저전송속도 CELP 부호화기에서 여기신호의 개선

권철홍

대전대학교 정보통신공학과

Improving The Excitation Signal For Low-rate CELP Speech Coding

Chul-Hong Kwon

Dept. of Information & Communication Eng., Taejon Univ.

Abstract

In order to enhance the performance of a CELP coder at low bit rates, it would be necessary to make the CELP excitation have the peaky pulse characteristic. In this paper we introduce an excitation signal with peaky pulse characteristic. It is obtained by using a two-tap pitch predictor. Samples of the signal have different gains according to their amplitudes by the predictor. In voiced sound the signal has the desirable peaky pulse characteristic, and its periodicity is well reproduced. Particularly, peaky pulses at voiced onset and a burst of plosive sound are clearly reconstructed.

I. Introduction

One way to lower the coding rate of code-excited linear prediction (CELP) coders below 4.8 kbits/s is to lengthen the analysis frame size for excitation parameters. However, in this case problems can occur, which may be summarized as follows. First, it is very difficult to reconstruct a burst of plosive sound. Without its proper reconstruction, intelligibility of the sound would get much damaged. Secondly, at voiced onset it is difficult to reproduce peaky (or sharpened) pulses in linear predictive coding (LPC) residual signal, and the convergence rate of pitch periodicity is slow. Lastly, in voiced sound the ability to reconstruct a periodic excitation with the peaky pulse characteristic is greatly reduced. Therefore, in order to enhance the performance of the CELP coder at low bit rates, it would be necessary to make the CELP excitation have the peaky pulse characteristic.

In this paper we introduce an excitation signal with peaky pulse characteristic. It is obtained by using a two-tap pitch predictor. Samples of the excitation have different gains according to their amplitudes by the predictor. The excitation obtained by our method has excellent pulse characteristic. Listening tests show that the roughness of the conventional CELP coder at low bit rates is eliminated, and the buzziness does not occur.

II. General Characteristics of LPC Prediction Residual

We now consider how the features of the CELP excitation are related to those of the LPC residual in voiced sound. We first consider the prediction residual in LPC analysis. For a speech signal $s(n)$, the LPC prediction residual $r(n)$ is given by

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

where p and a_k are the order and LPC coefficients of a short-term predictor, respectively. It is known that in voiced sound the LPC residual $r(n)$ has the following three main features that are perceptually important in speech coding [1]-[3]. First, $r(n)$ has large pulses at the beginning of each pitch period. In other words, **major excitation** occurs at the instant of glottal closure. Secondly, **formant structure** of speech signal, particularly in a low-frequency region, remains in $r(n)$ because of inherent shortcomings of LPC analysis.

The CELP excitation consists of an adaptive source represented by a pitch predictor and a stochastic source by a sequence of a Gaussian codebook. The adaptive source produces the major excitation in the LPC residual because it reconstructs the pitch periodicity of speech signal. The major excitation is represented by an impulse train in conventional LPC vocoders. However, these LPC vocoders take no account of the formant structure in the LPC residual [1]. But, examining the excitation search process in CELP coding, we observe that the adaptive source also reconstructs the formant structure of input speech. That is, CELP coding introduces the formant structure similar to that of the input speech to the adaptive source by comparing original and synthesized speech in closed-loop fashion. The weighting filter that accentuates low-frequency and high-energy components in CELP coders also introduces the formant structure to the source.

This observation in the adaptive source and the first two features of the LPC residual as mentioned above can be confirmed in Figs. 1 and 2. Fig. 1 shows the DFT spectrum of a speech signal, and Fig. 2 shows the DFT spectra of the corresponding LPC residual and the adaptive source obtained in CELP coding. The analysis frame size for the source is 40 samples (or 5 msec). From Figs. 1 and 2 we can see that the LPC residual has some of the remaining formant structure similar to that of the corresponding speech signal. The formant structure is faithfully reconstructed by the adaptive source in a low-frequency region (frequencies below 2000 Hz) in Fig. 2. We can also see that the spectrum of the adaptive source shows a harmonic structure, which is reproduced by the major excitation.

