LOW-RATE CELP SPEECH CODING USING AN IMPROVED WEIGHTING FUNCTION

Chul Hong Kwon

Digicom Institute of Telematics
Digicom Corporation
828-10 Yeoksamdong Kangnamku
Seoul 135-080 Korea

Chong Kwan Un

Department of Electrical Engineering
Korea Advan. Inst. of Science and Tech.
373-1 Kusungdong Yusungku
Taejon 305-701 Korea

ABSTRACT

Below 4.8 kbits/s, code-excited linear predictive (CELP) coders in general suffer from two kinds of perceptually important degradation. They are inter-harmonic noise and high frequency mismatch. To remedy these degradations, we propose in this paper an improved weighting function which utilizes the spectral weighting methodology and also takes into account the periodic character in voiced sound. The function can adapt to variation of pitch by itself without any pitch estimation in voiced sound and is also applicable to all speech segments without any voiced/unvoiced discrimination algorithm. Simulation results show that the performance of the CELP coder with the proposed weighting function is better than that of the conventional CELP coder.

1. INTRODUCTION

Below 4.8 kbits/s, the quality of a CELP coder in general suffers from two perceptually important degradations [1]. One is the presence of noise-like components between adjacent harmonics, that is, inter-harmonic noise, which gives rise to roughness in voiced sounds. The other is the crude approximation of speech signal at high frequencies, that is, high frequency mismatch.

The presence of the noisy components can be explained by considering the excitation search procedure. Since the squared-error criterion in CELP coding tends to ignore small and evenly-spread errors, the coder can generally choose the correct pitch period regardless of the noise [2]. Considering now the high frequency mismatch, the CELP coder uses a perceptual weighting filter which amplifies the frequency band with high energy (i.e., at lower frequencies) and attenuates the frequency band with low energy (i.e., at higher frequencies). Therefore, these high frequency components tend to be poorly reproduced by the conventional CELP coder.

In general, the performance of the speech coder is heavily dependent on the selection of error criterion. However, the fact that the constrained excitation approach [4] or the pitch sharpening approach [2] improves the subjective quality illustrates the inadequacy of the error criterion typically used in CELP analysis. On the other hand, from the fact that the subjective quality is improved by the comb filtering approach [1] or by the harmonic noise weighting approach [3], we can know the importance of preserving the periodic character in speech signal. However, the perceptual weighting typically used in the conventional CELP coder did not take into account the spectral fine structure of the speech signal. A better understanding of the hearing mechanism may lead to a more appropriate error criterion. Therefore, an improved weighting function needs to be found.

2. FORMULATION OF THE WEIGHTING FUNCTION

In this work, we use a weighted squared-error distortion measure in the frequency domain, defined as

$$d(\boldsymbol{X}, \hat{\boldsymbol{X}}) \equiv (\boldsymbol{X} - \hat{\boldsymbol{X}})^T \boldsymbol{W} (\boldsymbol{X} - \hat{\boldsymbol{X}}) \qquad (1)$$

where $X(n)$ and $\hat{X}(n)$ are discrete Fourier transforms (DFTs) of input speech $x(m)$ and reconstructed output speech $\hat{x}(m)$, respectively, and $\boldsymbol{W}$ is an $N \times N$ weighting matrix.

The choice of $\boldsymbol{W}$ is extremely important for speech coding because of the nonflat frequency characteristic of human perception. Previous methods of weighting in the conventional CELP coder considered only one aspect of applying perceptual

criteria and did not take into account the pitch periodicity of speech signal. However, the weighting function should be chosen in such a manner that quantization noise is most effectively masked by the speech signal. For this purpose we propose an improved weighting function which utilizes the spectral weighting methodology and also accentuates the periodic character in voiced sound. An effective choice of the weighting function can be of the form

$$w(n) = |X(n)|^{2\gamma}, \quad n = 0, \cdots, N-1 \qquad (2)$$

and

$$\boldsymbol{W} = \begin{bmatrix} |X(0)|^{2\gamma} & . & . & . & 0 \\ 0 & . & . & . & . \\ . & . & . & . & . \\ . & . & |X(n)|^{2\gamma} & . & . \\ . & . & . & . & . \\ . & . & . & . & 0 \\ 0 & . & . & 0 & |X(N-1)|^{2\gamma} \end{bmatrix} \qquad (3)$$

where $|X(n)|$ is the spectral magnitude of transform component $X(n)$, and γ is a parameter that can be experimentally chosen and is varied between 0 and -1. Our experiments led to the choice of γ = -0.25.

