S1.5

WIDEBAND CELP SPEECH CODING AT 16 KBITS/SEC

Guylain Roy[1] and Peter Kabal[1,2]

[1] Electrical Engineering
McGill University
Montreal, Quebec, H3A 2A7

[2] INRS-Télécommunications
Université du Québec
Verdun, Quebec, H3E 1H6

Abstract

This paper investigates the use of CELP (Code Excited Linear Prediction) in coding wideband speech signals at an operating rate of 16 kbits/sec The wideband signals under consideration are bandlimited to 7500 Hz and sampled at 16 kHz. In order to achieve a low operating rate, the coding places more emphasis on the lower frequencies (0 – 4 kHz), while the higher frequencies are coded less precisely, but with little perceived degradation To this effect, the basic CELP model is modified to operate in a split-band mode

1. Introduction

In recent years, CELP coders have been developed for narrowband systems, and have achieved high quality speech reproduction at rates from 4 8 kbits/sec to 9 6 kbits/sec [1] However, since roughly 80% of the perceptually important speech spectral information lies within the baseband (0 2 – 3.2 kHz) [2], it is reasonable to assume that the incremental cost of coding the extra bandwidth found in a wideband signal should be relatively small The added bandwidth increases the perceived speech quality, and helps discriminate between fricatives (e g "f" vs "s") Potential applications for this type of coder include mobile telephone, high-quality videoconferencing and voice-mail services

In this paper, the basic CELP structure is first reviewed Then, the mathematical derivation for the more general split-band CELP structure is presented This structure can operate in either split or full-band mode In Section 5, both structures are compared while subjected to an operating rate of 16 kbits/sec Various parameter coding issues are discussed and simulation results are presented Finally, both structures are subjectively compared to a 16 kbits/sec narrowband coder

2. Basic CELP coding

CELP coding falls in the ***analysis-by-synthesis*** category of linear predictive systems These coders offer a full parametric representation of speech signals, and can produce communications quality output at rates as low as 4 8 kbits/sec [3,4] The term ***analysis-by-synthesis*** means that the speech coding analysis is done at the transmitter by synthesizing speech signals using pre-determined synthesis parameters (i e quantized LPC coefficients, lag values, pitch parameters and residual waveforms) The synthesis parameters that yield the best match between the original and coded speech signals are sent to the receiver

In CELP coders, since both formant (long-term) and pitch prediction (short-term) are used, the residual excitation signal is noise-like. The residual waveform is coded using B bits pointing to an entry in a codebook of 2^B waveforms A simple CELP coder structure is shown in Figure 1.

An LPC analysis is first used to obtain the LPC coefficients a_k At the synthesis stage, the formant frame (e.g 20 ms) is divided into pitch sub-frames (e g 5 ms) For each sub-frame, the parameter selection is performed by scanning the codebook, one waveform $\hat{r}_i(n)$ at a time For each waveform, the gain G, the pitch lag M and the pitch coefficients β_i are computed such that the weighted error signal $e_w(n)$, defined below, is minimized in the mean-square sense The index of the waveform yielding the smallest error energy is sent to the receiver, along with the other synthesis parameters. The scaled excitation waveform $\hat{r}_i(n)$ is fed into the pitch synthesis filter $G(z) = 1/(1 - P(z))$ to generate the formant residual signal $\hat{d}(n)$, which in turn excites the formant synthesis filter $H(z) = 1/(1 - F(z))$ to yield the coded speech signal $\hat{s}(n)$ (lower branch of Fig 1) The formant and pitch prediction filters are respectively defined as

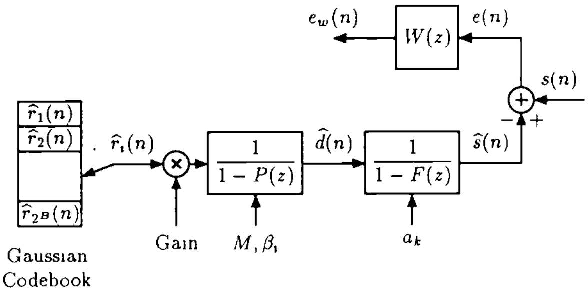


Fig. 1 Basic CELP coder

$$F(z) = \sum_{k=1}^{N_f} a_k z^{-k}, \tag{1}$$

and

$$P(z) = \sum_{k=1}^{N_p} \beta_i z^{-(M+i)}, \tag{2}$$

where N_f and N_p are the number of LPC and pitch coefficients respectively

The weighted error signal $e_w(n)$ is obtained by passing the error

$$e(n) = s(n) - \hat{s}(n)$$

through the noise shaping filter $W(z)$, defined as

$$W(z) = \frac{H(\gamma z)}{H(z)} = \frac{H'(z)}{H(z)} \tag{3}$$

where the bandwidth expansion factor, $\gamma = 1/0\ 75$, effectively concentrates the coding noise in the formant regions where it is not as perceptible [5]

