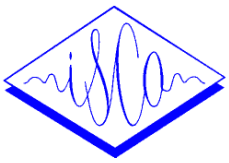

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

A LOW BIT-RATE CELP CODER BASED ON MULTI-PATH SEARCH METHODS

Maurizio Copperi

SIP Headquarters, R&D Department
Via San Dalmazzo 15, Torino, Italy 10122

ABSTRACT

This paper deals with the design and performance evaluation of a CELP codec at 3.5 kbit/s based on multi-path search methods. The rationale of this approach stems from the fact that delayed-decision coding algorithms outperform conventional techniques, in which the parameters to be transmitted are optimized sequentially. Tree codes have been applied to both the spectral information, quantized via a multi-stage vector quantizer, and the excitation codevectors, along with relative optimal gains. Preliminary results of this study show that the proposed scheme is a viable approach for achieving good speech quality at bit-rates below 4 kbit/s.

10.21437/ICSLP.1992-9

1. INTRODUCTION

Important applications, such as mobile radio communications, are undergoing rapidly expanding demands which require an efficient use of the available channel capacity. Therefore, advanced speech compression algorithms are currently studied to ensure high quality speech at rates around 4 kbit/s. Codebook Excited Linear Prediction (CELP) coders [1,2,3] are promising candidates to succeed in this difficult task, if specific techniques are incorporated in the basic analysis-by-synthesis scheme, in order to prevent quality degradations and artifacts that can be experienced in the low bit-rate region of interest.

Significant improvements of CELP speech quality at 4.0-4.8 kbit/s have been achieved exploiting some form of delayed decision coding, in which vectors and parameters describing the compressed information to be transmitted are globally optimized over each frame [4,5,6]. This is in contrast with conventional CELP schemes that perform a sequential optimization of the various parameters, disregarding any effect of the current choice over the successive ones. In fact, when the subframes are short (e.g. 10 ms or less), the ringing of filters, caused by previous excitation signals, affects significantly the parameter selection on the current subframe.

To cope with this problem, multi-path search coding can be applied, with different levels of computational complexity, to various excitation parameters, such as innovation vectors and long-term prediction parameters. This is done by storing a list of coding hypotheses across the subframes, and then a final choice is made by minimizing the cumulative distortion on the entire frame.

For what concerns the short-term parameters, a similar approach can also be used to minimize the spectral distortion of cascaded vector quantizers, where several candidate codevectors are listed for each stage and the best combination is eventually selected.

This paper reports on work that addresses the issue of multi-path search methods to design a good quality CELP coder at 3.5 kbit/s. In particular, these methods have been applied to the quantization of spectral parameters and to the selection of excitation vectors, along with relative gains.

2. LPC CASCADED VECTOR QUANTIZATION

In the proposed CELP scheme, the input speech is lowpass filtered at 3.4 kHz and sampled at 8 kHz. A 10-pole LPC analysis is performed on 30-ms signal frames, using the autocorrelation method (Hamming window) without preemphasis. Reflection coefficients are converted to log-area-ratio (LAR) parameters, to be quantized with a vector quantizer (VQ). Typically, a 24-bit VQ is necessary to quantize the spectral information without introducing perceptible distortions. In order to reduce storage as well as computational cost, a cascaded (multi-stage) structure must be used.

Codebooks of LAR vectors (first stage) and of difference vectors (subsequent stages) have been designed using a modified LBG algorithm with specific strategies to cope with outliers and empty cells, thus reducing the overall mean squared error (mse). The training sequence of the first stage is composed of 25,000 LAR vectors derived from twelve different speakers (six male and six female). This sequence is obtained by sliding the 30-ms analysis window along the speech signal by 15-ms steps (50 % overlap), so that a large number of spectra present in the speech database is captured.

In the coder simulation, the optimal codevectors can be selected with either a standard sequential technique or a multi-path search method, to be discussed next.

2.1 Single Path Cascaded VQ

In the single path or sequential case (no delayed decision), each codebook is full searched to find out the template which gives the minimum mse in that stage. The final quantized spectral vector is then simply the sum of the chosen codevectors. When designing a cascaded VQ, the first stage(s) should be allocated the maximum number of bits that can be reasonably afforded. We have investigated two different bit allocations in a 3-stage VQ, using (10,10,4) bits and (8,8,8) bits respectively. The performance of the two multi-stage VQs considered here is depicted in Table I, in terms of spectral distortion for two out-of-training speakers. The decibel spectral distortion is computed as

$$SD = 6.142\ (d_{LR})^{0.5} \qquad (1)$$

where d_{LR} is the average likelihood ratio measure. For a given speech frame, the likelihood ratio distortion measure is

defined by

$$d'_{LR} = (\mathbf{a}^T\mathbf{R}\mathbf{a} / e) - 1 \qquad (2)$$

where **R** is the autocorrelation matrix of the input speech, e is the corresponding LPC residual energy and **a** is the quantized LPC vector.

The better performance of the first VQ with (10,10,4) bits is obtained at the cost of higher storage and computational complexity. In particular, the number of distortion calculations per input vector, for the two VQs, is 2064 and 768 respectively.