Figs. 1 and 2 show the DFT spectra of speech signal and the corresponding weighting function in voiced sound of male voice and female voice, respectively. The weighting functions follow the formant structure of the speech spectrum as well as the fine structure (pitch harmonics). Furthermore, our proposed weighting function can adapt to variation of pitch by itself without any pitch estimation. Fig. 3 shows the DFT spectra of speech signal and the corresponding weighting function in unvoiced sound. The weighting function follows the formant structure of speech spectrum. However, the function does not show the periodic character because of the random character in unvoiced sound.

We now consider computational complexities of the codebook search method used in the conventional CELP coder and our proposed scheme. Note that we experimented with 256-point FFT considering the resolution of the DFT spectrum and the computational complexity. Assuming that the codeword length is 60 samples (or 7.5 msec) and the codebook size is 1024, the number of multiply-add operations of one frame search in the conventional CELP coder is 2,181,120 [5]. In our proposed scheme, the codebook made of the codewords in the frequency domain is prepared off-line before the search process. The DFT computations of the input speech and the impulse response are performed once per frame. We also use the property that a convolution operation in the time domain is reduced to a single complex multiplication for each sample in the frequency domain and the Fourier transform of a real sequence is conjugate symmetric. Consequently, the number of multiply-add operations in our proposed scheme is 1,320,960 and the speed-up factor of the proposed method is 1.65. We can reduce the computational complexity further by applying the frequency domain search method proposed by Lee and Un [5].

3. COMPUTER SIMULATION RESULTS AND DISCUSSION

To evaluate the performance of the CELP coder with our proposed weighting function, we implemented a 4.8 kbits/s CELP coder with the following parameters. Table 1 summarizes the bit allocation scheme for this coder.

3.1 Comparison Between the New Weighting Function and Other Weighting Functions

The frequency responses of a comb filter [1] and a harmonic weighting filter [3] show a harmonic structure but do not have spectral envelope of speech signal. Thus, the filters should be cascaded by a spectral weighting filter used in a conventional CELP coder. On the other hand, we implement a spectral envelope and harmonic weighting by a single filter as shown in Fig. 1.

In the comb and the harmonic weighting filter approach, a pitch estimate is needed prior to filtering. However, an incorrect pitch estimate results in undesirable pitch harmonics in the filter output. The accuracy of pitch estimation is thus very important but is much deteriorated at low bit rates. But, in our proposed weighting function any parameters related to pitch are not included. Therefore, it is not necessary to estimate pitch in implementing the weighting function and suffer from the incorrect filtering.

Many voiced segments show pitch variation in the frequency domain. In other words, pitch values between a low-frequency and a high-frequency region may be slightly different. However, the comb and the harmonic weighting filter have a single pitch value over the total frequency range, and thus the use of the filters introduces the buzziness in output speech. Application of these filters to unvoiced and transient segments also introduces unwanted pitch periodicity, leading to the buzzy speech quality. On the other hand, our proposed weighting function takes spectral magnitudes at equi-distant frequencies as its component values. Therefore, the function in voiced region follows pitch harmonics of input speech but in unvoiced region

does not show periodic characteristic. Hence, our proposed weighting function does not introduce unwanted buzziness in some voiced and unvoiced segments unlike the comb or the harmonic weighting filter.

3.2 Performance Comparison Between Proposed Coder and Conventional CELP Coder

We used the segmental SNR, SNR_{seg}, as an objective speech quality measure for its mathematical simplicity, and we used informal listening tests as a supporting measure. The output SNR_{seg} of a conventional CELP coder at 4.8 kbits/s was 10.9 dB. Although the performance improvement by objective testing appears to be not very impressive, the SNR_{seg} of a 4.8 kbits/s CELP coder with the proposed weighting function was 11.6 dB, 0.7 dB higher than that of the conventional CELP coder. With the proposed weighting function, SNR improvement in case of male voice was 0.5 dB and in case of female voice was 0.9 dB. It is natural that SNR improvement of female voice is higher than that of male voice because pitch periodicity of female voice is more important than that of male voice. The improvement in the output speech with the proposed weighting function was also confirmed by informal listening tests. According to our informal listening tests, the output speech generated by the CELP coder with the proposed weighting function has less roughness and sounds cleaner than that of the conventional CELP coder.

