ of the LSF coefficients described in [5] is compared to the following two VQ techniques.

4.1. Multi-stage vector quantization

The multi-stage vector quantization (MSVQ) implemented in this work is similar to the one used in the MELP (*Mixed Excitation Linear Prediction*) coder [6, 7]. A modification was introduced in the calculation of weights used to determine the average distortion.

The VQ used is a four stages one. Each quantized vector $\hat{\mathbf{w}} = [\hat{w}_1 \ldots \hat{w}_{10}]^T$ is obtained by the sum of the components in each stage, where the number of levels are 128, 64, 64 and 64. Therefore, a total of 25 bits is used to quantize each vector $\mathbf{w} = [w_1 \ldots w_{10}]^T$, while in the SQ 32 bits were used. Table 1 shows the target vectors of each stage and its number of levels and bits. The quantized vector $\hat{\mathbf{w}}$ is

Table 1. Target vectors and number of levels and bits for each stage.

Stage	Target vector	Levels	Bits
1	$\mathbf{w}$	128	7
2	$\Delta\mathbf{w} = \mathbf{w} - \hat{\mathbf{w}}$	64	6
3	$\Delta\Delta\mathbf{w} = \Delta\mathbf{w} - \Delta\hat{\mathbf{w}}$	64	6
4	$\Delta\Delta\Delta\mathbf{w} = \Delta\Delta\mathbf{w} - \Delta\Delta\hat{\mathbf{w}}$	64	6
Total	25 bits		

obtained by

$$\hat{\mathbf{w}} = \hat{\mathbf{w}} + \Delta\hat{\mathbf{w}} + \Delta\Delta\hat{\mathbf{w}} + \Delta\Delta\Delta\hat{\mathbf{w}}, \tag{3}$$

where $\hat{\mathbf{w}}$,$\Delta\hat{\mathbf{w}}$,$\Delta\Delta\hat{\mathbf{w}}$,$\Delta\Delta\Delta\hat{\mathbf{w}}$ are the best vectors obtained in each stage.

One of the most important caracteristics of the VQ is the distortion measure used. For the MSVQ in this work it is used the weighted Euclidian distance

$$d(\mathbf{w}, \hat{\mathbf{w}}) = \sqrt{\sum_{i=1}^{10} W_i (w_i - \hat{w}_i)^2}, \tag{4}$$

where the weights $W_1, \ldots, W_{10}$ are given by

$$W_i = [|H(e^{jw_i})|]^{0.6}, \quad 1 \leq i \leq 10, \tag{5}$$

and $|H(e^{jw_i})|$ represents the magnitude response of $H(e^{jw})$ in the frequency w_i.

The search procedure is an M-best approximation to a full search, in which the $M = 8$ best indexes from each stage are saved for use with the next stage [8].

4.2. Split vector quantization

In this article many tests and splitting factors were performed concerning to SVQ [4, 8]. The VQ codebooks were divided in four different configurations: 5/5, 6/4, 4/6 and 4/3/3.

The search procedure in all configurations is very similar, beginning with the search in the first codebook using (4) to find the best vector. The weights are given by

$$W_i = \frac{1}{w_i - w_{i-1}} + \frac{1}{w_{i+1} - w_i}, \quad 1 \leq i \leq 10, \tag{6}$$

were $w_0 = 0.0$ and $w_{11} = 0.5$. Then, the search is performed in the second codebook, taking the precaution to keep the synthesis filter stable, i.e., the LSF must be in a growing order.

Table 2 shows the number of codevectors used in each codebook and the total number of bits used in every configuration implemented.

Table 2. Number of codevectors and bits used in each configuration.

Codevectors	Num. of Bits
5/5, 6/4, 4/6	
1024/1024	20
2048/1024	21
4096/1024	22
2048/2048	22
4096/2048	23
4/3/3	
256/128/64	21
128/128/128	21
256/128/128	22
256/256/128	23
512/128/128	23
256/256/256	24
512/256/128	24

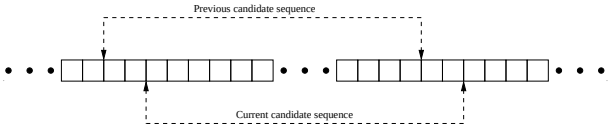

Fig. 3. Overllaping process.

5. THE EXCITATION PARAMETERS

5.1. The adaptive codebook

The adaptive codebook used has integer delays in a range from 20 to 148 samples. During the coding process, the output of $H(z)$ for each codevector $x_{aD}(n)$, that will be called $y_{aD}(n)$, is correlated to the signal that is wished to be reproduced, called $t(n)$. This correlation is obtained by

$$C = \sum_{n=0}^{N-1} y_{aD}(n)t(n), \tag{7}$$

where N is the number os samples of the sequence (the size of a subframe). Only codevectors that causes $C > 0$ are used. If $C \leq 0$ for all the codevectors of a subframe, the index D is set to zero to indicate the excitation for this subframe will be composed only by the fixed codebook. This procedure improves the reconstructed signal quality because it produces a more consistent evaluation of the delay.

