Improved Excitation Modeling for Low-Rate CELP Speech Coding

*Chul-Hong Kwon

Abstract

In this paper, we propose a weighting dependent mixed source model (WD-MSM) coder that is an improved version of a CELP-based mixed source model (C-MSM) coder. The coder classifies speech segments into three types : voiced, unvoiced and mixed. The excitation for a voiced frame is an adaptive source, and the excitation for an unvoiced frame is a stochastic source. The coder has a modified mixed source for a mixed frame. We apply different weighting functions for three classes. Simulation results show that the proposed coder at 4 kbits/s yields very good performance both subjectively and objectively.

I. Introduction and review of previous works

I.1 Consideration on excitation modeling

Excitation modeling typically used in CELP coders is such that an excitation signal consists of signals from an adaptive source which models long-term correlation of speech signal and a stochastic source which has random character. Such an excitation modeling approach becomes inadequate to represent the time-varying characteristics of a speech signal at low bit rates. Therefore, different excitation modeling of speech segments having different characters is desired for the reduction of bit rate while maintaining a given level of speech quality.

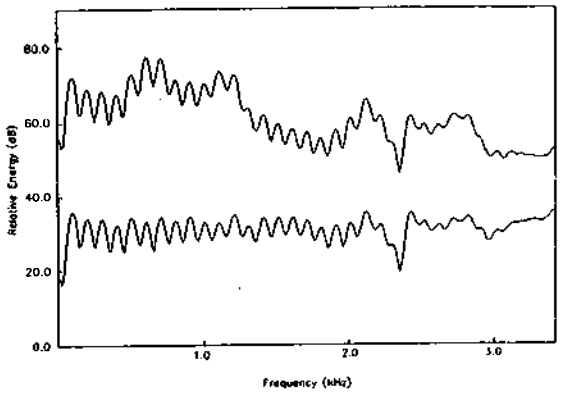

Figure 1. Speech spectrum (top) and residual spectrum (bottom) for vowel /a/.

Careful observation of speech spectra shows that there are many voiced segments where a high-frequency range is almost completely devoiced. This character of speech signal must be taken care of in speech coding since the human hearing system can discriminate between frequency regions dominated by pitch harmonics and those dominated by noise-like energy. Typical examples are shown in Figs. 1 and 2. Fig. 1 shows the speech spectrum of 180 samples of an 8 kHz sampled signal corresponding to a portion of vowel /a/ and that of the corresponding residual signal. And, Fig. 2 shows the speech spectrum of consonant /n/ and that of the corresponding residual signal. The residual signals were obtained by inverse filtering the speech signal with a 10th order linear prediction inverse filter. The speech samples are voiced and sound quite normal. However, we can find partial devoicing above 2 kHz in Fig. 1 and above 1.5

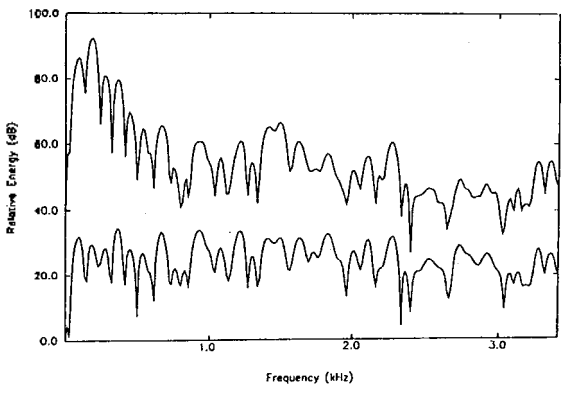

Figure 2. Speech spectrum (top) and residual spectrum (bottom) for consonant /n/.

* Taejon University
Manuscript Received : February 8, 1999

kHz in Fig. 2. We conclude that the spectra of the speech signal may contain some spectral regions that are periodic and others that are aperiodic (mixed frame).