The resulting speech quality is a function of the codebook size and parameter selection Codebooks containing as little as 32 waveforms can yield communications quality coded speech From a practical standpoint, a fully optimal parameter selection is not possible In particular, the LPC coefficients cannot be easily optimized, due to the feedback of the formant synthesis filter They must remain as derived at the analysis stage However, the pitch parameters (gain, lag and coefficients) can be re-optimized. When the sub-frame size is smaller than the pitch lag, the pitch lag in the pitch synthesis feedback loop reduces the optimization to solving a set of linear equations

3. Split-band CELP

Consider the split-band CELP model shown in Figure 2 The codebooks have been left out for clarity while the noise shaping filter $W(z)$ has been absorbed into each branch The excitation source now consists of two separate signals $\hat{r}_L(n)$ and $\hat{r}_H(n)$, respectively for the

CH2977-7/91/0000-0017 $1.00 © 1991 IEEE

low and high band. For generality, each band is given its own set of gain, lag and pitch synthesis parameters. In previously reported work on split-band CELP [6], each sub-band had its own spectral model (using QMF filters) and a dynamic bit allocation scheme was used. In this research, the spectral envelope is derived from the full-band speech signal, is coded using a fixed bit allocation and there is no down-sampling

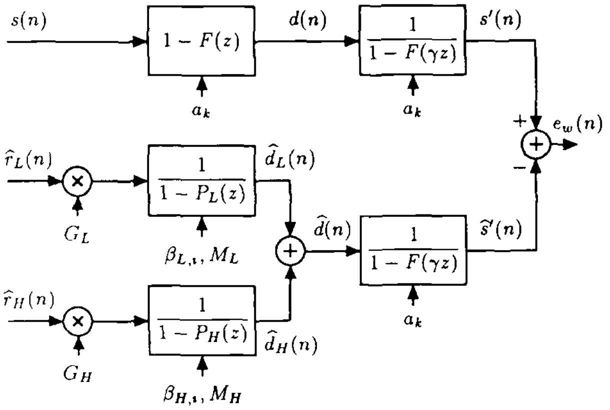


Fig. 2 Split-band CELP structure

The parameters subjected to optimization are the low and high band pitch coefficients $\beta_{L,i}$ and $\beta_{H,i}$, pitch lags M_L and M_H and gain factors G_L and G_H This structure provides flexible control over the number of pitch coefficients in each band, N_{p_L} and N_{p_H}

3.1 Mathematical description

Given the excitation signals, the goal is to minimize the energy of the error $e_w(n)$ for every pitch sub-frame of N samples. This error is the difference between the bandwidth expanded original and reconstructed speech signals $s'(n)$ and $\hat{s}'(n)$

$$\begin{aligned} s'(n) &= \sum_{k=-\infty}^{\infty} d(k)h'(n-k), \\ \hat{s}'(n) &= \sum_{k=-\infty}^{\infty} \hat{d}(k)h'(n-k). \end{aligned} \tag{4}$$

where $h'(n)$ is the impulse response of the bandwidth expanded formant synthesis filter $H'(z)$ This impulse response is time-varying. However, since the minimization is done at the pitch sub-frame level, $h'(n)$ is fully known and held constant for the duration of the sub-frame. The signal $s'(n)$ is also known for the duration of the sub-frame. Therefore, the summation limits can be changed to 0 and $N-1$, provided the contributions of past sub-frame excitation samples (i.e. $k < 0$) are preserved as initial conditions for the current sub-frame. This is achieved by saving the formant synthesis filter internal memory from one sub-frame to the next

