CODE-EXCITED LINEAR PREDICTION (CELP): HIGH-QUALITY SPEECH AT VERY LOW BIT RATES

Manfred R. Schroeder
Drittes Physikalisches Institut
University of Goettingen, F. R. Germany
and AT&T Bell Laboratories
Murray Hill, New Jersey 07974

Bishnu S. Atal
AT&T Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

We describe in this paper a code-excited linear predictive coder in which the optimum innovation sequence is selected from a code book of stored sequences to optimize a given fidelity criterion. Each sample of the innovation sequence is filtered sequentially through two time-varying linear recursive filters, one with a long-delay (related to pitch period) predictor in the feedback loop and the other with a short-delay predictor (related to spectral envelope) in the feedback loop. We code speech, sampled at 8 kHz, in blocks of 5-msec duration. Each block consisting of 40 samples is produced from one of 1024 possible innovation sequences. The bit rate for the innovation sequence is thus 1/4 bit per sample. We compare in this paper several different random and deterministic code books for their effectiveness in providing the optimum innovation sequence in each block. Our results indicate that a random code book has a slight speech quality advantage at low bit rates. Examples of speech produced by the above method will be played at the conference.

INTRODUCTION

Performance of adaptive predictive coders for speech signals using instantaneous quantizers deteriorate rapidly at bit rates below about 10 kbits/sec. Our past work has shown that high speech quality can be maintained in predictive coders at lower bit rates by using non-instantaneous stochastic quantizers which minimize a subjective error criterion based on properties of human auditory perception. [1]. We have used tree search procedures to encode the innovation signal and have found the tree codes to perform very well at 1 bit/sample (8 kbits/sec.). The speech quality is maintained even at 1/2 bit/sample when the tree has 4 branches at every node and 4 white Gaussian random numbers on each branch [2].

The tree search procedures are suboptimal and the performance of tree codes deteriorates significantly when the innovation signal is coded at only 1/4 bit/sample (2 kbits/sec). Such low bit rates for the innovation signal are necessary to bring the total bit rate for coding the speech signal down to 4.8 kbits/sec - a rate that offers the possibility of carrying digital speech over a single analog voice channel.

Fehn and Noll [3] have discussed merits of various multipath search coding procedures: code-book coding, tree coding, and trellis coding. Code-book coding is of particular interest at very low bit rates. In code-book coding, the set of possible sequences for a block of innovation signal is stored in a code book. For a given speech segment, the optimum innovation sequence is selected to optimize a given fidelity criterion by exhaustive search of the code book and an index specifying the optimum sequence is transmitted to the receiver. In general, code-book coding is impractical due to the large size of the code books. However, at the very low bit rates we are aiming for, exhaustive search of the code book to find the best innovation sequence for encoding short segments of the speech signal becomes possible [4].

SPEECH SYNTHESIS MODEL

The speech synthesizer in a code-excited linear predictive coder is identical to the one used in adaptive predictive coders [1]. It consists of two time-varying linear recursive filters each with a predictor in its feedback loop as shown in Fig. 1. The first feedback loop includes a long-delay (pitch) predictor which generates the pitch periodicity of voiced speech. The second feedback loop includes a short-delay predictor to restore the spectral envelope.

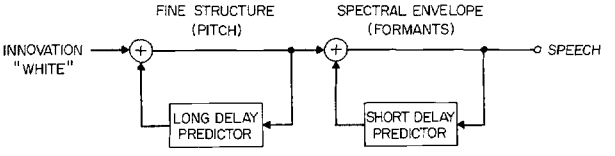

Fig. 1. Speech synthesis model with short and long delay predictors.

The two predictors are determined using procedures outlined in References 1 and 5. The short-delay predictor has 16 coefficients and these are determined using the weighted stabilized covariance method of LPC analysis [1,5] once every 10 msec. In this method of LPC analysis, the instantaneous prediction error is weighted by a Hamming window 20 msec in duration and the predictor coefficients are determined by minimizing the energy of the weighted error. The long-delay (pitch) predictor has 3 coefficients which are determined by minimizing the mean-squared prediction error after pitch prediction over a time interval of 5 msec [2].

