A Low-Delay 8 Kb/s Backward-Adaptive CELP Coder

L.G. Neumeyer, W.P. LeBlanc and S.A. Mahmoud
Dept. of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, K1S 5B6, Canada
Phone: 613-788-5753
Fax: 613-788-5727

ABSTRACT*

Code-excited linear prediction coding is an efficient technique for compressing speech sequences. Communications quality of speech can be obtained at bit rates below 8 Kb/s. However, relatively large coding delays are necessary to buffer the input speech in order to perform the LPC analysis. In this paper we introduce a low–delay 8Kb/s CELP coder in which the short-term predictor is based on past synthesized speech. A new distortion measure that improves the tracking of the formant filter is discussed. Formal listening tests showed that the performance of the backward-adaptive coder is almost as good as the conventional CELP coder.

INTRODUCTION

Recent advances in linear prediction coding have made it possible to achieve communications quality of speech at bit rates below 8 Kb/s. Practical real-time implementations are possible due to efficient algorithms based on Code-Excited Linear Prediction (CELP) [1]. In these coders, the residual is vector quantized using an analysis-by-synthesis search procedure. The excitation vector (or codevector) is chosen from a large codebook. All the codevectors are passed through the synthesis filters and compared with the original speech vector. The index of the codevector that minimizes an objective distortion measure between original and quantized speech is sent through the channel. The parameters of the synthesis filters (gain, long- and short-term LPC coefficients, and pitch lag) are sent to the decoder as side information. Gain, pitch lag, and long-term predictor coefficients can be optimized using closed-loop procedures. Ideally, the formant filter could also be optimized in a closed-loop procedure but this would lead to a mathematically untractable set of non-linear equations [2]. Therefore, the short-term predictor coefficients are calculated using an open-loop solution based on the original speech. In order to obtain a reliable linear prediction filter, approximately 20 ms of speech samples are buffered. The one-way delay of the coder, although highly dependent on real-time implementations, could be as high as 60 ms. The delay could be reduced by using only past speech (no buffering). However, the linear prediction analysis would be unreliable, resulting in poor speech quality. This problem can be overcome by updating the LPC parameters at a higher rate. This would require more bits/sample, thereby increasing either the total bit rate or the distortion.

In this paper, we present an 8 Kb/s CELP coder in which the short-term linear prediction parameters are updated in a backward-adaptive manner. That is, the linear prediction analysis is performed on past synthesized speech which is available, assuming no transmission errors, at both ends of the channel. Therefore, the LPC parameters are not sent through the channel and can be updated at high rates, even in a sample-by-sample basis. Speech quality is as good as in the conventional (or forward-adaptive) coder. A new distortion measure is introduced to prevent predictor mistracking.

A diagram of the encoder is shown in figure

* This work has been sponsored by the Telecommunication Research Institute of Ontario (TRIO).

1. The synthesis filters are separated into their zero-input and zero-state components. The minimum pitch is constrained to be always greater than the block size. Therefore the transfer function of the zero-state pitch synthesis filter is unity. The weighting filter is moved from its original location (filtering the error between original and synthesized speech) to both of its input branches.