The ouputs of the low and high band pitch synthesis filters are combined to form the regenerated formant residual signal $\hat{d}(n)$, expressed as

$$\begin{aligned} \hat{d}(n) = {} & G_L\hat{r}_L(n) + \sum_{i=1}^{N_{p_L}} \beta_{L,i}\hat{d}_L(n - M_L - i) \\ & + G_H\hat{r}_H(n) + \sum_{i=1}^{N_{p_H}} \beta_{H,i}\hat{d}_H(n - M_H - i). \end{aligned} \tag{5}$$

The pitch lags M_L and M_H must both be larger than the pitch sub-frame size. This prevents any feedback which causes the equations to become non-linear Then, $\hat{d}(n)$ can be viewed as a linear combination of all the known waveforms $\hat{r}_L(n)$, $\hat{r}_H(n)$, $\hat{d}_L(n - M_L - i)$ and $\hat{d}_L(n - M_H - i)$ The bandwidth expanded regenerated speech $\hat{s}'(n)$ can then be expressed as.

$$\begin{aligned} \hat{s}'(n) = {} & \sum_{k=-\infty}^{-1} \hat{d}_L(k)h'(n-k) + \sum_{k=-\infty}^{-1} \hat{d}_H(k)h'(n-k) \\ & + \sum_{k=0}^{\infty} \hat{d}_L(k)h'(n-k) + \sum_{k=0}^{\infty} \hat{d}_H(k)h'(n-k), \end{aligned} \tag{6}$$

The anti-causal terms in the above equation are the zero-input responses of $H'(z)$, and account for the initial conditions of each band at the pitch sub-frame boundaries. The impulse response $h'(n)$ is causal, and the upper limit in both causal terms summations can be set to $N-1$ Defining the following terms,

$$\begin{aligned} x_L(n) &= \sum_{k=0}^{N-1} \hat{r}_L(k)h'(n-k), \\ x_H(n) &= \sum_{k=0}^{N-1} \hat{r}_H(k)h'(n-k), \end{aligned} \tag{7}$$

and

$$\begin{aligned} y_{L,i}(n) &= \sum_{k=0}^{N-1} \hat{d}_L(k - M_L - i)h'(n-k), \\ y_{H,i}(n) &= \sum_{k=0}^{N-1} \hat{d}_H(k - M_H - i)h'(n-k), \end{aligned} \tag{8}$$

the weighted error $e_w(n)$ can be expressed as:

$$\begin{aligned} e_w(n) = {} & s^*(n) - G_L x_L(n) - \sum_{i=1}^{N_{p_L}} \beta_{L,i} y_{L,i}(n), \\ & - G_H x_H(n) - \sum_{i=1}^{N_{p_H}} \beta_{H,i} y_{H,i}(n), \end{aligned} \tag{9}$$

where $s^*(n)$ contains all the terms not subjected to optimization·

$$s^*(n) = s'(n) - \sum_{k=-\infty}^{-1} \hat{d}(k)h'(n-k) \tag{10}$$

The optimization is done in the mean-square sense. Let the energy of the weighted error signal $e_w(n)$ in the pitch sub-frame be:

$$\xi = \sum_{n=0}^{N-1} e_w(n)^2. \tag{11}$$

Differentiating the above equation with respect to the gain and the pitch coefficients and setting it equal to 0 yields, for any given pitch lag values, a linear system of equations. This is best represented in matrix form, $\mathbf{\Phi v} = \mathbf{b}$, where $\mathbf{\Phi}$, $\mathbf{v}$ and $\mathbf{b}$ are as follows[†]:

$$\mathbf{\Phi} = \langle \mathbf{q}\mathbf{q}^{\mathrm{T}} \rangle, \tag{12}$$

where

$$\mathbf{q} = \begin{pmatrix} x_L(n) \\ x_H(n) \\ y_{L,1}(n) \\ \vdots \\ y_{L,N_{p_L}}(n) \\ y_{H,1}(n) \\ \vdots \\ y_{H,N_{p_H}}(n) \end{pmatrix}, \tag{13}$$

$$\mathbf{v} = \begin{pmatrix} G_L \\ G_H \\ \beta_{L,1} \\ \vdots \\ \beta_{L,N_{p_L}} \\ \beta_{H,1} \\ \vdots \\ \beta_{H,N_{p_H}} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \langle s^*(n)x_L(n) \rangle \\ \langle s^*(n)x_H(n) \rangle \\ \langle s^*(n)y_{L,1}(n) \rangle \\ \vdots \\ \langle s^*(n)y_{L,N_{p_L}}(n) \rangle \\ \langle s^*(n)y_{H,1}(n) \rangle \\ \vdots \\ \langle s^*(n)y_{H,N_{p_H}}(n) \rangle \end{pmatrix}. \tag{14}$$

