IMPROVING PERFORMANCE OF MULTI-PULSE LPC CODERS AT LOW BIT RATES

Sharad Singhal and Bishnu S. Atal

Acoustics Research Department
AT&T Bell Laboratories, Murray Hill, NJ 07974

ABSTRACT

The multi-pulse excitation model provides a method for producing natural-sounding speech at medium to low bit rates. Multi-pulse analysis obtains the all-pole filter excitation by minimizing a spectrally-weighted mean-squared error between the original and synthetic speech signals. Although the method provides high quality speech around 10 kbits/sec, speech quality suffers if the bit rate is lowered. In this paper, we focus on problems encountered in attempting to maintain speech quality while synthesizing speech using multi-pulse excitation at lower bit rates.

INTRODUCTION

The multi-pulse model of LPC excitation [1] provides a method for producing natural sounding speech. This model replaces the traditional pitch pulse and white noise excitation used in vocoders by a sequence of pulses. The optimum pulse sequence is chosen to minimize a spectrally-weighted mean-squared error between the original and synthetic speech signals. The optimization is carried out sequentially - one pulse at a time - over short time segments typically 5 to 10 msec long during which the speech signal can be considered quasi-stationary.

Although the multi-pulse excitation is capable of producing high quality synthetic speech at rates around 10 kbits/sec, the speech quality deteriorates rapidly at much lower bit rates. There are several reasons for this loss of speech quality at lower bit rates.

First, present LPC analysis procedures are not perfect. Even when the all-pole model provides a reasonable approximation to the speech spectrum, these procedures introduce errors in estimating the filter parameters. At higher bit rates, these errors can be compensated by using additional pulses in the multi-pulse excitation. This however is not possible at lower bit rates, where the number of pulses have to be kept at a minimum. We thus need a more accurate LPC analysis procedure to achieve a significant reduction in the bit rate for the multi-pulse excitation. An important source of error in present LPC analysis procedures is introduced by the rectangular window used for computing the mean-squared prediction error. The rectangular window causes the resultant LPC parameters to become sensitive to the position of the analysis frame in the utterance [2]. In this paper, we describe an improved LPC analysis procedure to avoid this problem.

Second, in the multi-pulse analysis method [1], the amplitudes and locations of the pulses are determined in successive stages - one pulse at a time. For closely spaced pulses, successive optimization of individual pulses is inaccurate and often requires additional pulses to compensate for inaccuracies introduced earlier. Again, these additional pulses result in a higher bit rate than is necessary. This problem is avoided by using a procedure in which the amplitudes of the pulses are kept optimum at each stage of multi-pulse analysis.

Third, it has been our experience that the multi-pulse excitation needs approximately 8 pulses per pitch period to synthesize high quality speech. Thus, for proper synthesis of speech with an average pitch of 10 msec, an average pulse rate of 800 per second for the excitation is necessary. However, this also means that a much higher pulse rate is needed for speech where the pitch period is shorter. The number of pulses in a pitch period can be reduced significantly by exploiting the periodic nature of voiced speech. We describe in this paper a method to achieve this reduction by using pitch prediction on the multi-pulse excitation signal.

IMPROVED LPC ANALYSIS PROCEDURE

Linear predictor coefficients are determined by minimizing the mean-squared prediction error between the current value of the speech signal and its predicted value based on the past samples. The prediction error $\epsilon(n)$ is defined as

$$\epsilon(n) = s(n) - s'(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k), \qquad (1)$$

where $s(n)$ is the speech sample at the nth sampling instant, $s'(n)$ is the predicted value of the nth speech sample, a_k is the kth predictor coefficient, and p is the number of predictor coefficients. For stationary signals, whose underlying spectral characteristics do not change with time, the mean-squared prediction error is given by

$$E = \lim_{N\to\infty} 1/N \sum_{n=1}^{N} \epsilon^2(n). \qquad (2)$$

For quasi-stationary signals, the sum over n in Eq. (2) is usually carried out over a time interval during which the spectral characteristics of the signal are approximately constant. The prediction error as defined in Eq. (2) shows considerable variation even when the position of the analysis frame is changed slightly. The predictor coefficients are determined by minimizing E and consequently show similar variations. Figure 1 shows a portion of the utterance "Joe brought a young girl" spoken by a male speaker, and a plot of the first 4 partial correlation coefficients as a function of the position of the analysis frame. LPC analysis was performed using the stabilized covariance method [3] with a 20 msec frame. Each data point in the plot represents a new set of LPC parameters obtained by shifting the analysis frame by one sampling interval (125 μsec at the sampling frequency of 8 kHz).