We now consider which components of the LPC residual in one pitch period contribute to the formant structure. In Fig. 3, we show the DFT spectrum (dashed line) of the LPC residual in Fig. 2 with the samples zeroing except for the major excitation, and the DFT spectrum (solid line) of the LPC residual with only the major excitation zeroing. The dashed line shows a typical spectrum of an impulse train, and the solid line shows the formant structure in the LPC residual. As seen in this figure, we find that samples except for the major excitation at glottal closure is mainly responsible for the formant structure. Thus, we can divide the LPC residual into two components: one is **the major excitation signal part** and the other is **the formant excitation signal part** that is related to the formant structure. Here, by the major excitation signal we mean the signal that has the largest amplitude within a pitch period, and by the formant excitation signal we mean the signal that reconstructs the formant structure remained in the LPC residual.

III. Formulation of The Proposed Model

So far, we have considered the relation between the LPC prediction residual and the CELP excitation. In this work we will focus on the improvement of the adaptive source, developing a new adaptive source which reconstructs well the major excitation as well as the formant structure remained in the LPC residual.

We first consider the case of the excitation analysis frame size equal to 40 samples (or 5 msec) in which the output of the CELP coder is of toll-quality. Almost all pitch values taken in real speech exceed this value (we assume that pitch delays in a long-term predictor are between 20 and 147 samples). In this case each analysis frame may be in the region that includes the major excitation or in the region without it. Thus, the major and the formant excitation are well reconstructed in separate analysis frames. On the other hand, in case of the excitation analysis frame size equal to 80 samples, many pitch values are less than this value. Therefore, each analysis frame may contain one or more pitch periods. As a result, the quality of the CELP coder can become deteriorated because the adaptive source should reconstruct the major and the formant excitation simultaneously within one frame.

In this paper we propose an adaptive source that approximates the major and the formant excitation separately within one frame in case of the excitation analysis frame size equal to 80 samples. The source is based on a two-tap pitch predictor. The new adaptive source, $e_P(n)$, is given by

$$e_P(n) = e_m(n) + e_f(n) \qquad (2)$$

and

$$e_P(n) = e_m(n) = \beta_1 \Delta, \ n = n_1 \qquad (3)$$

$$e_P(n) = e_f(n) = \beta_2 \, e_P(n-P), \ otherwise \qquad (4)$$

where Δ is a sample with the largest amplitude of $e_P(n-P)$, n_1 gives its position, P is a pitch delay, β_1 and β_2 are the corresponding gain factors, respectively. The first term of right-hand side in eq. (2) represents the major excitation, and the second term models the formant excitation. We assign different gains to a sample with the largest amplitude and to the rest other samples in our proposed source.

The conventional CELP coder searches for the adaptive source parameters by minimizing the perceptually weighted mean-squared error. In the source search process it compares original speech with synthesized speech that is the sum of the major excitation contribution and the formant excitation contribution. However, in this coder the major and the formant excitation have a single identical gain, and it is possible to obtain a lower MSE by matching original speech to the formant excitation contribution rather than to the major excitation contribution. Hence, the source has poor pulse characteristic. But, in our proposed model we can obtain the source with desirable pulse characteristic because the major and the formant excitation have different gains.

Note that β_1, β_2 and P can be obtained by using the same procedure as in searching for the parameters for the two-tap pitch predictor in the conventional CELP coders. The position n_1 is available at the decoder without its transmission by examining past samples of the CELP excitation.