5.2. The fixed codebook

The fixed codebook is composed of 1082 samples, forming 512 codevectors through the overlapping process, as shown in Figure 3.

Table 3. Bits allocation for the quantization of the excitation parameters.

Parameter	Range	Num. of bits
FC gain	-0,05 a 0,05	5
FC index	0 a 511	9
AC gain	0 a 2	4
AC index	0 a 127	7
Total	25 bits	

The fixed codebook was clipped in order to have 90% of zero samples in each codevector. Besides a good quality, this kind of codebook allows a faster search for the best excitation.

5.3. Gains and bits allocation

The gains G_a and G_f, calculated as described in [5], are scalar quantized with 4 and 5 bits, respectivally, after been optimized at the end of the analysis-by-synthesis procedure for each subframe. The non-uniform quantizer for each coefficient was obtained through the LLoyd training algorithm. Table 3 shows the way the excitation parameters are quantized.

6. RESULTS

The following tests were performed in a database composed of 20 sentences, spoken by 10 speakers, 5 male and 5 female. The choice of number of bits for the LSF per frame for each method is done in such a way that the subjective tests imply similar quality.

6.1. VQ evaluation

The evaluation of the SQ, MSVQ and SVQ was made in terms of spectral distortion (SD)

$$SD_i = \sqrt{\frac{1}{\pi}\int_0^{\pi} [10\log_{10}|H_i(e^{jw})| - 10\log_{10}|\hat{H}_i(e^{jw})|]^2 dw} \quad (dB), XS \tag{8}$$

where $|H_i(e^{jw})|$ and $|\hat{H}_i(e^{jw})|$ are respectively the magnitude responses of $1/A_i(e^{jw})$ and $1/\hat{A}_i(e^{jw})$ for the i-th frame. For sake of evaluation, it was used the average spectral distortion ($\bar{SD}$) and number of outliers [8].

Table 4 shows the test results. The following quantizations were compared: SQ with 32 bits, MSVQ using 25 bits and SVQ with the number of bits ranging from 20 to 24 divided into four configurations: 5/5, 6/4, 4/6 and 4/3/3.

It can be seen that SQ reaches the best quality mostly because this quantization employs 32 bits.

Table 4. Spectral distortion performance and for each implemented quantization.

Codevectors	SD	$\%SD \in [2;4]$	$\%SD > 4$
SQ			
-	1.08	3.15	0.00
MSVQ			
-	1.63	21.09	0.55
SVQ			
$5/5$			
$1024/1024$	1.48	11.72	0.05
$2048/1024$	1.40	9.14	0.00
$4096/1024$	1.34	6.92	0.00
$2048/2048$	1.31	5.86	0.00
$4096/2048$	1.25	4.44	0.00
$6/4$			
$1024/1024$	1.49	11.36	0.00
$2048/1024$	1.38	7.93	0.00
$4096/1024$	1.28	4.80	0.00
$2048/2048$	1.33	6.57	0.00
$4096/2048$	1.23	3.69	0.00
$4/6$			
$1024/1024$	1.60	19.95	0.15
$2048/1024$	1.56	18.59	0.15
$4096/1024$	1.53	16.72	0.15
$2048/2048$	1.43	11.77	0.00
$4096/2048$	1.39	10.71	0.00
$4/3/3$			
$256/128/64$	1.47	11.57	0.00
$128/128/128$	1.46	10.45	0.00
$256/128/128$	1.39	8.54	0.00
$256/256/128$	1.30	5.56	0.00
$512/128/128$	1.31	6.36	0.00
$256/256/256$	1.24	4.65	0.00
$512/256/128$	1.22	3.99	0.00

Table 5. PSSNR, CD and ID for the CELP coder with each implemented quantization for the synthesis filter parameters.

Codevectors	PSSNR(dB)	CD(dB)	ID(dB)
SQ			
-	17.61	2.95	1.08
MSVQ			
-	15.90	3.30	1.35
SVQ			
$5/5$			
$1024/1024$	15.99	3.28	1.33
$2048/1024$	16.05	3.25	1.31
$4096/1024$	16.11	3.24	1.31
$2048/2048$	16.09	3.24	1.30
$4096/2048$	16.14	3.20	1.27
$6/4$			
$1024/1024$	15.86	3.24	1.30
$2048/1024$	15.94	3.22	1.29
$4096/1024$	16.03	3.18	1.26
$2048/2048$	15.94	3.21	1.28
$4096/2048$	16.09	3.18	1.26
$4/6$			
$1024/1024$	16.11	3.32	1.37
$2048/1024$	16.19	3.32	1.38
$4096/1024$	16.24	3.32	1.37
$2048/2048$	16.26	3.26	1.32
$4096/2048$	16.28	3.26	1.32
$4/3/3$			
$256/128/64$	16.02	3.26	1.32
$128/128/128$	15.91	3.28	1.33
$256/128/128$	16.03	3.23	1.30
$256/256/128$	16.05	3.21	1.28
$512/128/128$	16.07	3.22	1.29
$256/256/256$	16.01	3.19	1.26
$512/256/128$	16.13	3.20	1.27

6.2. CELP coder evaluation

Three objective measures were used to evaluate the CELP coder: perceptual segmented signal-to-noise ratio (PSSNR), the cepstral distance (CD) and the Itakura distance (ID).