SELECTION OF OPTIMUM INNOVATION SEQUENCE

Let us consider the coding of a short block of speech signal 5 msec in duration. Each such block consists of 40 speech samples at a sampling frequency of 8 kHz. A bit rate of 1/4 bit per sample corresponds to 1024 possible sequences (10 bits) of length 40 for each block. The procedure for selecting the optimum sequence is illustrated in Fig. 2. Each member of the code book provides 40 samples of the innovation signal. Each sample of the innovation signal is scaled by an amplitude factor that is constant for the 5 msec block and is reset to a new value once every 5 msec. The scaled samples are filtered sequentially through two recursive filters, one for introducing the voice periodicity and the other for the spectral envelope. The regenerated speech samples at the output of the second filter are compared with the corresponding samples of the original speech signal to form a difference signal. The difference signal representing the objective error is further processed through a linear filter to attenuate those frequencies where the error is perceptually less important and to amplify those frequencies where the error is perceptually more important. The transfer function of the weighting filter is given by

CH2118-8/85/0000-0937 $1.00 © 1985 IEEE

$$W(z) = \frac{1 - \sum_{k=1}^{p} a_k z^{-k}}{1 - \sum_{k=1}^{p} a_k \alpha^k z^{-k}}, \quad (1)$$

where a_k are the short-delay predictor coefficients, $p=16$ and α is a parameter for controlling the weighting of the error as a function of frequency. A suitable value of α is given by

$$\alpha = e^{-2\pi.100/f_s} \quad (2)$$

where f_s is the sampling frequency. The weighted mean-squared error is determined by squaring and averaging the error samples at the output of the weighting filter for each 5-msec block. The optimum innovation sequence for each block is selected by exhaustive search to minimize the weighted error. As mentioned earlier, prior to filtering, each sample of the innovation sequence is scaled by an amplitude factor that is constant for the 5-msec block. This amplitude factor is determined for each code word by minimizing the weighted mean-squared error for the block.

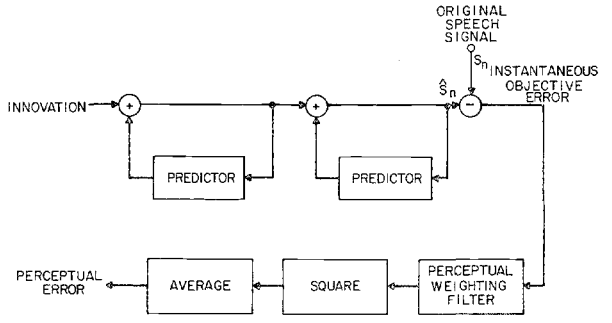

Fig. 2. Block diagram illustrating the procedure for selecting the optimum innovation sequence.

CONSTRUCTION OF OPTIMUM CODE BOOKS

A code book, within the limitation of its size, should provide as dense a sampling as possible of the space of innovation sequences. In principle, the code words could be block codes that are optimally placed on a hypersphere in the 40-dimensional space (representing 40 samples in each 5-msec block). Fehn and Noll [3] have argued that random code books (code books with randomly selected code words) are less restrictive than deterministic code books Random code books, in some sense, provide a lower bound for the performance at any given bit rate. A deterministic code book, if properly constructed, should provide a performance that is at least equal to - if not better than - that of the random code books and the deterministic nature of the code book should make it easier to find the optimum innovation sequence for each block of speech. However, it is generally very difficult to design an optimum deterministic code book.