These sample-to-sample fluctuations are not usually obvious when the LPC parameters are computed every 10 to 20 msec; in this case, the high-frequency fluctuations are folded back in the low frequency range. Let us consider this point in more detail. The prediction error defined in Eq. (1) has the same bandwidth as the speech signal. The squared prediction error has a bandwidth which is twice the bandwidth of the speech signal. The rectangular window used in averaging the squared prediction error as shown in Eq. (2) has sharp edges and thus a poor side-lobe response at high frequencies [4]. The sharp edges of the rectangular window cause the mean-squared prediction error E to vary rapidly as a function of the position of the analysis frame. Thus the rectangular window is not appropriate for computing the average squared prediction error.

CH1945-5/84/0000-0003 $1.00 © 1984 IEEE

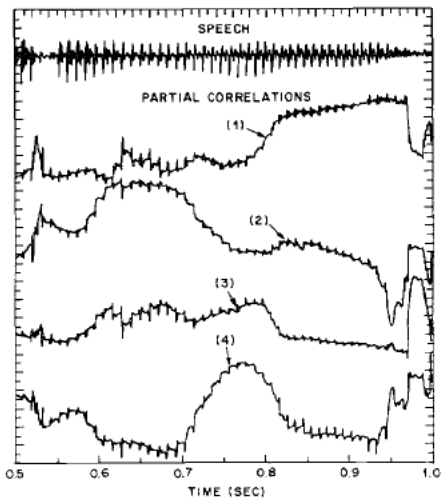

Fig. 1. The speech waveform and the first 4 partial correlation coefficients obtained with traditional LPC analysis.

This problem can be avoided by using a non-rectangular window, such as a Hamming window [4], which has lower side lobes at high frequencies. We now define a more general form for the mean-squared prediction error as

$$E(n) = \sum_{i=-m}^{m} \epsilon^2(n-i)w(i), \tag{3}$$

where $E(n)$ is the mean squared-prediction error at the nth sampling instant and $w(i)$ is the ith sample of the Hamming window (or any other appropriately tapered window) with a duration of $2m+1$ samples. Equation (3) can be rewritten as

$$E(n) = \sum_{i=-m}^{m} [s(n-i) - \sum_{k=1}^{p} a_k s(n-i-k)]^2 w(i). \tag{4}$$

The predictor coefficients at the nth sampling instant are determined by minimizing the prediction error $E(n)$. We then obtain a set of linear equations

$$\sum_{k=1}^{p} a_k \phi_{kr} = \psi_r, \quad 1 \leqslant r \leqslant p, \tag{5}$$

where

$$\phi_{kr} = \sum_{i=-m}^{m} w(i)s(n-i-k)s(n-i-r), \quad 1 \leqslant k \leqslant p, \; 1 \leqslant r \leqslant p,$$

and

$$\psi_r = \sum_{i=-m}^{m} w(i)s(n-i)s(n-i-r). \tag{6}$$

These equations are solved to provide a new set of predictor coefficients. The extension of the above approach to determine partial correlations (or reflection coefficients) using the stablized covariance method is straightforward. The covariance matrix $[\phi_{kr}]$ is first expressed as the product of a lower triangular matrix $L = [L_{kr}]$ and its transpose L^t. Next, a set of linear equations

$$\sum_{k=1}^{p} L_{kr} q_k = \psi_r, \quad 1 \leqslant r \leqslant p, \tag{7}$$

is solved. The partial correlation at delay m is given by

$$r_m = q_m / (\psi_0 - \sum_{i=1}^{m-1} q_i^2)^{1/2} \tag{8}$$

where ψ_0 is the weighted energy of the speech signal as defined in Eq. (6). It is also desirable to apply a correction for the high frequency roll-off of the low pass filter used in analog-to-digital

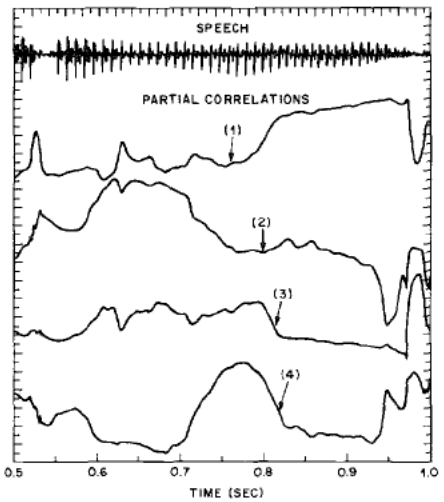

Fig. 2. The partial correlation coefficients obtained with the improved LPC analysis procedure for the same speech segment as in Fig. 1.