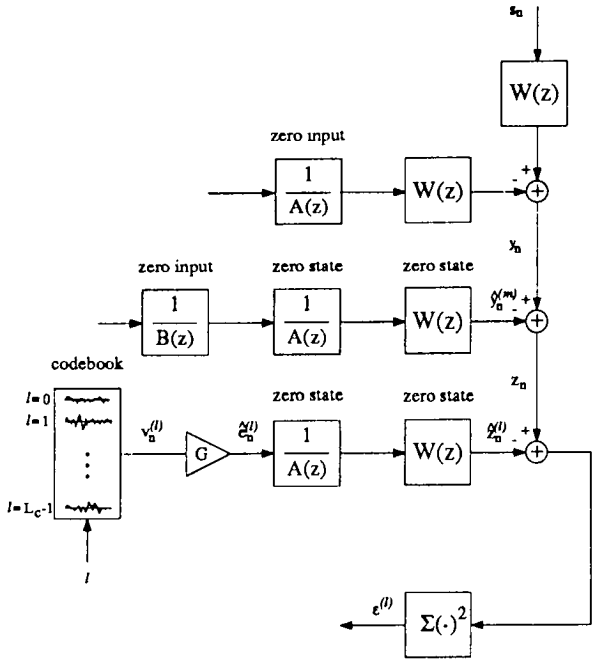

Figure 1 CELP Encoder

CODER DESIGN

Backward adaptation

In pure backward-adaptive 16 Kb/s CELP coders [3] [4] only the excitation vector index is sent through the channel. The rest of the parameters are computed in a backward-adaptive mode. The vector dimension (block of samples) is 4–5 samples and the one-way delay is below 2 ms. Unfortunately, as the bit rate decreases the quantization effects become more pronounced, leading to poor filter tracking and to further distortion of the original speech. As a result, in the BA-CELP coder only the short-term predictor is computed in a backward-adaptive manner. Pitch filter parameters and gain are optimized in closed-loop procedures and sent through the channel as side information. The three-tap pitch filter plays an important role, not only in the fine structure but also in the shape of the spectrum of the reconstructed speech.

In our BA-CELP coder, all the parameters are updated at the end of each block of samples. The bit allocation scheme is shown in table 1. The vector dimension is 26 samples and the sampling rate is 8 Khz. Consequently, the total delay (typically 4 times the vector dimension) is around 13 ms. The short-term LPC analysis is performed using the modified covariance method. The length of the frame is four times the vector dimension. Note that the frames are highly overlapped and relatively short. This is necessary to improve the tracking of the adaptive filter, specially when rapid changes of the spectrum occur. The autocorrelation method proved to be inefficient in this application. This is because the windowing process weights the error in the middle of the frames higher than at the edge of the frames. As a result, spectral match is poor in regions of rapid spectrum variations.

Parameter	bits/block	Kbits/sec
Formant filter	0	0.0
Pitch filter	5	1.5
Pitch lag	7	2.2
Gain	5	1.5
Excitation vec.	9	2.8
Total	26	8.0

Table 1 Bit allocation and corresponding bit rate. The vector dimension is 26 samples and the sampling rate is 8 Khz.

Perceptual weighting filter

Psychoacoustical studies show that the human auditory system can tolerate more errors in the formants of the speech spectrum than in the valleys. Therefore, we can obtain a more subjective distortion measure by weighting the spectrum of the error. Regions of the error spectrum that correspond to valleys between formants in the speech spectrum are de-emphasized and regions corresponding to the formants are emphasized. Using a weighting function $W(z)$ we can

write,

$$\epsilon_w = \frac{1}{2\pi}\int_0^{2\pi} \left|S\left(e^{j\omega}\right) - \hat{S}\left(e^{j\omega}\right)\right|^2 W\left(e^{j\omega}\right) d\omega \quad (1)$$

where ϵ_w is the noise-weighted mean-squared error (NWMSE). A general weighting filter is discussed in [3] and [5].

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \qquad 0 < \gamma_2 < \gamma_1 \leq 1 \quad (2)$$

A good choice for the parameters is γ_1=*0.9* and γ_2=*0.4*. Note that in conventional CELP coders only one LPC analysis is necessary for the synthesis and weighting filters. Conversely, the backward-adaptive approach "requires" two separate LPC analyses. One based on reconstructed speech for the synthesis filter and the other one based on the original speech for the weighting filter.

Mixed distortion

Further improvement in filter tracking can be achieved by using a *mixed distortion measure* in the excitation vector search procedure. The proposed mixed distortion combines mean-squared error with a subjectively meaningful LPC distortion measure.

Log-likelihood ratio distortion measure.