As a start, we have chosen a random code book in which each possible code word is constructed of white Gaussian random numbers with unit variance. We have chosen the Gaussian distribution since our earlier work has shown that the probability density function of the prediction error samples (after both short-delay and long-delay predictions) is nearly Gaussian [1]. Figure 3 shows a plot of the first-order cumulative amplitude distribution function for the prediction residual samples and compares it with the corresponding Gaussian distribution function with the same mean and variance. A closer examination of the prediction error shows that the Gaussian assumption is valid almost everywhere except for stop bursts of unvoiced stop consonants and for a few pitch periods during the transition from unvoiced or silence regions to voiced speech.

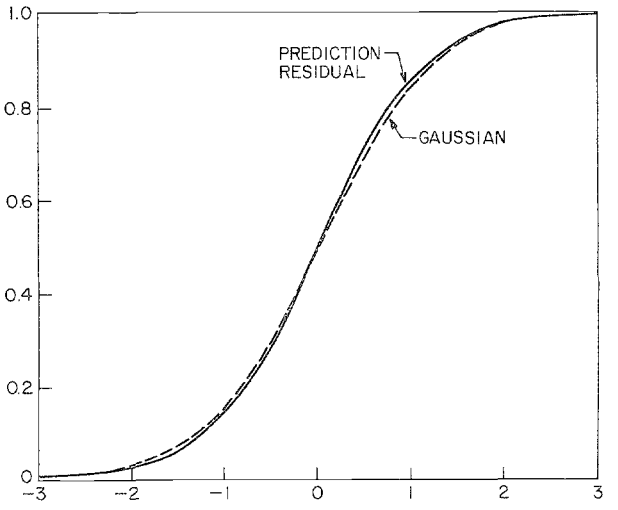

Fig. 3. First-order cumulative probability distribution function for the prediction residual samples (solid curve). The corresponding Gaussian distribution function with the same mean and variance is shown by the dashed curve.

Each sample v_n of the innovation sequence in a Gaussian code book can be expressed as a Fourier series of N cosine functions ($N=20$):

$$v_n = \sum_{k=0}^{N-1} c_k \cos(\pi kn/N + \phi_k), \; n=0,1,...,2N-1, \quad (3)$$

where c_k and ϕ_k are independent random variables, ϕ_k is uniformly distributed between 0 and 2π, and c_k is Rayleigh distributed with probability density function

$$p(c_k) = c_k \exp(-c_k^2/2), \; c_k > 0. \quad (4)$$

The function of the innovation sequence in the synthesis model of Fig. 1 is to provide a correction to the filter output in reproducing the speech waveform within the limitation of the size of the code book. Using the Fourier series model of Eq. (3), the correction can be considered separately for the amplitude and phase of each Fourier component. Do we need both amplitude and phase corrections for high-quality speech synthesis? Are the two types of corrections equally important? These questions can be answered by restricting the variations in the amplitudes and phases of various Fourier components in Eq. (3). For example, a code book can be formed by setting the amplitudes c_k to a constant value and by keeping the phases ϕ_k uniformly distributed between 0 and 2π. Another code book is formed by setting the phases to some constant set of values and by keeping the amplitudes Rayleigh distributed in accordance with Eq. (4).

We have also used a code book in which the different innovation sequences are obtained directly from the prediction error (after normalizing to unit variance) of speech signals. The amplitudes and phases are no longer distributed according to Rayleigh and uniform density functions, respectively, but reflect the distributions represented in the actual prediction error.

RESULTS

As we mentioned earlier, the random code book provides a base line against which we can compare other code books. We have synthesized several speech utterances spoken by both male and female

speakers (pitch frequencies ranging from 80 Hz to 400 Hz) using the different code books discussed in the previous section. The random code book (with 1024 code words) provided unexpectedly good performance. Even in close pair-wise comparisons over head phones, only occasional small differences were noticeable between the original and synthetic speech utterances. These results suggest that a 10-bit random code book has sufficient flexibility to produce high-quality speech from the synthesis model shown in Fig. 1.