Based on above observation, we divided speech segments into three types of class in a CELP-based mixed source model (C-MSM) coder shown in Fig. 3 [1]. They are voiced, unvoiced and mixed. The spectral fine structure of an unvoiced frame is random and does not show any periodic behavior. Hence, the excitation in unvoiced segments can be modeled by the stochastic source only. However, the spectral fine structure of a voiced frame has periodic characteristic and is determined by a pitch period. Thus, we can use only the adaptive source in voiced segments. The spectrum of a mixed frame is divided into periodic low-frequency and random high-frequency regions as shown above. Therefore, we can model the periodic low-frequency region by the adaptive source and the devoiced high-frequency region by the stochastic source. We refer to the combination of the two sources as a mixed source.

I.2 Consideration on weighting function

In order to make quantization noise perceptually inaudible in a speech coder, it is necessary to consider the spectrum of the quantization noise and its relation to the speech spectrum. The theory of auditory masking suggests that noise in the frequency regions where speech energy is concentrated (such as formant regions) would be partially or totally masked by speech signal[2]. In other words, the perceived noise mainly comes from those frequency regions where the speech energy is low. Consequently, adaptive shaping of the spectrum of the quantization noise is essential. But, the question remains to be unsolved as to what the most appropriate form of weighting is for the optimum subjective performance of the coder. In general, this weighting function should be dynamic, that is, weighting for voiced speech should be different from that for unvoiced speech. Therefore, the weighting function should be chosen in such a manner that the quantization noise is most effectively masked by the speech signal.

In a conventional CELP coder the excitation search process typically uses the following for perceptual weighting[3]:

$$W(z)=\frac{1+\sum_{i=1}^{p}a_i z^{-i}}{1+\sum_{i=1}^{p}a_i \gamma^i z^{-i}} \qquad (1)$$

where $\{ a_i \}$ and p are the coefficients and the order of a short-term predictor, respectively, and γ is a parameter between 0 and 1 that controls the amount of perceptual weighting applied to an error signal. The error signal between an input and a corresponding synthesized speech is processed by a perceptual weighting filter $W(z)$ in order to attenuate the frequency band where the error is perceptually less important and to amplify the frequency band where the error is perceptually more important. Therefore, the perceptual weighting function $W(z)$ allows the spectrum of the coded noise to follow the spectral envelope of the input speech. Therefore, it is adequate for speech segments with random character (such as unvoiced frames) where the reproduction of spectral envelope is particularly important.

However, the weighting function of (1) does not take into account the spectral fine structure of speech signal, that is, it does not have terms to consider pitch harmonics of the speech signal. Thus, it may be inappropriate for speech segments with periodic characteristic. In the previous paper[4], we proposed an improved weighting function as the following:

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

where $X(n)$ is the discrete Fourier transform (DFT) of input speech $x(m)$, $|X(n)|$ is the spectral magnitude of the transformed component $X(n)$, and γ is a parameter that can be experimentally chosen and is varied between 0 and -1. This weighting function utilizes the spectral weighting methodology like the perceptual weighting filter of (1) and accentuates the periodic character in voiced region. We showed in [4] that the weighting function of (2) is more adequate for voiced region than the function of (1).

In this paper we propose a weighting dependent mixed source model (WD-MSM) coder that is an improved version of the C-MSM coder. The WD-MSM coder has a modified mixed source for mixed frame. Parameters of the excitation source for each class are obtained by applying different weighting to each class.

The remainder of this paper is organized as follows. In section II, the proposed coder at 4 kbits/s is formulated. In section III, simulation results of the proposed coder are shown and discussed. Finally we make a conclusion in section IV.

II. Description of the proposed model

The WD-MSM coder consists of four major components: a short-term predictor, an adaptive source, a stochastic source, and a modified mixed source. Bit allocation for the proposed coder at 4 kbits/s is shown in Table 1. The frame length for spectral analysis is 160 samples or 20 msec. The frame length for excitation analysis is 80 samples or 10 msec. 1 bit to classify mixed and non-mixed frame for input speech, 1 bit to classify voiced and unvoiced for non-mixed frame, and 2 bits to specify the cutoff frequency for mixed frame are required.

Table 1. Bit allocation for the 4 kbits/s WD-MSM coder (bits) (frame length = 20 msec, subframe length = 10 msec).