[†] For clarity, all summation symbols are left out. Thus $\langle x(n) \rangle$ refers to $\sum_{n=0}^{N-1} x(n)$

The matrix Φ is symmetric and can be solved using the Cholesky factorization technique. Since the solution vector $\mathbf{v}$ depends on the pitch lags (from Eq. 8), the overall optimal solution is obtained through an exhaustive search of all possible lag values. Then, for each codebook entry, the linear system in Eq. 12 is solved. The index of the codewords yielding the lowest error is then transmitted to the receiver.

4. Parameter design and selection

This section briefly discusses each coding parameter. In most cases, the parameter configuration and selection are based of simulation results. The wideband CELP coders are simulated in floating-point on a general purpose workstation. The wideband speech signals are sampled at 16 kHz and are bandlimited to 7500 Hz. Also, where applicable, the parameters for the full-band and split-band structures are dealt with separately.

4.1 Frame and sub-frame sizes

The frame and sub-frame sizes control the update rate of all the coding parameters and are set to 320 and 40 samples (50 and 400 Hz) respectively. These update rates correspond to typical frame and sub-frame durations (20 ms and 5 ms) found in narrowband CELP coders. Faster update rates directly improve the quality of the coded speech.

4.2 LPC coefficients coding

The LPC coefficients a_k are coded using Line Spectral Frequencies (LSF's) [7]. These are a transformation of the direct form coefficients a_k. Moreover, LSF's are always ordered for stable synthesis filters and thus, stability can easily be ensured after quantization. For wideband speech, 16 coefficients are used to model the spectral envelope, and a non-uniform differential scalar quantization scheme is used [8]. Since the LSF's are related to the formants positions, allocating more bits for the lower LSF's emphasizes the perceptually important lower frequencies. During the simulations, 50 to 60 bits/frame are used. This figure could be reduced through inter and intra-frame interpolation [1].

4.3 Pitch coefficients coding

The computed optimal pitch coefficients are coded with non-uniform scalar quantizers. Quantization is done before the error energy ξ (Eq. 11) is calculated. The quantization error is thus accounted for within the optimization. For the full-band structure, 3 pitch taps are used, while for the split-band approach, there is 1 tap in each band. In both cases, a higher number of pitch taps increases the perceived quality of the coded speech. Also, since the sampling rate is 16 kHz, the pitch parameter resolution is finer than that found in narrowband systems and has an effect similar to fractional pitch determination [9].

4.4 Lag estimate and coding

The optimal solution for Eq. 12 is computationally heavy. Indeed, the system of equations must be solved for all lag values within the predefined range and all codebook waveforms. A slightly less optimal, yet more efficient approach, is to solve for the optimal lag values with the gains set to zero [1], thus eliminating any contribution from the current excitations. In essence, this amounts to letting the pitch synthesis filters *free-wheel* (or self-excite) with past regenerated formant residuals. This eliminates the computational burden induced by nesting exhaustive lag and waveform index searches. The loss in performance is small [1] since the contributions to the pitch structure primarily come from the past regenerated formant residuals and not from the current excitations.

In the split-band structure, a single lag value is used for both bands. Simulation results show that this does not reduce the perceived quality of the regenerated speech.

4.5 Codeword design and selection

For the full-band structure, the codebook consists of normalized *iid* Gaussian sequences. The optimal codeword is selected by solving the linear system of Eq. 12 for each codeword entry, and keeping the index of the codeword that yields the smallest error energy ξ. The codeword length is always the same as the sub-frame size. The quality of the reproduced speech improves with the size of the codebook and the number of codeword is set to 1024.

For the split-band structure, separate excitations are required for each band and thus, a low and a high-band codebook are used. The codebooks can either be normalized *iid* Gaussian sequences (as in the full-band case), or band-limited normalized Gaussian sequences. Band-limiting the codebooks is done at design time by filtering a Gaussian sequence with a low or high-pass filter. Experimental results show that the best configuration consists in a full-pass codebook for the low-band combined with a sharp cutoff (4 kHz) high-pass codebook for the high-band. This prevents the high-band excitation from contributing to the low-band regenerated speech and provides the best harmonic match to the original speech.