IV. Results And Discussion

To evaluate the performance of the CELP coder with our proposed excitation model, we have done simulations with the following parameters. Simulation was done using a speech data file of 70 sec long. The speech data file consisted of 24 sentences uttered by four male and four female speakers. Speech samples were band-limited with a lowpass filter having 3.2 kHz cutoff frequency and the sampling rate was 8 kHz. The frame length of spectral analysis was 160 samples (or 20 msec). Spectral parameters were obtained by the autocorrelation method. We used a codebook of 1024 codewords that contains samples of a zero-mean, unity-variance, white Gaussian sequence. The range of delay in a pitch predictor was between 20 and 147 samples. We implemented three CELP coders as follows. Reference coder 1 had a two-tap pitch predictor with the excitation analysis frame size equal to 40 samples. Reference coder 2 had a two-tap pitch predictor with the excitation analysis frame size equal to 80 samples. The proposed coder is the same as the reference coder 2 except for the adaptive source model. To see the effectiveness of our proposed source, we did not quantize all the parameters.

The average segmental SNRs of the reference coders 1 and 2 were 12.1 and 9.1 dB, respectively. We can see that the performance of the CELP coder significantly drops in case of lengthening the excitation analysis frame size (i.e., from 40 to 80 samples). The average segmental SNR of our proposed coder was 9.2 dB. This SNR value is similar to that of the reference coder 2. This result can be explained as the following. The CELP coder is a kind of waveform coder. Some parts of speech get distorted if others are heavily accentuated without considering direct waveform matching. Therefore, if two coders with similar structures have the same bit rate, it is expected that the output SNRs should be almost identical. But the output quality can be different perceptually.

Though the average segmental SNRs of both coders are almost the same, simulation results showed that local SNRs in particular sounds were different to a considerable extent. Speech sounds of the proposed coder that were superior to that of the reference coder 2 in terms of SNR were unvoiced sound, particularly a burst of plosive sound, transitions from unvoiced to voiced sound and from voiced to unvoiced sound, and vowels that have strong peaky pulses in the LPC residual. For these sounds it is difficult to reconstruct in case of lengthening the excitation analysis frame size as mentioned in I. Introduction. On the other hand, the reference coder 2 yielded better sound in nasal and nasalized vowels. In these sounds the pulse characteristic of the LPC residual is weak. Because both coders have the average segmental SNRs higher than 15 dB in these sounds, the output quality of both coders was perceptually undistinguishable.

Fig. 4 shows a speech signal, the corresponding LPC residual and the segmental SNRs of the reference coder 2 and the proposed coder in transition from unvoiced to voiced sound. We can see that SNRs of the proposed coder are higher than those of the reference coder 2 in both unvoiced and voiced sounds. Particularly at voiced onset (samples between 8480 and 8560) SNR improvement is almost 4 dB, which is remarkable. Fig. 5 illustrates the excitation waveforms of the reference coders 1, 2 and our proposed coder. The reference coder 2 cannot reconstruct large major pulses in the LPC residual, but our proposed source has excellent pulse characteristic. From the SNR comparison of Fig. 4 (bottom) and the excitation waveforms of Fig. 5 (middle and bottom) we can confirm that the pulse excitation reconstructs satisfactorily the regular structure of speech signal at voiced onset. Note that the steady-state part of vowel sounds continues to have such a pulse characteristic of voiced onset by the feedback search process of a pitch predictor in our proposed coder.

Fig. 6 shows a speech signal, the corresponding LPC residual and the segmental SNRs of the reference coder 2 and the proposed coder in voiced sound. We can see that the proposed coder has higher SNRs than the reference coder 2 over all voiced segments. Fig. 7 illustrates the excitation waveforms of the reference coders 1, 2 and the proposed coder. The LPC residual shows the periodic characteristic of voiced sound and consists of a few large pulses surrounded by a number of small samples within each pitch period. However, in the excitation signal of the reference coder 2 the major pulse is small and samples surrounding the major pulse are larger than those of the LPC residual. This source provides a cause of rough output quality. Therefore, we can conclude that the reference coder 2 cannot well reconstruct the major as well as the formant excitation. On the other hand, the major pulse in the proposed source is outstanding, and samples surrounding the major pulse are smaller than in the reference coder 2. Note that the reference coder 1 not only has excellent pulse characteristic but also reconstructs the formant excitation faithfully.