Parameter	Voiced	Unvoiced	Mixed
Adaptive source	(7+6)/subframe x 2 subframes	0	(7+5)/frame
Stochastic source	0	(8+5)/subframe x 2 subframes	(8+5)/frame
Mixed/non-mixed classification	1/frame	1/frame	1/frame
V/uv classification	1/frame	1/frame	0
Cutoff frequency Fc	0	0	2/frame
Short-term predictor	24/frame	24/frame	24/frame

A. Short-term predictor

For the spectral parameters we used the 24-bit split vector quantizer proposed by Paliwal and Atal[5] like in the C-MSM coder.

B. Unvoiced frame excitation

We use only the stochastic source for unvoiced frame excitation as mentioned in Section I.1. As for the weighting function to be used in the excitation search process, we note that the spectral fine structure of unvoiced sound does not have pitch harmonics but has random characteristics. Therefore, we do not need the weighting function that takes into the periodic character account. Hence, we conducted experiments with the weighting function of (1) for the unvoiced frame.

Since we implemented the coder at 4 kbits/s, more bits can be allocated to excitation parameters of the unvoiced frame than in the 3 kbits/s C-MSM coder. We experimented with various bit allocations, that is, larger codebook size, more bits for gain, and reduced frame rate. From this experiment we conclude that it is important to keep the excitation frame length small for unvoiced sound. Consequently, this leads to divide an unvoiced frame into two subframes. For each frame of the unvoiced frame, the stochastic source is a scaled code vector of the stochastic codebook which needs 8 bits to specify 256 random Gaussian sequences, and 5 bits to encode the gain.

C. Voiced frame excitation

For speech segments with periodic characteristic, we mentioned in Section I.2 that the weighting function of (2) makes a good performance. Therefore, we apply the improved weighting function of (2) in the excitation search process for the voiced frame. The voiced frame excitation should have the periodic characteristic, and therefore we test four excitation types as shown in Table 2. First, we can use the same adaptive source as used in the C-MSM coder. In this case, a frame is divided into two subframes and the sources are obtained separately for each subframe. The adaptive source is generated from a third-order long-term predictor. Second, we can have delta encoding of pitch delay of the adaptive source. Dividing a frame into four subframes, an optimal pitch delay for the first subframe is obtained. Then, for the subsequent subframes pitch delays are found within a predefined search range around pitch delay of the first subframe. Here we use the fact that strong stationarity between adjacent subframes in voiced sound exists. Third, we test excitation of the self-excited vocoder (SEV) proposed by Rose and Barnwell[6]. In this work the SEV gets the excitation signal from two third-order long-term predictors. Finally, we can apply pulse excitation like in the multipulse LPC coder. In this case the adaptive source is first obtained, and then a pulse is found. We can see in section III that the second case yields the best result.

D. Mixed frame excitation

If a frame is classified into mixed, the cutoff frequency which divides the entire frequency range into voiced and unvoiced region should be determined. The closed-loop method we propose in this work is as follows. The determination of the cutoff frequency that divides the spectrum into voiced/unvoiced (v/uv) region is done by choosing a proper excitation source for the frequency band instead of examining the periodicity of the spectrum. First, the full band is divided into several bands with equal bandwidth. Then, the minimum-squared errors for the stochastic source are obtained in each band. The minimum-squared errors for the adaptive source are also obtained in each band. The v/uv decision for each band is made by comparing the minimum-squared error for the stochastic source with that for the adaptive source and by choosing the source with smaller minimum-squared error.

That is, if the former is smaller than the latter, the band is regarded as unvoiced. Examining v/uv decision for all the bands, unvoiced bands below voiced band are regarded as voiced. The reason for this way of decision is that the stochastic source is more likely to excite higher frequencies. As for the cutoff frequency, it is determined as a boundary point between voiced and unvoiced regions. If all the bands are declared as voiced, the frame is regarded as voiced frame. On the other hand, if all the bands are declared as unvoiced, the frame is determined as unvoiced frame.