The objective results of the tests performed over the CELP coder implemented respectivally with SQ, MSVQ and SVQ are presented on Table 5.

It can be seen that the objective results of the coder implemented with SQ are better then the others because it uses more bits to the quantization of the synthesis filter coefficients.

6.3. Subjective tests

Two informal subjective tests were performed: the first with 22 people and the other with 20 people. In the first each of the 20 listeners heard two sentences, one quantized with the SQ and another with the MSVQ. They chose the best among them. The results of this test are in Table 6. They could choose one of the options: SQ better, MSVQ better or similar. It can be seen that the quality is almost every time the same between the SQ and the MSVQ. In the second test they first heard four sentences, one of each configuration of SVQ quantized with the highest number of bits. They could choose one as the best or they could say they are all similar. The results of this test are in Table 7. Then, after finding the best configuration ($5/5$), they heard six sentences, five quantized with the five different codebooks in that configuration and one scalar quantized. They could also choose one as the best or all similar. The results of this test are in Table 8.

Table 6. Results of the subjective test between the SQ and the MSVQ.

Sentence	MSVQ	Similar	SQ
M1	7	12	3
F1	4	16	2
M2	4	17	1
F2	2	12	8
M3	5	14	3
F3	1	19	2
M4	3	15	4
F4	2	14	6
M5	2	12	8
F5	1	18	3

Table 7. Results of the first subjective test with SVQ.

Sentence	$5/5$	$6/4$	$4/6$	$4/3/3$	Similar
M1	3	1	2	0	14
F1	1	3	2	2	12
M2	3	0	1	1	15
F2	4	0	2	1	13
M3	1	1	3	1	14
F3	2	2	1	2	13
M4	1	1	2	1	15
F4	2	0	1	1	16
M5	3	2	3	1	11
F5	2	1	1	1	15

The first subjective test showed that MSVQ has a quality very similar to the SQ with the advantage that it uses 25 bits against 32 bits used in the SQ. The first part of the second subjective test showed that $5/5$ configuration performs the best quality. The second part showed that using 22 bits the quality is still very similar to the SQ.

7. CONCLUSIONS

In this paper it was presented a CELP coder algorithm and the implementation of three different quantizations for the synthesis filter coefficients: scalar quantization (SQ) with 32 bits, multi-stage vector quantization (MSVQ) with 25 bits and the split vector quantization (SVQ) with number of bits ranging from 20 to 24. The objective tests showed that the SQ quantization with 32 bits per frame is the best one and the SVQ is better than MSVQ. The subjective tests showed that MSVQ using 25 bits has a similar quality when compared with SQ using 32 bits and that SVQ with the spliting factor $5/5$ using 22 bits has also a similar quality when compared with SQ. Due to this fact, the SVQ is implemented in this CELP coder, making the bit rate go from 4.40 kbps with SQ to 4.06 kbps with SVQ.

Table 8. Results of the second subjective test with SVQ.

Codevectors	No. of bits	Chosen
$1024/1024$	20	1
$2048/1024$	21	1
$4096/1024$	22	4
$2048/2048$	22	3
$4096/2048$	23	4
SQ	32	5
Similar	-	2

8. REFERENCES

[1] M. R. SCHROEDER AND B. S. ATAL. Code-excited linear prediction (CELP): high-quality speech at very low bit rates. In "Proceedings of the IEEE Int. Conf. Acoustics, Speech, and Signal Processing", pages 937–940, Tampa, FL (1985).

[2] N. SUGAMURA AND N. FARVARDIN. Quantizer design in LSP analysis-synthesis. *IEEE Journal on Selected Areas in Communications* **6**(2), 432–440 (Feb 1988).

[3] L.M. DA SILVA. "Contribuições para a melhoria da codificação CELP a baixas taxas de bits". PhD thesis, PUC-RJ (Feb 1996).

[4] K.K. PALIWAL AND B.S. ATAL. Efficient vector quantization of LPC parameters at 24 bits/frame. In "Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing", pages 661–664 (May 1991).

[5] R. S. MAIA, C. B. RIBEIRO, F. G. RESENDE, AND S. L. NETTO. Um sistema CELP para a codificação da fala a 4,4 kbps. In "Congresso Brasileiro de Automática - CBA2000", Florianópolis, SC (Set. 2000).

[6] MCCREE AND T.P. BARNWEL. A mixed excitation LPC vocoder model for low bit rate speech coding. *IEEE Transactions on Speech and Audio Processing* **3**(4), 242–250 (July 1995).

[7] MCCREE, K. TRUONG, E.B. GEORGE, T.P. BARNWELL, AND V. VISWANATHAN. A 2,4 kbits/s MELP coder candidate for the new u.s. federal standard. In "Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing", pages 200–203 (1996).

[8] W.B. KLEIJN AND K.K. PALIWAL, editors. "Speech Coding and Synthesis". Elsevier (1995).