Table I - Spectral distortion SD of 3-stage VQs with single path search

Speaker	VQ [bit]	SD [dB]	Frames with SD>2 dB
male-7	10,10,4	1.08	2.1 %
female-7	10,10,4	0.98	2.0 %
male-7	8,8,8	1.37	8.4 %
female-7	8,8,8	1.17	4.3 %

2.2 Multi-Path Cascaded VQ

An experiment has been carried out to verify the effectiveness of the multi-path search in the preceding 3-stage VQs of LARs. The 2nd and 3rd stage should now be trained with a sequence containing all the candidates corresponding to each vector fed into the previous stage (i.e. the 1st and 2nd stage, respectively). In other words, the training procedure should be performed using the same structure chosen for the actual quantization process. Several tree codes have been examined, computing their performance versus complexity curves. Considering the relatively large size of the codebooks, an acceptable quality/complexity trade-off is provided by a binary tree, in which two best candidates are selected for each input vector at both the 1st and 2nd stage. In the 3rd stage, only one branch per node, populated with the nearest neighbor, is determined, so that the code is formed by (2,4,4) nodes. The distance measure used to full search each codebook is again the squared Euclidean distance between vectors, as in the sequential case, but now the optimal path, among the four possible alternatives stored in the map, is chosen by minimizing the spectral distortion SD.

The results for out-of-training speakers, summarized in Table II, show that the multi-path strategy with binary tree search provides, in comparison to the sequential method, a distortion reduction of 0.10-0.14 dB and, more importantly, a significant reduction of frames with SD greater than 2 dB. The distortion computations for the two multi-stage VQs are now 3136 and 1792 respectively.

To reduce the memory requirements of previous multi-stage VQs, we have also attempted a 4-stage VQ incorporating smaller codebooks of 6-bit size. In this case, to get a low average distortion, it is necessary increasing the number of nodes and branches per stage in the tree code. An efficient method is given by the M-L algorithm, in which only M best candidates per stage (those yielding the lowest cumulative distortions) are retained. However, even for M=8 the performance was worse than that of the (8,8,8)-bit VQ. Therefore, the (10,10,4)-bit VQ has been selected for the codec simulation.

Table II - Spectral distortion SD of 3-stage VQs with delayed decision based on a binary tree

Speaker	VQ [bit]	SD [dB]	Frames with SD>2 dB
male-7	10,10,4	0.99	0.9 %
female-7	10,10,4	0.88	0.3 %
male-7	8,8,8	1.24	3.6 %
female-7	8,8,8	1.03	0.9 %

3. LONG-TERM PREDICTION

Pitch parameters are computed on the actual LPC residual signal (open loop analysis) every 10 ms. A 3rd order prediction filter has been chosen in order to obtain an interpolated non-integer effective lag and get an SNR improvement of about 1.5 dB over a single-tap filter. The predictor coefficients are treated as vectors and quantized using a 6-bit VQ. Pitch lags are coded with 7 bits in the range 30-120.

4. EXCITATION AND GAIN

4.1 Excitation Codebook

The excitation codebook used in this study has a dimension equal to the subframe length, that is 80 samples, and a size fixed at 9 bits. Two different classes of templates have been tested, based on either ternary or residual-like samples.

The employed ternary excitation is a kind of regular pulse excitation in which every pulse can only be equal to 1 or -1. Since the constant duration between any two consecutive pulses is filled with zeroes, the codevectors look like upsampled versions of a full binary sequence with samples of amplitude 1 and -1. Important benefits of this structure are a fast search of the codebook and a small storage requirement. This class of excitation can be specified by a shape codebook and the phase, that indicates the position of the first pulse. Two ternary structures are investigated, the former based on 8 bits for the shape codebook (1:2 upsampling) and 1 bit for the phase, the latter using 7 bits for the shape (1:4 upsampling) and 2 bits for the phase. The sign of each pulse is determined by a random-number generator during the codebook design.

A different excitation source is obtained by populating a codebook with normalized residual samples, extracted from a speech database. This way, in addition to noise-like vectors, we can also include typical pitch pulse waveforms, time-shifted along the vector length, in order to feed the synthesis filter with a better excitation signal when the long-term predictor is unable to produce a correct output (for example, at the beginning of voiced segments). The residual-like codebook is designed through a particular procedure described in [7]. Codevectors are center-clipped in order to increase their sparseness and lower the search effort. The performance of these codebooks is discussed in section 5.

4.2 Multiple Gain

When using vectors of 10 ms in duration, one single gain value per vector may be critical, because the possible level mismatching on shorter speech segments can cause perceptible artifacts. We have devised an effective procedure in which every codevector half is scaled by the corresponding rms value of the actual LPC residual, prior to enter the weighted synthesis loop. Thus, for each subframe we compute only one correlation between the weighted synthetic speech and the target signal, as in standard CELP schemes, but using codevectors pre-scaled on a 5-ms basis according to the energy of the actual excitation signal. Using the rms values computed on the LPC residual is advisable to take advantage of the presence of pitch pulses (higher energy epochs).