Fig. 4 illustrates the FFT spectra of the original voiced segment and its coded version with the conventional CELP coder at 4.8 kbits/s. Pitch periodicity near 700 Hz gets irregular and the regions between harmonics near 2 kHz are filled with noise. Also, pitch harmonic mismatch is significant above 1.5 kHz. That is, we can see that there is high frequency mismatch between the spectra. Spectral envelope mismatch sometimes occurs. The FFT spectrum of the same segment coded by the CELP coder with the proposed weighting function is shown in Fig. 5. The noise between harmonics is now removed and pitch harmonics is reproduced faithfully at lower frequencies as well as higher frequencies. Spectral envelope of the coded spectrum is well matched with that of the original spectrum over the entire range of frequency. Also, high frequency mismatch does not occur.

Finally, let us examine the excitation sources that are produced in the conventional CELP coder and the CELP coder with the proposed weighting. Fig. 6 (a) and (b) show the input speech and the LPC residual in the voiced sound, respectively. The LPC residual shows the periodic character of voiced sound and has a few large pulses followed by a number of small pulses in each pitch period. The excitation source generated by a conventional 4.8 kbits/s CELP coder is illustrated in Fig. 6 (c). We can see that the large pulses in the excitation are not clearly defined and pulses occurring subsequently are larger than those of the original LPC residual. The excitation produced by the CELP coder with the proposed weighting is shown in Fig. 6 (d). It is seen that the large pulses in the excitation are clearer and other smaller pulses are weaker than in the conventional CELP coder.

IV. CONCLUSION

In this paper, we proposed an improved weighting function in CELP coding. The proposed function exploits the spectral masking methodology and also takes into account the periodic character in voiced sound. Its salient feature is that it can adapt to variation of pitch by itself without any pitch estimation in voiced sound and is applicable to all speech segments without any voiced/unvoiced discrimination algorithm.

Spectra of speech segment generated by the proposed method reveal that the noise between adjacent harmonics is removed and high frequency mismatch does not occur. The periodic pulse behavior of the excitation source produced by the proposed method is more apparent than that of the conventional CELP coder.

REFERENCES

[1] S. Wang and A. Gersho, "Improved excitation for phonetically-segmented VXC speech coding below 4 kb/s," Proc. IEEE Global Telecomm. Conf., pp. 946-950, 1990.

[2] T. Taniguchi, M. Johnson and Y. Ohta, "Pitch sharpening for perceptually improved CELP, and the sparse-delta codebook for reduced computations," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 241-244, 1991.

[3] I. A. Gerson and M. A. Jasiuk, "Techniques for improving the performance of CELP-type speech coders," IEEE J. Selected. Areas Commun., Vol. 10, No. 5, pp. 858-865, 1992.

[4] Y. Shoham, "Constrained-stochastic excitation coding of speech at 4.8 kb/s," Advances in Speech Coding, Kluwer Academic Publishers, Massachusetts, pp. 339-348, 1991.

[5] J. I. Lee and C. K. Un, "On reducing computational complexity of codebook search in CELP coding," IEEE Trans. Commun., Vol. 38, No. 11, pp. 1935-1937, 1990.