The waveforms of the original and synthetic speech signals were found to match closely for voiced speech and reasonably well for unvoiced speech. The signal-to-noise ratio averaged over several seconds of speech was found to be approximately 15 dB. Examples of speech waveforms are shown in Fig. 4. The figure shows (a) original speech, (b) synthetic speech, (c) the LPC prediction residual, (d) the reconstructed LPC residual, (e) the prediction residual after pitch prediction, and (f) the coded residual from a 10-bit random code book. As expected, the Gaussian code book is not able to reproduce a sharp impulse in the coded residual waveform. The absence of the sharp impulse produces appreciable phase distortion in the reconstructed LPC prediction residual. However, this phase distortion is mostly limited to frequency regions outside the formants.

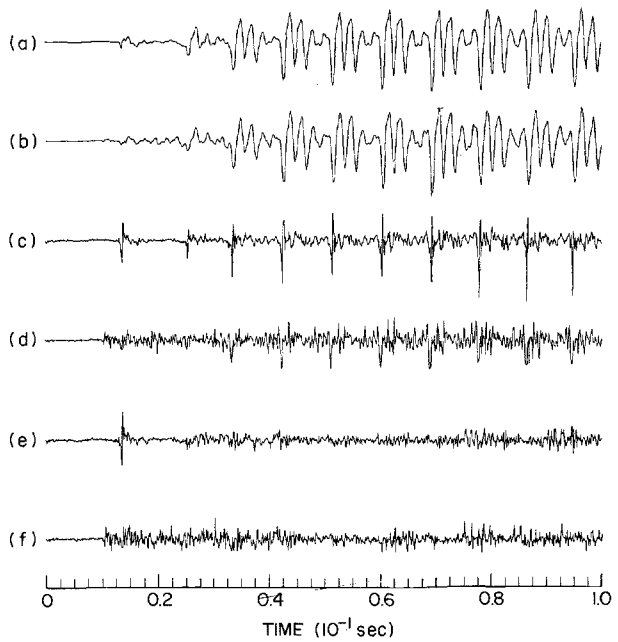

Fig. 4. Waveforms of different signals in the coder: (a) the original speech, (b) the synthetic speech, (c) the LPC prediction residual, (d) the reconstructed LPC residual, (e) the prediction residual after pitch prediction, and (f) the coded residual from a 10-bit random code book. Waveforms (c) and (d) are amplified 5 times relative to the speech signal. Waveforms (e) and (f) are amplified by an additional factor of 2.

We have also examined the distribution of the reconstruction error amongst various code words. Figure 5 shows a plot of the number of code words which produced a given amount of rms error in a particular 5-msec block of speech. The behavior shown is typical of what we observed in several blocks. The minimum rms error for this block was 30 and only 5 code words (out of a total of 1024) produced an rms error less than 33. This indicates that the size of the code book cannot be reduced significantly without producing substantial increase in the error.

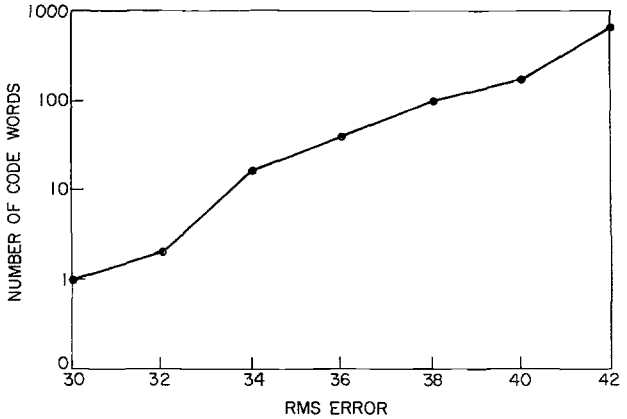

Fig. 5. Distribution of error amongst the various code words in a Gaussian code book.