In linear prediction theory, the minimum residual energy for a particular speech frame is given by

$$\alpha = r_0 - \mathbf{a}^T\mathbf{r} \quad (3)$$

where $\mathbf{r}$ is the correlation vector, r_0 is the energy of the segment and $\mathbf{a}$ is the optimum LPC coefficient vector. If the same frame is passed through a non-optimum inverse filter then the residual energy β must be greater than α,

$$\beta = r_0 - 2\hat{\mathbf{a}}^T\mathbf{r} + \hat{\mathbf{a}}^T\mathbf{R}\hat{\mathbf{a}} \geq \alpha \quad (4)$$

where $\mathbf{R}$ is the correlation matrix. Equality holds when $\mathbf{a} = \hat{\mathbf{a}}$. The log-likelihood ratio (LLR) distortion measure is defined as

$$d_{LLR}\left(A(z), \hat{A}(z)\right) = \log\left(\frac{\beta}{\alpha}\right) \quad (5)$$

which is equivalent to the difference of the logarithmic prediction gains. The LLR distortion measure has proved to be subjectively meaningful [6][7].

Figure 2 shows the filtering operation. The two input sequences are *s(n)* and $\hat{s}$*(n)*. The corresponding p^{th} order inverse filters are *A(z)* and $\hat{A}$*(z)*. When one of the input sequences, say *s(n)* is passed through the filters, the resulting residual energies are α and β. A different distortion would be obtained by using $\hat{s}$*(n)* instead of *s(n)* as the input sequence. This difference shows the asymmetric nature of the likelihood ratio.

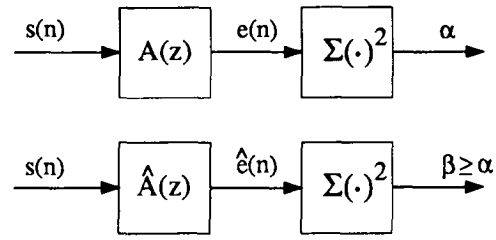

Figure 2 Computation of the residual energies in the log-likelihood ratio distortion measure.

Search procedure. The optimum excitation vector is searched in two sequential steps. First, a conventional search algorithm finds the best n_c excitation vectors that minimizes the NWMSE. The best n_c candidates are used in the second stage in order to minimize the mixed distortion measure.

The convolution of the l^{th} codevector $\mathbf{v}(l)$ with the impulse response of the weighted synthesis filter can be written in matrix form as,

$$\hat{\mathbf{z}}(l) = G\mathbf{H}\mathbf{v}(l) \quad (6)$$

where G is the gain, $\mathbf{H}$ is a lower triangular toeplitz matrix containing the impulse response in its first column and K is the vector dimension,

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & 0 & \cdots & 0 \\ h_1 & h_0 & 0 & \cdots & 0 \\ h_2 & h_1 & h_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{K-1} & h_{K-2} & h_{K-3} & \cdots & h_0 \end{bmatrix} \quad (7)$$

The NWMSE is given by,

$$\epsilon_w(l) = \|\mathbf{z} - \hat{\mathbf{z}}(l)\|^2$$
$$= \|\mathbf{z}\|^2 - 2G\mathbf{z}^T\mathbf{H}\mathbf{v}(l) + G^2\mathbf{v}^T(l)\,\mathbf{H}^T\mathbf{H}\mathbf{v}(l) \quad (8)$$

Taking the derivative with respect to G we get the minimum NWMSE and the optimum gain for the l^{th} codevector,

$$\frac{\partial \epsilon_w}{\partial G} = -2\mathbf{z}^T\mathbf{H}\mathbf{v}(l) + 2G\mathbf{v}^T(l)\mathbf{H}^T\mathbf{H}\mathbf{v}(l) = 0$$
$$\Rightarrow \quad G_{opt}(l) = \frac{\mathbf{z}^T\mathbf{H}\mathbf{v}(l)}{\mathbf{v}^T(l)\mathbf{H}^T\mathbf{H}\mathbf{v}(l)}$$
$$\Rightarrow \quad \epsilon_{w_{min}}(l) = \|\mathbf{z}\|^2 - \frac{\left(\mathbf{z}^T\mathbf{H}\mathbf{v}(l)\right)^2}{\mathbf{v}^T(l)\mathbf{H}^T\mathbf{H}\mathbf{v}(l)} \quad (9)$$