Fig. 8 shows the DFT spectra of an original and the corresponding reconstructed speech of the reference coder 2 in samples between 5920 and 6080 of Fig. 6. At frequencies below 1500 Hz, spectral envelope mismatches often appear. At frequencies above 1500 Hz, pitch periodicity gets deteriorated and the spectral envelope is much smaller than that of the original speech spectrum. The DFT spectrum of the same speech segment reconstructed by our proposed coder is shown in Fig. 9. Pitch harmonics is reproduced faithfully at lower and higher frequencies. The spectral envelope matches excellently over the total frequency range.

V. Conclusion

In this paper, we have introduced a method to generate CELP excitation signals with peaky pulse characteristic, which is based on a two-tap pitch predictor. In voiced sound the excitation has the desirable peaky pulse characteristic, and its periodicity is well reproduced. Particularly, peaky pulses at voiced onset and a burst of plosive sound are clearly reconstructed. According to subjective quality tests, the synthesized speech by the proposed excitation has little roughness and the clearness is greatly improved.

The proposed model has been conceived based on our observation that the adaptive source of a CELP coder reconstructs the major excitation at glottal closure and the formant structure of the LPC residual.

References

1. G. S. Kang and S. S. Everett, "Improvement of the excitation source in the narrow-band linear prediction vocoder," IEEE Trans. Acoustic., Speech, Signal Processing, vol. 33, no. 2, Apr. 1985, pp. 377-386.

2. T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," IEEE Trans. Acoustic., Speech, Signal Processing, vol. 27, no. 4, Aug. 1979, pp. 309-319.

3. J. L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd ed. New York: Springer-Verlag, 1972, pp. 184-186.

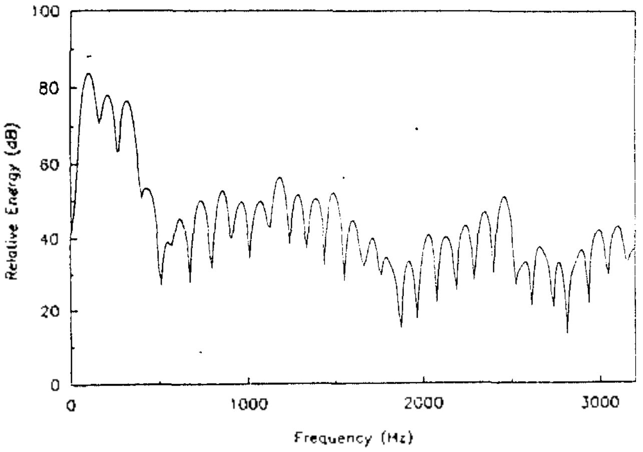

Fig. 1 Spectrum of a speech signal in voiced sound.

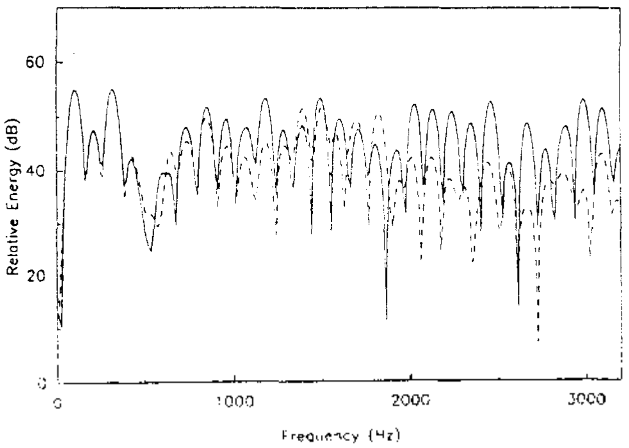

Fig. 2 Comparison of spectra of an LPC residual (solid) and the corresponding signal of adaptive source in a CELP coder (dashed).