Due to the random nature of code books, different Gaussian code books produced different innovation sequences. However, we did not hear any audible difference between the speech signals reconstructed from these different code books. Figure 6 shows several examples of the innovation sequences selected from several different Gaussian code books for one 5-msec block. The innovation sequences for other previous blocks were kept the same; thus, the filter coefficients and the filter memories were identical at the beginning of the block. The coded innovation sequences show very little similarity to each other. The amplitude spectrum for the different sequences is shown in Fig. 6(b). Again, there is no obvious common pattern amongst the different amplitude spectra. The corresponding phase responses are shown Fig. 6(c).

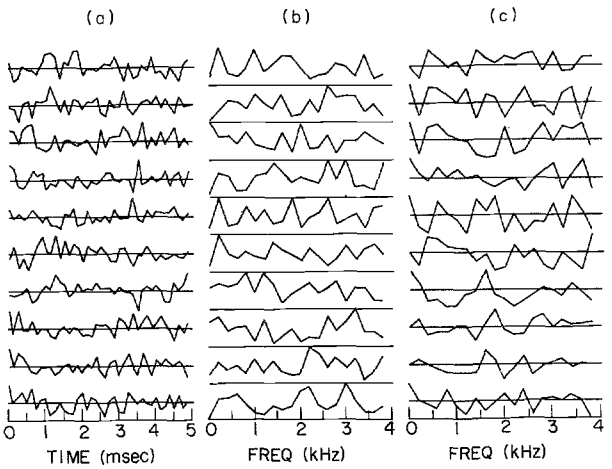

Fig. 6. (a) Waveforms of different innovation sequences for a particular 5-msec block, (b) amplitude spectra of innovation sequences, and (c) phase responses of innovation sequences.

The code book with constant amplitude but uniformly distributed phases performed nearly as well as the Gaussian code book. The signal-to-noise ratio decreased by about 1.5 dB and there was an audible difference between the two code books. The code book with

constant phases but Rayleigh-distributed amplitudes performed very poorly, both in the signal-to-noise ratio and in listening to synthetic speech. The code book based on the prediction residual signals derived from speech performed as well as the Gaussian code book.

CONCLUDING REMARKS

Our present work with the code-excited linear predictive coder has demonstrated that such coders offer considerable promise for producing high quality synthetic speech at bit rates as low as 4.8 kbits/sec. The random code book we have used so far obviously does not provide the best choice. The proper design of the code book is the key to success for achieving even lower bit rates than we realized in this study. We have so far employed a fixed code book for all speech data. A fixed code book is somewhat wasteful. Further efficiency could be gained by making the code book adaptive to the time-varying linear filters used to synthesize speech and to weight the error. The coding procedure is computationally very expensive; it took 125 sec of Cray-1 CPU time to process 1 sec of the speech signal. The program was however not optimized to run on Cray. Most of the time was taken up by the search for the optimum innovation sequence. A code book with sufficient structure amenable to fast search algorithms would lead to real time implementation of code-excited coders.

REFERENCES

[1] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Commun.* vol. COM-30, pp. 600-614, April 1982.

[2] M. R. Schroeder and B. S. Atal, "Speech coding using efficient block codes," in *Proc. Int. Conf. on Acoustics, Speech and Signal Proc.*, vol. 3, pp. 1668-1671, May 1982.

[3] H. G. Fehn and P. Noll, "Multipath search coding of stationary signals with applications to speech," *IEEE Trans. Commun.* vol. COM-30, pp. 687-701, April 1982.

[4] B. S. Atal and M. R. Schroeder, "Stochastic coding of speech signals at very low bit rates," in *Proc. Int. Conf. Commun. - ICC84*, part 2, pp. 1610-1613, May 1984.

[5] S. Singhal and B. S. Atal, "Improving performance of multi-pulse LPC coders at low bit rates," in *Proc. Int. Conf. on Acoustics, Speech and Signal Proc.*, vol. 1, paper no. 1.3, March 1984.