conversion of the speech signals [3]. Figure 2 shows a plot of the first 4 partial correlation coefficients based on the above procedure using a 20 msec. Hamming window for the same speech segment as in Fig. 1. Even though the effective duration of the Hamming window is only half of the rectangular window used in the traditional LPC analysis, the sample-to-sample fluctuations are reduced significantly with the new procedure.

OPTIMAL AMPLITUDE EXCITATION

Multi-pulse excitation is obtained by minimizing a weighted mean-squared error between the original and synthetic speech. The LPC synthesizer produces synthetic speech, which is compared to the original speech to produce the error signal e_n. The error signal is weighted to de-emphasize the error in the formant regions by using a linear filter of the form

$$W(z) = [1 - P(z^{-1})]/[1 - P(\mu z^{-1})], \tag{9}$$

where

$$1 - P(z^{-1}) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{10}$$

is the LPC inverse filter and μ is a fraction between zero and one. The parameter μ determines the degree to which the error is de-emphasized in the formant regions. Since the weighting filter is linear, the weighting can be done before the error computation. Both the speech signal $s(n)$ and the synthetic speech signal $\hat{s}(n)$ can be passed through $W(z)$ to produce the weighted signals $y(n)$ and $\hat{y}(n)$, respectively, and the weighted error signal computed as $y(n) - \hat{y}(n)$. Let $h(n)$ be the overall impulse response of the all-pole filter in cascade with the weighting filter at the nth sampling instant. Let input pulses of amplitude $\beta_1, \beta_2, \ldots, \beta_m$ occur at instances $n_1, n_2, \ldots, n_m$, respectively. The synthetic speech signal $\hat{y}(n)$ can then be expressed as

$$\hat{y}(n) = \sum_{k=1}^{m} \beta_k h(n-n_k). \tag{11}$$

The amplitudes and locations of the pulses are obtained by minimizing the total squared error E between $y(n)$ and $\hat{y}(n)$, where

$$E = \sum_{n=0}^{N-1} [y(n) - \hat{y}(n)]^2. \tag{12}$$

Minimizing E with respect to β_j, $1 \leqslant j \leqslant m$, leads to

$$\sum_{k=1}^{m} \beta_k \alpha(n_k, n_j) = \chi(n_j), \quad 1 \leqslant j \leqslant m, \tag{13}$$

where

$$\alpha(i,j) = \sum_{n=0}^{N-1} h(n-i)h(n-j) \tag{14}$$

and

$$\chi(m) = \sum_{n=0}^{N-1} y(n)h(n-m). \tag{15}$$

The minimum error $E_{\min}$ is given by

$$E_{\min} = \sum_{n=0}^{N-1} y^2(n) - \sum_{k=1}^{m} \beta_k \chi(n_k). \tag{16}$$

Minimization with respect to the pulse locations leads to a set of nonlinear equations which do not have a closed form solution. Equations (13) cannot be solved directly, because they involve the unknown pulse locations. Thus suboptimal solutions have to be used. In [1], Atal and Remde obtained the pulse amplitudes and locations in successive stages, one pulse at a time. At stage j, all pulse amplitudes and locations up to stage $j-1$ were assumed to be known and only the pulse location n_j and the pulse amplitude β_j were computed. While this procedure is fast, it has several shortcomings. Successive optimization of individual pulses is inaccurate for closely spaced pulses. Additional pulses are often required to compensate for the inaccuracy introduced in earlier stages. As a result, the signal-to-noise ratio (SNR) saturates after a few pulses have been placed.

These problems can be avoided to a large extent if the amplitudes of *all* pulses found in earlier stages are kept optimal along with the amplitude of the current pulse during the search. Thus at stage j we assume that only the pulse locations $n_1, n_2, \ldots, n_{j-1}$ remain constant. Equations (13) and (16) are then used to compute the pulse amplitudes $\beta_1, \ldots, \beta_{j-1}$, as well as β_j and n_j. Since the amplitudes of all pulses can now be modified, the pulse amplitudes remain accurate even when they are closely spaced and the SNR does not saturate.