The gain quantization scheme is as follows. For each frame (30 ms), we compute one rms value R every 5-ms segment of the LPC residual, and the mean G of these six terms. The mean value G', version of G quantized in the log domain with a 5-bit quantizer, is used to normalize the gain vector g for each template

$$g = (R_j\ S/G',\ R_{j+1}\ S/G') \qquad (3)$$

where S is the absolute value of the normalized correlation between the weighted synthetic speech and the target signal for the given subframe, and R_j, R_{j+1} are the proper open-loop scaling factors. Vectors g are quantized with a 3-bit VQ in the log domain.

4.3 Multi-Path Search

Optimal codevectors and gains are selected via a multi-path method over the entire 30-ms frame. In particular, the tree code, used in our scheme to implement the delayed decision, retains the four best excitation signals in the first subframe (four nodes) and eight corresponding signals in the second subframe (two branches per node). These eight branches are then pruned to keep only the four minimum distortion paths. At the last subframe, the survived branches are extended computing only one optimal excitation signal per branch, in order to complete the search within the current frame, thus avoiding additional coding delay. This procedure yields better speech quality and an average SNR improvement of 1.4 dB in comparison to instantaneous (conventional) coding of each subframe.

5. RESULTS AND CONCLUSIONS

A fully quantized CELP codec, based upon the multi-search methods described in the previous sections, has been simulated to evaluate its performance. Coding parameters and relative bit allocations, for a total rate of about 3.5 kbit/s, are shown in Table III. Spectral parameters are quantized with the (10,10,4)-bit VQ.

As described in section 4, three excitation codebooks have been tested in the same CELP algorithm, namely a ternary codebook with 7-bit shape and 2-bit phase (CELP-A), a ternary codebook with 8-bit shape and 1-bit phase (CELP-B), and a residual-like codebook with 9-bit shape (CELP-C). Table IV shows the segmental SNR, computed between original and reproduced signals, for out-of-training speech material. Since the coder exploits a perceptual weighting factor of 0.85, the SNR figures are lower than those achievable without the noise shaping mechanism. However, the perceptual quality is improved significantly.

Table III - Coding parameters and bit allocation

Parameter	bit/frame
Excitation codebook	27
Average gain G	5
Multiple gain g	9
Excitation sign	3
Pitch lag	21
Pitch predictor taps	18
Spectrum	24
Total	**107**

Table IV - Coder performance in terms of segmental SNR

Coder	SNRseg [dB]
CELP-A	7.9
CELP-B	8.5
CELP-C	9.4

Worth noting is that the drop in SNR from CELP-C to CELP-A, that makes use of the simplest codebook, is of 1.5 dB. In order to get more insights into the trade-off between quality and complexity, we have carried out a further test to see how the speech quality depends on the bit allocation between shape and gain. An 8-bit shape codebook, populated with residual-like samples, has been designed along with a 4-bit VQ for the multiple gain g. The resulting bit-rate is the same as before, but this shape codebook requires a search over half the number of vectors in comparison to CELP-C. The performance of the CELP algorithm with this excitation is 8.7 dB, that is almost equivalent to the CELP-B scheme. Informal listening tests have revealed that the four systems provide good reproduced speech, with slight perceived difference in quality between the codebook structures.

From these experiments, it is clear that the optimal choice between the examined excitation schemes should be done taking also into account the relative computational complexity and the capacity of the available hardware.

In conclusion, three major issues have been investigated. The first is the efficient vector quantization of the spectral parameters, exploiting multi-stage VQ with delayed decision based on tree codes. The second is the evaluation of various excitation structures along with optimal gain quantization, and the third is the determination of the quality enhancement given by a multi-path search method, applied to the excitation on a frame-by-frame basis. The result of this study is a CELP codec at 3.5 kbit/s that seems to be a viable approach for achieving good speech quality, suitable for many important applications ranging from mobile radio systems to voice mail services.

References

[1] M.Copperi and D.Sereno, "Improved LPC excitation based on pattern classification and perceptual criteria", Proc. Seventh International Conference on Pattern Recognition, pp. 860-862, Montreal, 1984

[2] M.Copperi and D.Sereno, "Vector quantization and perceptual criteria for low-rate coding of speech", Proc. ICASSP, pp. 252-255, Tampa, 1985

[3] M.Copperi and D.Sereno, "CELP coding for high quality speech at 8 kbit/s", Proc. ICASSP, pp. 1685-1688, Tokyo, 1986

[4] K.Mano and T.Moriya, "4.8 kbit/s delayed decision CELP coder using tree coding", Proc. ICASSP, pp. 21-24, Albuquerque, 1990

[5] K.Ozawa and T.Miyano, "4 kbit/s improved CELP coder with efficient vector quantization", Proc. ICASSP, pp. 213-216, Toronto, 1991

[6] H.Su and P.Mermelstein, "Delayed decision coding of pitch and innovation signals in code-excited linear prediction (CELP) speech codecs", IEEE Workshop on Speech Coding for Telecommunications, Whistler, 1991

[7] M.Copperi, "Efficient excitation modeling in a low bit-rate CELP coder", Proc. ICASSP, pp. 233-236, Toronto, 1991