An optimal codewords selection method for the split-band structure can be computationally intensive due to the nested searches. A less optimal approach is to let both codebooks have the same size. A single index chooses both the low and the high-band codeword. Since most of the error energy ξ comes from the low-band contribution and since the optimization is done by minimizing ξ, in most cases the effects of the low-band codebook predominate. Yet, this codewords selection method remains flexible enough to accomodate cases where the high frequency contents of the signal significantly contributes to the error energy. Experimental results show that this approach induces little degradation in the reconstructed speech and that there is no perceived difference between codebooks of size 512 and 1024. The success of this method shows that the high frequencies found in a wideband signal need not be coded precisely.

4.6 Gain estimate and coding

For both strucures, a differential quantizer with a leaky predictor (1 tap $\alpha = 0.9$) is used to code the difference in successive sub-frame gain magnitudes. An extra bit codes the sign. The computed gains are quantized before calculating the error energy ξ. This ensures the overall best solution under quantization constraints.

In split-band mode, distinct gains G_L and G_H are computed and coded. Experimental results show that separate gains, as opposed to a common gain for both bands, help reduce high frequency hiss and improve the Segmental SNR of the regenerated speech signal.

5. Comparison of full-band and split-band wideband CELP

Based on the simulation results, the best full and split-band wideband CELP coders are now compared while subjected to a maximum operating rate of 16 kbits/sec. During the simulations, the emphasis has been put on studying the model structures rather than developing elaborate parameter quantization methods. To this effect, the comparison is done with no quantization other than that introduced by the codeword selection. The operating rate calculations use estimated bit requirements for each parameter based on existing narrowband CELP implementations, except for the LPC coefficients, which use estimates based on LSF coding experimentations done on wideband signals. Two coder implementations are considered and listed in Tables 1 and 2.

Parameter	Bits	Update rate (Hz)	Bits/sec
LPC coefficients	48	50	2400
β_1	5	400	2000
β_2	3	400	1200
β_3	3	400	1200
gain G	6	400	2400
lag M	7	400	2800
codebook index	10	400	4000
		Total	16000

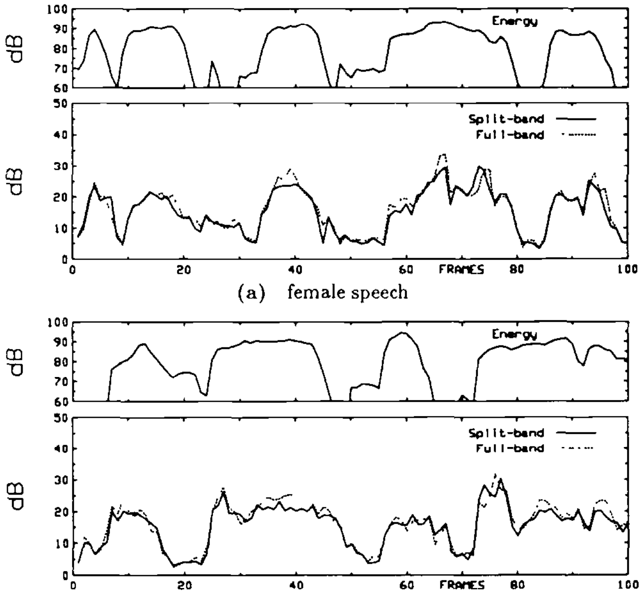
Table 1 Full-band coder configuration

Both coders yield high quality reconstructed speech (unquantized parameters). In terms of SegSNR, the full-band implementation is about 0.5 dB higher than the split-band approach. The SegSNR tracks of Figure 3 show little overall difference between the two methods.

Perceptually, there are some cases where the full-band implementation suffers from a slight hollowness and from a certain hiss around fricatives. The split-band implementation does not exhibit these problems and generally produces a richer sound than the full-band method.

Parameter	Bits	Update rate (Hz)	Bits/sec
LPC coefficients	48	50	2400
β_L	5	400	2000
β_H	3	400	1200
gain G_L	6	400	2400
gain G_H	4	400	1600
lag M	7	400	2800
codebook index	9	400	3600
		Total	16000

Table 2 Split-band coder configuration.