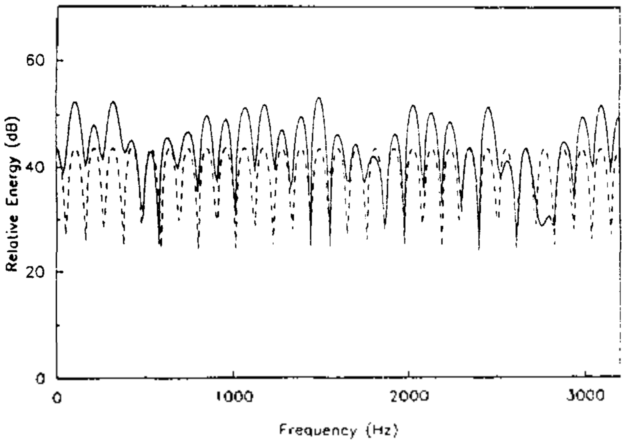

Fig. 3 Spectra of the LPC residual of Fig. 2 with only the major excitation zeroing (solid) and with other samples zeroing except for the major excitation (dashed).

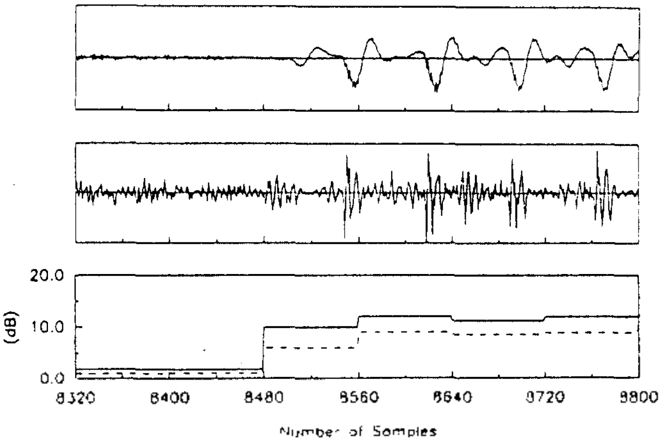

Fig. 4 Waveforms of a speech signal (top) and the corresponding LPC residual (middle), and the SNRs of the reference coder 2 (dashed line of bottom) and our proposed coder (solid line of bottom) at voiced onset. (The LPC residual has been amplified five times.)

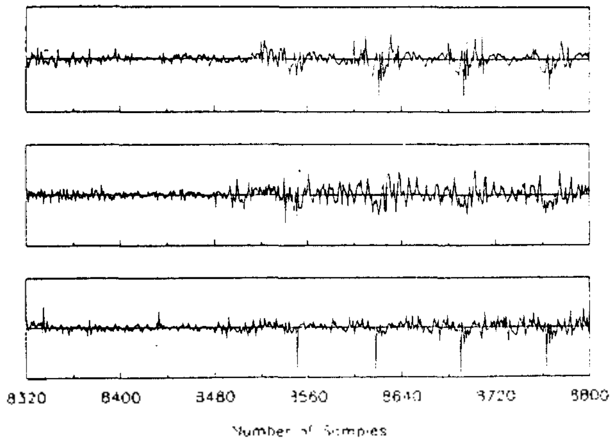

Fig. 5 Excitation signal waveforms of the reference coder 1 (top), the reference coder 2 (middle) and our proposed coder (bottom).