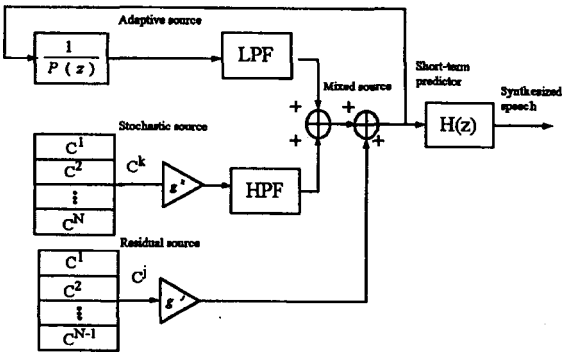

Figure 3. CELP based mixed source model $(P(z)=1+\sum_{i=-1}^{1} b_i z^{-(D+i)})$.

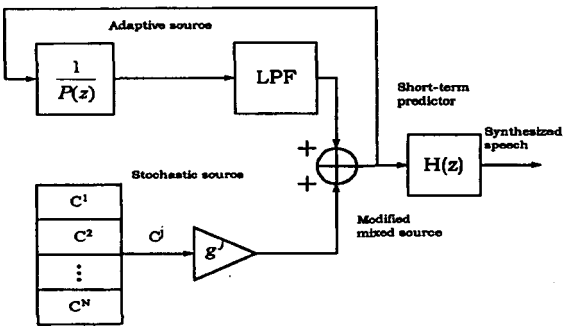

Figure 4. Proposed modified mixed source model $(P(z)=1+b\,z^{-D})$.

Here, we propose a modified mixed source for the mixed frame excitation as shown in Fig. 4. A low-frequency region of the spectrum is generated from the lowpass filtered adaptive source and the remaining region is generated from the stochastic source. Comparing with the mixed source in Fig. 3, the modified mixed source does not have the highpass filtered stochastic source.

Let us discuss the weighting function to be used for each source. The adaptive source which excites a low-frequency region of the modified mixed source models periodic components of an excitation signal. Hence, it is desirable to use the weighting function which accentuates periodic characteristics. The improved weighting function of (2) would be a good choice. As for the stochastic source, we use a highpass-filtered version of the weighting function $W(z)$ of (1) as the following:

$$W_{HP}(z)=(1-\xi z^{-1})W(z) \tag{3}$$

where ξ=0.4 k_1 and k_1 is the first LPC reflection coefficient. The mixed source of the C-MSM coder has a highpass-filtered stochastic source and a residual source shown in Fig. 3. However, by utilizing the highpass-filtered weighting function, we can simplify the highpass filtered stochastic and residual sources of the C-MSM coder by one source, i.e., the stochastic source.

The procedure to obtain parameters for the adaptive source of the modified mixed source is as follows. $X(n)$ represents the DFT of an input speech vector after subtracting the zero input filter response. $X_L(n)$ denotes its lowpass-filtered version with the cutoff frequency F_c described above. Let $Y(n)$ be the DFT of the output of a long-term predictor with a pitch delay D,

$$Y(n)=bY(n)z^{-D} \tag{4}$$

where b is the gain of the long-term predictor. Let $Y_L(n)$ represent a lowpass-filtered version of $Y(n)$ with the same cutoff frequency F_c. Then, the reconstructed speech $X_L(n)$ generated by an input $Y_L(n)$ is expressed as

$$\hat{X}_L(n)=bS_L(n) \tag{5}$$

where $S_L(n)$ is the Fourier transform of a convolution of the filter input $y_L(m)$ and its impulse response $h(m)$, that is,

$$s_L(m)=h(m)*y_L(m) \tag{6}$$

and

$$S_L(n)=H(n)Y_L(n) \tag{7}$$

where the asterisk denotes convolution operation, $H(n)$ and $Y_L(n)$ represent Fourier transforms of $h(m)$ and $y_L(m)$, respectively. The total squared error E is expressed in frequency-domain notations as

$$E=\sum_n |X_L(n)-bS_L(n)|^2 \cdot w(n) \tag{8}$$

where $w(n)$ is the improved weighting function of (2). The optimum gain b that minimizes E can easily be

obtained by setting $\partial E / \partial b = 0$, and is given by

$$b = \frac{\mathrm{Re}(\sum_n X_L^*(n) S_L(n) \cdot w(n))}{\sum_n |S_L(n)|^2 \cdot w(n)} \quad (9)$$

where $X_L^*(n)$ is a complex conjugate of $X_L(n)$ and Re( ) represents a real part. The optimum pitch delay D is obtained by finding the minimum of E for a predefined range of D.