In order to find the n_c optimum excitation vectors out of the L-level codebook it is necessary to maximize the second term of the minimum error:

$$find \quad l_{opt} \Rightarrow \max_{l=0\ldots L-1} \frac{\left(\mathbf{z}^T\mathbf{H}\mathbf{v}(l)\right)^2}{\mathbf{v}^T(l)\mathbf{H}^T\mathbf{H}\mathbf{v}(l)} \quad (10)$$

The computational complexity of the search is reduced by using a shift-symmetric codebook [8].

In the second stage of the search, the optimum codevector is chosen out of the n_c candidates. The objective is to choose a codevector that minimizes the distortion between the original LPC model $A(z)$ and the backward-adaptive LPC model $\hat{A}(z)$ one vector into the future. The original LPC model has already been computed for the weighting filter. To calculate the corresponding $\hat{A}(z)$ for each candidate, we compute the next block of speech samples and perform the LPC analysis on the updated synthesized speech sequence. The mixed distortion is defined as,

$$d_{mix}^{(i)} = d_{LLR}^{(i)}\left(A(z), \hat{A}(z)\right) + \eta \log \frac{\epsilon_w^{(i)}}{\epsilon_w^{(min)}}$$
$$i = 1 \ldots n_c \quad (11)$$

where $\epsilon_w^{(min)}$ is the minimum NWMSE of the candidates and η is a parameter to be optimized in subjective tests. In equation 11, as η goes to infinity the mixed distortion measure becomes equivalent to the NWMSE. On the other hand, as η approaches zero the LPC distortion of future frames decreases at the cost of accuracy in the current block of samples.

SIMULATION RESULTS

Computer simulations results were obtained for the BA-CELP coder and for a conventional forward-adaptive version. The conventional 8 Kb/s CELP coder computed the LPC analysis on 20 ms of buffered speech. The autocorrelation method was used to calculate the LPC coefficients which were transformed to linear-spectrum pairs and quantized. For the BA-CELP coder we used the mixed distortion parameters η=*1* and n_c=*16*. For these values, the NWMSE was greater than the minimum in 20% of the speech blocks. The shift-symmetric excitation codebook was optimized using a 10–minute speech database.

Formal listening tests were conducted following the CCITT recommendations in [9]. The stimulus material contained six different sentences spoken by different males and females. Twenty listeners evaluated speech quality under five different conditions, 2 coders and 3 references. The reference conditions consisted of the original speech (PCM 64 Kb/s) and speech corrupted with random noise which has amplitude proportional to the instantaneous signal amplitude. The distorted speech is specified according to the modulated noise reference unit (MNRU) [10]. Mean opinion scores and 95% confidence intervals are shown in table 2. According to our results, speech quality in the forward-adaptive coder is only 0.1 points in the MOS scale better than the BA-CELP coder.

Condition	Mean	Error
PCM 64 Kb/s	4.24	0.15
Forward	3.42	0.20
Backward	3.33	0.19
MNRU 20 dB	2.39	0.18
MNRU 15 dB	1.68	0.17

Table 2 Mean opinion score test.
Mean and 95% confidence intervals.

Figure 3 shows noise-weighted signal-to-noise ratio as a function of η for a 30 second segment of speech. The dashed line represents the NWSNR for the conventional search (no LLR distortion). Observe that for values of η between 0.2 and 3 the global NWSNR is greater than for the regular search NWSNR. This shows how the global NWMSE was reduced by using the suboptimal (in a NWMSE sense) mixed distortion measure. In other words, an increase in the error of one block of samples helps in filter tracking and therefore improves the overall performance of the coder. Figure 4 shows the log-likelihood ratio distortion measure for different values of η.