Table 1. BIT ALLOCATION SCHEME FOR THE 4.8KBITS/S CELP CODER

	Spectral Analysis	Pitch Search	Codebook Search
Update Rate	30 msec	30/4 = 7.5 msec	30/4 = 7.5 msec
Bits/Frame	34 bits, 10 LSPs	4 index : {7,5,7,5} 4 prediction gain : {5,5,5,5}	4 index : {10,10,10,10} 4 codeword gain : {5,5,5,5}
Bit Rate	1133.3 bits/s	1466.7 bits/s	2000 bits/s

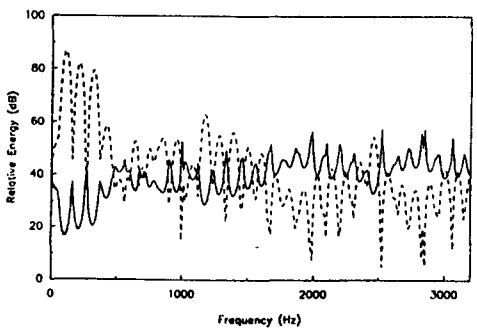

Fig. 1 Spectra of a speech signal (dashed) and the corresponding weighting function (solid) for voiced sound of male speech.

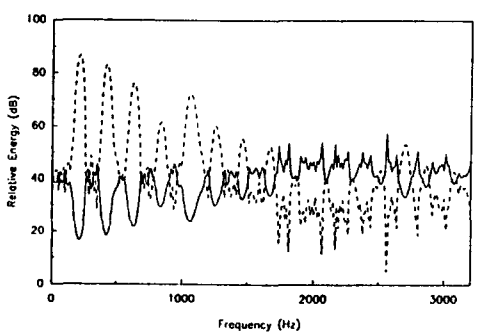

Fig. 2 Spectra of a speech signal (dashed) and the corresponding weighting function (solid) for voiced sound of female speech.

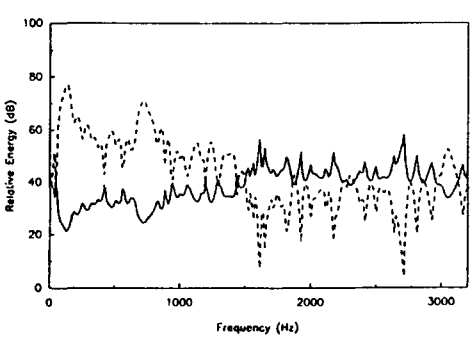

Fig. 3 Spectra of a speech signal (dashed) and the corresponding weighting function (solid) for unvoiced sound of male speech.

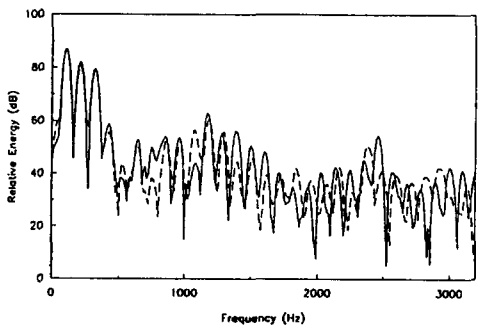

Fig. 4 Comparison of spectra of an original speech (solid) and the coded speech by the conventional 4.8 kbits/s CELP coder (dashed).

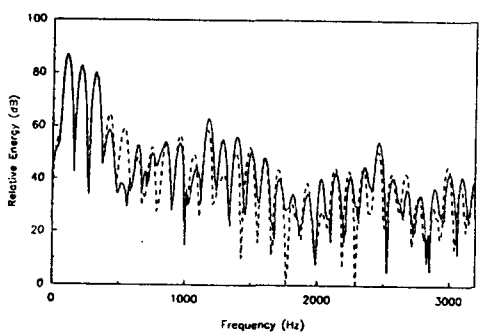

Fig. 5 Comparison of spectra of an original speech (solid) and the coded speech by the CELP coder with the proposed weighting function (dashed).

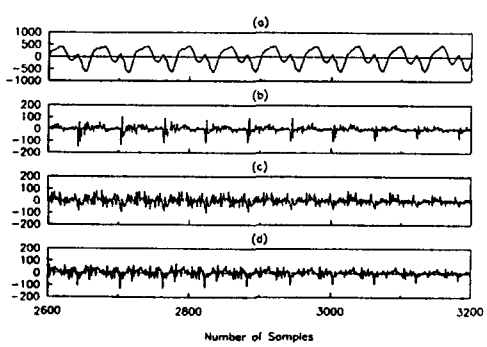

Fig. 6 Waveforms of (a) an original speech, (b) its LPC residual, (c) the corresponding excitation waveform of the conventional 4.8 kbits/s CELP coder and (d) that of the CELP coder with the proposed weighting function.