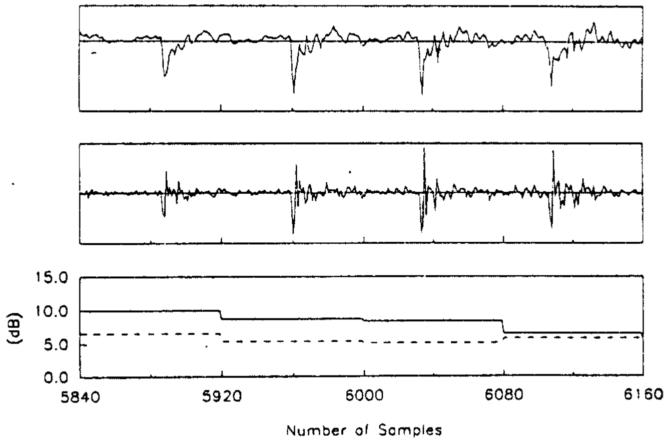

Fig. 6 Waveforms of a speech signal (top) and the corresponding LPC residual (middle), and the SNRs of the reference coder 2 (dashed line of bottom) and our proposed coder (solid line of bottom) in voiced sound. (The LPC residual has been amplified five times.)

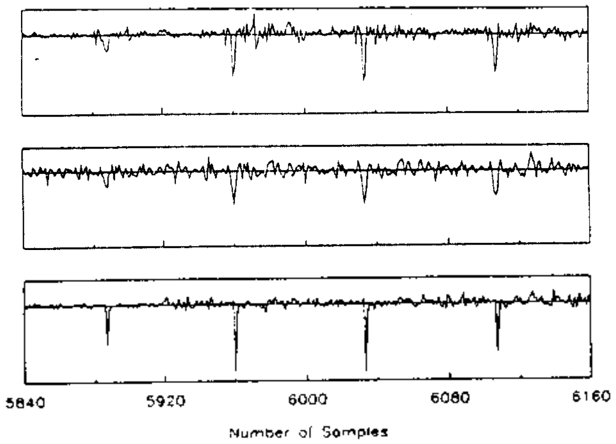

Fig. 7 Excitation signal waveforms of the reference coder 1 (top), the reference coder 2 (middle) and our proposed coder (bottom).

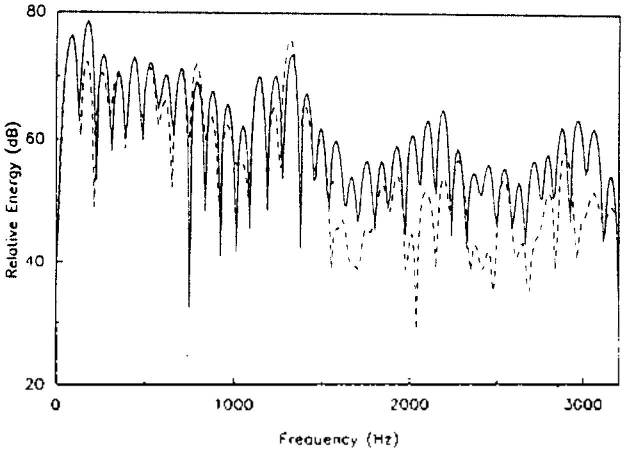

Fig. 8 Comparison of spectra of original speech (solid) and reconstructed speech by the reference coder 2 (dashed).

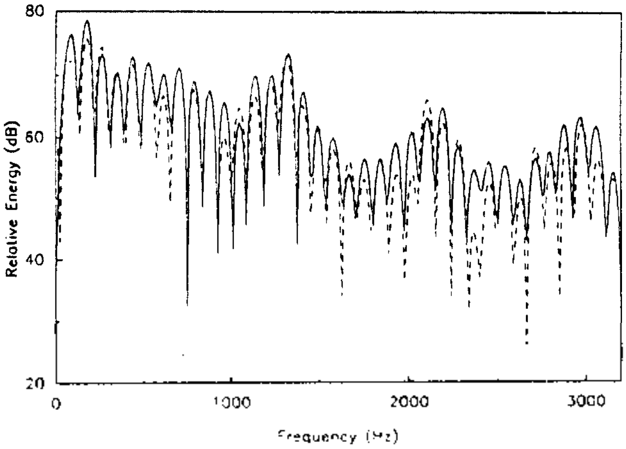

Fig. 9 Comparison of spectra of original speech (solid) and reconstructed speech by our proposed coder (dashed).