For the adaptive source, 7 bits to specify 128 different pitch delays between 20 and 147 samples and 5 bits to encode the gain of a first-order long-term predictor are required. For the stochastic source, 8 bits to specify 256 code vectors of the stochastic codebook and 5 bits to encode the scale factor are sufficient. For determination of the cutoff frequency Fc, we divide the full band into four bands. Therefore, 2 bits are required to specify the cutoff frequency.

III. Simulation results and discussion

To evaluate the performance of the proposed model, we have conducted experiments with the following parameters. We used speech samples of 70 sec long 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.

Distributions of the gain of the adaptive source are shown in Figs. 5 and 6. For voiced frame the gain is concentrated near 1. This reflects strong stationarity of voiced frames. The gain is distributed in a narrow region and is symmetrical in both sides of 1. Thus, we can quantize the gain more accurately for the same number of

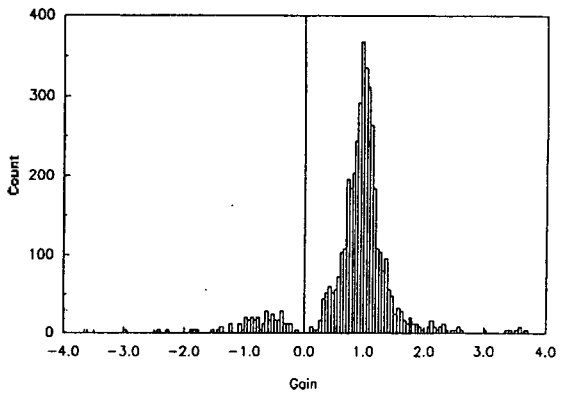

Figure 5. Distribution of the gain of the adaptive source for voiced frame.

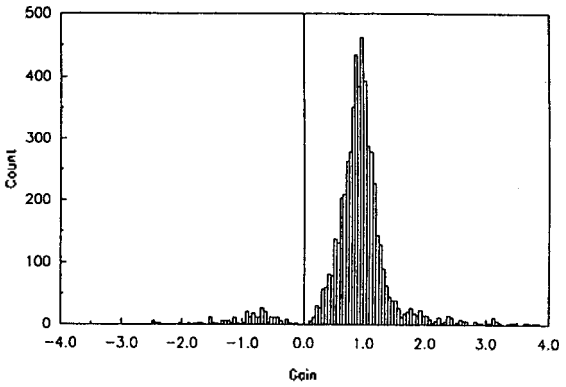

Figure 6. Distribution of the gain of the adaptive source for mixed frame.

bits than in the conventional CELP coder (We assume that the gain in the conventional CELP coder is distributed between -1 and 2). Note that for mixed frames the gain has distribution similar to that of voiced frames.

As shown in Table 1, 6 bits for voiced frames and 5 bits for mixed frames are allocated to encode the gain of the adaptive source. However, the gains of both types have similar distribution as shown in Figs. 5 and 6. The difference of bit allocation between two types of class is explained as follows. We can use a first-order or multiple-order long-term predictor for the adaptive source. However, we find the gain suitable for the low-frequency region for a mixed frame as mentioned above. Thus, we use a first-order long-term predictor with a gain encoded by 5 bits for the mixed frame. On the other hand, a constant single gain for a voiced frame can cause distortion in the high-frequency region. Thus, we use a multiple-order predictor for the voiced frame. The order of the predictor to represent the adaptive source efficiently is sufficient to be 3, and 3 coefficients are encoded using 6 bits by a vector quantizer.

In Section II.C, we explained four excitation types for the voiced frame. We obtained the segmental SNR, SNR_{seg}, to compare the performance of a conventional 4.8 kbits/s CELP coder with that of our 4 kbits/s coder using

Table 2. Segmental SNR for four voiced frame excitation types in Section II.C.