(a) female speech

(b) male speech

Fig. 3 Full versus split-band SegSNR

In particular, it does a better job of reproducing the baseband This, in turn, seems to be an essential condition for the overall good reproduction of a wideband signal. The split-band CELP structure proposed here follows this condition and emphasizes the low-band. The cost of the high-band is computed as follows. 1600 bits/sec for the extra gain factor G_H, 1200 bits/sec for the extra pitch tap β_H, 720 bits/sec for the LPC coefficients (assuming that 30% of the LPC bits are modeling the high-band), and finally 1400 bits/sec for the shared lag value This adds up to 4920 bits/sec, or roughly 30% of the overall operating rate.

5.1 Comparison with a 16 kbits/sec narrowband coder

Finally, both coders were compared to a low-delay CELP narrowband coder operating near toll-quality at 16 kbits/sec [10] For this comparison, the original wideband speech files were low-pass filtered at 3300 Hz, downsampled at 8 kHz and then processed by the narrowband coder. Informal tests were conducted with many different listeners to determine which of the two types of coders (narrowband *vs* wideband) was preferred. The wideband coders were always preferred over the narrowband one.

This test clearly demonstrates that the reproduced wideband speech is judged to be of better quality. The extra bandwidth yields a "fuller" sound, and also greatly enhances the perception of fricative sounds. The small degradations found when carefully listening to the coded wideband signals through headphones are not noticeable in an open environment such as a conference room. Even though the wideband CELP coders were not operating under full parameter quantization, these results nevertheless indicate that for a potential operating bit rate of 16 kbits/sec, the wideband CELP coders can yield a clearer, richer sound than their narrowband counterparts.

6. Conclusion

The feasibility of a wideband CELP speech coder operating at 16 kbits/sec has been demonstrated. To this effect, the basic CELP model has been extended to a more general split-band CELP model. This provides flexible control over the parameters found in each band. The split-band CELP coder yields a cleaner, richer sound than the full-band CELP coder. Simulations also helped determined that although they greatly improve the perceived quality of a coded speech signal, the high frequencies found in a wideband signal need not be coded precisely Finally, for the same operating rate of 16 kbits/sec, subjective tests showed that the wideband speech coder was preferred to a high-quality narrowband coder This wideband speech coder offers an attractive alternative to conventional narrowband coders at rates near 16 kb/s for many applications.

References

1 P. Kabal, J L Moncet and C. C. Chu, "Synthesis filter optimization and coding applications to CELP," *Proc IEEE Int. Conf. on Acoust., Speech and Signal Processing,* Section S4 2, pp 147–150, New York, April 1988

2 D O'Shaughnessy, *Speech Communication, Human and Machine,* Addison-Wesley, 1987

3. B. Atal and M R. Schroeder, "Code-excited linear prediction (CELP): high quality speech at very low bit rates," *Proc IEEE Int. Conf. on Acoust., Speech and Signal Processing,* pp 937–940, San Diego, March 1984.

4. P. Kabal, "Code excited linear prediction coding of speech at 4 8 Kbits/s," *Rapport technique de l'INRS-Télécommunications,* No. 87-36, July 1987.

5 B S. Atal and M. R. Shroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. ASSP-27, No. 3, June 1979, pp. 247–254.

6 R. Drogo, R. Montagna, F. Perosino and D. Sereno, "Some experiments of 7 kHz audio coding at 16 kbits/s," *Proc IEEE Int. Conf. on Acoust., Speech and Signal Processing,* pp. 192–195, Glascow, May 1989.

7. F. K. Soong and B. H. Juang, "Line spectrum pair (LSP) and speech data compression," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,* pp. 1.10.1–1.10.4, San Diego, March 1984.

8. N. Sugamura and N. Favardin, "Quantizer design in LSP speech analysis–synthesis," *IEEE Journal on Selected Areas in Communication,* Vol. 6, No. 2, pp.432–440, February 1988.

9. P. Kroon and B. S. Atal, "Pitch predictors with high temporal resolution," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,* pp. 661–664, Albuquerque, April 1990.

10 V. Iyengar and P. Kabal, "A low-delay 16 kbits/sec speech coder," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing,* pp. 243–246, New York, April 1988.