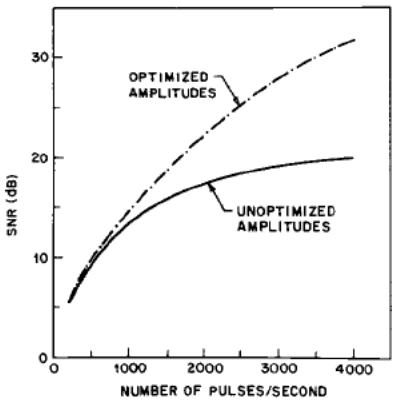

Fig. 3. Average SNR obtained with and without amplitude optimization

Figure 3 shows the SNR obtained with and without amplitude optimization as a function of the pulse rate for the utterance "Both teams started from zero," spoken by a female speaker. The saturation effect is clearly visible when the amplitudes are not optimized. It can be seen that with optimized amplitudes, the SNR increases with increasing number of pulses. For the same pulse rate, the new method provides a higher SNR. Conversely for the same SNR, the new method requires a smaller pulse rate.

Figure 4 illustrates the differences in the excitation obtained with and without amplitude optimization. A small section of speech is shown in the figure along with the excitation obtained with the two methods. The synthetic speech signals are also shown. It can be seen that amplitude optimization removes some superfluous pulses and corrects some pulse locations. It can also be seen that the excitation becomes more periodic for voiced speech.

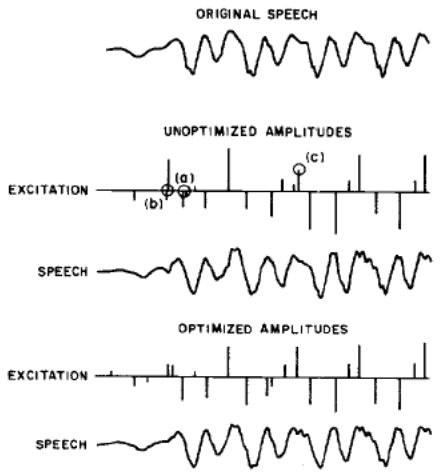

Fig. 4. Illustration of the differences in the excitation obtained with and without amplitude optimization. (a) A superfluous pulse caused by amplitude inaccuracy, (b) a location error, and (c) an irregular pitch pulse.

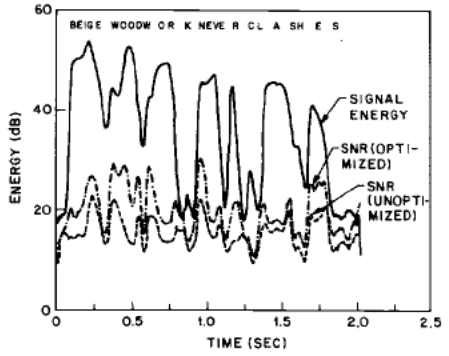

Fig. 5. The signal energy and the SNR obtained by with and without amplitude optimization for the utterance "Beige woodwork never clashes."

Figure 5 shows the signal energy and the SNR obtained with the two methods for a sentence at a pulse rate of 1600 pulses/sec. The improvement is consistent throughout the utterance and not limited to any one section of speech. The degree of improvement varies from 2 to 10 dB with an average of slightly over 3 dB.

BIT-RATE REDUCTION BY PITCH PREDICTION

It has been our experience that about 8 pulses are needed per pitch period in the multi-pulse excitation for high quality synthesis. Thus the bit rate increases as the pitch period is reduced. This implies that the bit rate required for multi-pulse excitation is higher for female speech than for male speech of the same speech quality. Alternately, for the same excitation bit rate speech quality is poorer for female speakers than for male speakers. Figure 6 shows the SNR obtained as a function of the pulse rate for a male and a female speaker. A difference of about 6 to 10 dB can be seen in the SNR for the same pulse rate. For voiced speech, multi-pulse excitation shows significant correlation from one period to the next. We use this correlation to reduce the number of pulses in the excitation substantially by pitch prediction on the multi-pulse excitation.