Excitation type	SNR_{seg}
First case	9.5 dB
Second case	10.9 dB
Third case	9.5 dB
Fourth case	9.4 dB

four excitations. The output SNR_{seg} of the 4.8 kbits/s CELP coder was 9.1 dB. As shown in Table 2, SNR_{seg} was 9.5 dB for the first case of four excitations, 10.9 dB for the second case, 9.5 dB for the third case, and finally 9.4 dB for the fourth case. Therefore, we see that the performance of our coders is slightly or much better than that of the 4.8 kbits/s CELP coder. Next, we compare the performance of four excitations in the voiced frame. We first consider excitations for the first and third cases. The first case is the system in which a single adaptive source updates twice per frame, and the third case is the system in which two adaptive sources update once per frame. The SNR_{seg}' s for the two cases are almost the same. Therefore, we can conclude that the performance of the coder has no relation with the excitation update rate if it has the same number of adaptive sources. We next examine whether a pulse excitation provides a more effective means of coding the periodic structure of speech segments than the long-term predictor by comparing the third and fourth cases. The two cases have similar SNR_{seg}' s, and also show almost undistinguishable subjective quality according to our listening tests. Hence, we can know that the pulse excitation yields the performance similar to the long-term predictor in reproducing voiced frames. The simulation result for the second case indicates that the second excitation yields the best performance of four excitation types, leading to conclude that it is better to update the excitation more frequently in the voiced frame. Therefore, we can see that since the long-term predictor models the effect of the glottis which changes faster than the vocal tract shape, it is important to update the pitch parameters more frequently than the formant parameters. Hence, the subframe length, which is the adaptation interval of the long-term predictor, needs to be reasonably small to handle short pitch periods.

IV. Conclusions

In this paper, we propose an weighting dependent mixed source model (WD-MSM) coder that is an improved version of a CELP-based mixed source model (C-MSM) coder. The WD-MSM coder has a modified mixed source for mixed frame. Parameters of the excitation source for each class are obtained by applying different weighting to each class. The excitation for a voiced frame is an adaptive source, and the excitation search process uses an improved weighting function. The excitation for an unvoiced frame is a stochastic source, and the spectral noise weighting function like in the conventional CELP coder is used. For a mixed frame we propose a modified mixed source which combines a lowpass-filtered adaptive source and a stochastic source. The excitation search process for the lowpass-filtered adaptive source uses a weighting function which accentuates periodic characteristics, that is, the improved weighting function. As for the stochastic source, we use a highpass-filtered version of the spectral weighting function of the conventional CELP coder. Simulation results show that the proposed coder at 4 kbits/s yields very good performance both subjectively and objectively.

References

1. C. H. Kwon and C. K. Un, "CELP based mixed-source model for very low bit rate speech coding," *IEE Electronics Letters*, vol. 29, no. 2, pp. 156-157, 1993.
2. J. L. Flanagan, M. R. Schroeder, B. S. Atal, R. E. Crochiere, N. S. Jayant, and J. M. Tribolet, "Speech coding," *IEEE Trans. on Commun.*, vol. COM-27, no. 4, pp. 710-737, Apr. 1979.
3. P. Kroon and E. F. Deprettere, "A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbits/s," *IEEE J. Select. Areas Commun.*, vol. 6, no. 2, pp. 353-362, 1988.
4. C. H. Kwon and C. K. Un, "Low-rate CELP speech coding using an improved weighting function," *IEEE Proceedings of ICASSP 97*, pp. 743-746, 1997.
5. K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Proceedings of ICASSP 91*, pp. 661-664, 1991.
6. R. C. Rose and T. P. Barnwell, "Design and performance of an analysis-by-synthesis class of predictive speech coder," *IEEE Trans. Acoustic., Speech, Signal Processing*, vol. 38, no. 9, pp. 1489-1503, 1990.

▲Chul-Hong Kwon

Chul Hong Kwon was born in Paju, Korea, in 1963. He received the B.S. degree in electronics engineering from Seoul National University in 1987, and the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science

and Technology(KAIST), Taejon, Korea, in 1989 and 1994, respectively. From 1989 to 1997 he was a Senior Researcher at the Digicom Institute of Telematics. He is with Department of Information and Communication Engineering, Taejon University, Taejon, Korea, where he is Professor. His current research interests include speech signal processing and wireless communication.