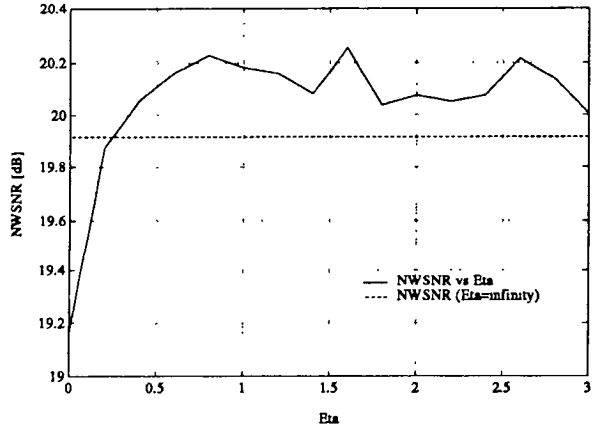

Figure 3 Noise-weighted signal-to-noise ratio versus η.

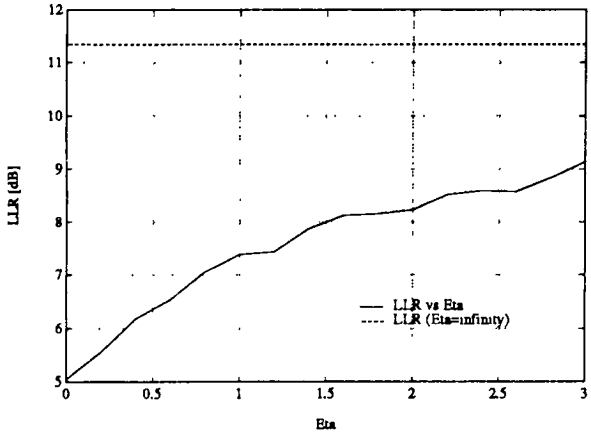

Figure 4 Log-likelihood ratio distortion vs η.

CONCLUDING REMARKS

In this paper we discussed how a delay of approximately 13 ms is achieved in the BA-CELP. Based on subjective MOS tests, speech quality has been found to be comparable to that of conventional forward-adaptive CELP coders. However, several LPC analyses are necessary to compute the mixed distortion measure. The number of candidates in the search procedure determines the computational complexity of the coder. Further reductions in complexity may be possible by using recursive LPC algorithms instead of block algorithms. To further reduce the delay, future investigation would include backward-adaptation of the remaining parameters.

REFERENCES

[1] M. Schroeder and B. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 937–940, 1985.

[2] P. Kabal, J. Moncet, and C. Chu, "Synthesis Filter Optimization and Coding: Applications to CELP," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 147–150, 1988.

[3] J. Chen, "A Robust Low-Delay CELP Speech Coder," *IEEE Global Communications Conference*, pp. 1237–1241, 1989.

[4] V. Cuperman, A. Gersho, R. Pettigrew, J. Shynk, and J. Yao, "Backward Adaptation for Low Delay Vector Excitation Coding of Speech at 16 Kbit/s," *IEEE Global Communications Conference*, pp. 1242–1246, 1989.

[5] J. Chen and A. Gersho, "Real-Time Vector APC Speech Coding at 4800 bps with Adaptive Postfiltering," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2185–2188, 1987.

[6] A. Gray and J. Markel, "Distance Measures for Speech Processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 5, pp. 380–391, October 1976.

[7] B. Juang, "On Using the Itakura-Saito Measures for Speech Coder Performance Evaluation," *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 8, pp. 1477–1498, October 1984.

[8] W. Kleijn, D. Krasinski, and R. Ketchum, "Improved Speech Quality and Efficient Vector Quantization in SELP," *IEEE Inter-*

national Conference on Acoustics, Speech and Signal Processing, pp. 155–158, 1988.

[9] CCITT Supplement No. 14 to Recommendation P.81, Blue Book, *Subjective Performance Assessment of Digital Processes Using the Modulated Noise Reference Unit*, 1988.

[10] CCITT Supplement No. 14 to Recommendation P.81, Blue Book, *Modulated Noise Reference Unit (MNRU)*, 1988.