The pitch predictor is shown in Fig. 7. The LPC synthesizer input $v(n)$ is given by

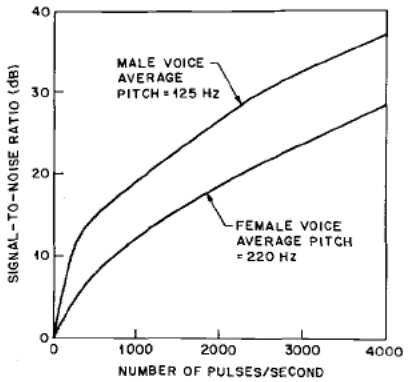

Fig. 6. Average SNR as a function of pulse rate for a female and a male speaker.

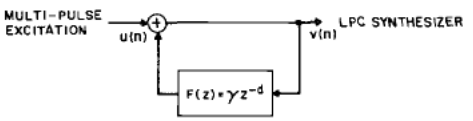

Fig. 7. The pitch predictor filter used for multi-pulse excitation.

$$v(n) = u(n)+\gamma v(n-d), \quad 0 \leqslant n < N, \tag{17}$$

where $u(n)$ is the nth sample of the multi-pulse excitation in the frame, γ is the predictor gain and d is the predictor delay. The predictor delay can be several pitch periods and in general is longer than the frame length N. The synthetic speech signal $\hat{y}(n)$ can now be expressed as

$$\hat{y}(n) = \sum_{k=0}^{N-1} h(k)u(n-k) + \gamma \sum_{k=0}^{n} h(k)v(n-k-d) + \hat{y}(-1)h(n+1). \tag{18}$$

As before, we minimize the mean-squared error E between the original and synthetic speech signals as in Eq. (12). The pitch predictor and the multi-pulse excitation are now obtained in two steps. First, we assume that the multi-pulse excitation is zero and minimize E to compute the predictor gain γ and the delay d. Let

$$\tilde{y}(n) = y(n) - \hat{y}(-1)h(n+1)$$

and

$$z_d(n) = \sum_{k=0}^{n} h(k)v(n-k-d). \tag{19}$$

The minimization of E with respect to γ results in

$$\gamma(d) = [\sum_{n=0}^{N-1} \tilde{y}(n)z_d(n)]/[\sum_{n=0}^{N-1} z_d^2(n)], \tag{20}$$

and

$$E_{\min}(d) = \sum_{n=0}^{N-1} \tilde{y}^2(n) - [\sum_{n=0}^{N-1} \tilde{y}(n)z_d(n)]^2/[\sum_{n=0}^{N-1} z_d^2(n)]. \tag{21}$$

Equation (21) is computed for each value of d in a given range and the value which minimizes $E_{\min}(d)$ is chosen as the predictor delay. The predictor gain γ is then found from Eq. (20). Next, the pitch predictor gain and delay are held constant, and the multi-pulse excitation is found as described in the last section. This procedure implies that the multi-pulse excitation only has to represent the uncorrelated part of the LPC filter excitation. For voiced speech, the excitation is highly correlated, and only a few pulses are needed in the multi-pulse excitation.

Figure 8 illustrates this difference for a segment of speech for a female speaker. The SNR improvement varies from about 2 dB at high bit rates to 5 dB at bit rates around 1000 pulses/sec. This corresponds to a typical savings of 500-600 pulses/sec.

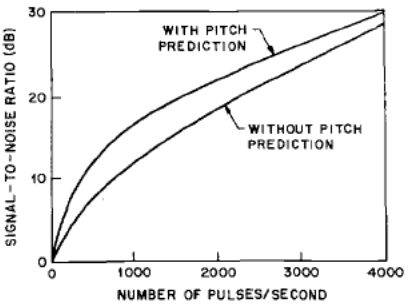

Fig. 8. Average SNR as a function of pulse rate with and without pitch prediction of the excitation for a 1 second segment of speech by a female speaker.

CONCLUSION

In this paper, we described three problems encountered by us in our attempt to reduce the bit rates for multi-pulse excited LPC speech synthesizers and the methods used by us to bypass these problems. Our results show that these synthesizers are capable of producing good quality speech even at low bit rates.

REFERENCES

[1] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Paris, France, 1982, pp. 614-617

[2] L. R. Rabiner, B. S. Atal and M. R. Sambur, " LPC prediction error- analysis of its variation with the position of the analysis frame," *IEEE Trans. Acoust.,Speech, Signal Processing*, vol. ASSP-25, no. 5, Oct. 1977, pp. 434-442

[3] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Commn.*, vol. COM-30, no. 4, April 1982, pp. 600-614

[4] L. R. Rabiner and R. W. Schafer, "*Digital Processing of Speech Signals*," Prentice Hall, 1978