CELP CODING BASED ON MEL-CEPSTRAL ANALYSIS

Kazuhito Koishida†, *Keiichi Tokuda*††, *Takao Kobayashi*† *and Satoshi Imai*†

†Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama, 227 Japan
††Department of Electrical and Electronic Engineering, Tokyo Institute of Technology, Tokyo, 152 Japan

ABSTRACT

In this paper, we propose a CELP coder based on mel-cepstral analysis. In the coder, since the transfer functions of perceptual weighting and postfiltering are defined through mel-cepstral coefficients, the effects of perceptual weighting and postfiltering should fit with the characteristics of the human auditory sensation. We use a basic CELP structure without adaptive codebook, and subjective speech quality of the proposed coder in terms of the opinion equivalent Q is measured and compared with that of the conventional CELP coder. It is shown that the improvement of more than 1.8dB is achieved by the proposed coder over the conventional CELP coder.

1. INTRODUCTION

In the past years, much has been done to improve the quality of speech coders at low bit rate. Code Excited Linear Prediction (CELP) [1] is one of the most effective coding method at low bit rate. CELP coding has used the AR spectral representation for short-term predictor. Although spectral zeros are important in some cases, AR modeling cannot represent them. On the other hand, cepstal modeling can represent spectral poles and zeros with equal weights. Furthermore, the spectrum represented by the the mel-cepstral coefficients has frequency resolution similar to that of the human ear which has high resolution at low frequencies. Therefore, it is expected that mel-cepstral representation is used for efficient spectral modeling in speech coders instead of the AR modeling. From the above view point, we have proposed a 16kb/s ADPCM coder which produces a high quality speech corresponding to that of CCITT 32kb/s G.721 [2].

In this paper, we propose a CELP coder based on the mel-cepstral analysis [3] and show effectiveness of the mel-cepstral representation in low bit rate speech coding. In the coder, since perceptual weighting and postfiltering are carried out through the mel-cepstral coefficients, the effects of perceptual weighting and postfiltering should fit with characteristics of the human auditory sensation. The subjective speech quality of the proposed coder in terms of the opinion equivalent Q is measured and compared with a conventional CELP coder. In the subjective test, we use a basic CELP structure without adaptive codebook. It is shown that the improvement of more than 1.8dB is achieved by the proposed coder over the conventional CELP coder.

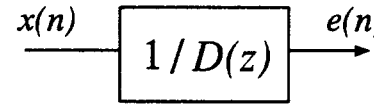


Fig. 1 Mel-cepstral analysis.

2. MEL-CEPSTRAL ANALYSIS

We model speech spectrum $D(e^{j\omega})$ by using the mel-cepstral coefficients $\tilde{c}(m)$ as follows:

$$D(z) = \exp \sum_{m=0}^{M} \tilde{c}(m)\tilde{z}^{-m} \tag{1}$$

where

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \qquad |\alpha| < 1, \tag{2}$$

and the gain factor of $D(z)$ is assumed to be unity. When the sampling frequency is 8kHz, the phase characteristics $\tilde{\omega}$ of the all-pass transfer function $\tilde{z}^{-1} = e^{-j\tilde{\omega}}$ for $\alpha = 0.31$ gives a good approximation to the mel frequency scale [4] based on subjective pitch evaluations.

In the mel-cepstral analysis [3], the coefficients $\tilde{c}(m)$ is determined in such a way that

$$\varepsilon = E[e^2(n)] \tag{3}$$

is minimized, where $e(n)$ is the output of the inverse filter $1/D(z)$, as shown in Fig. 1. To realize the $D(z)$, we rewrite (1) as

$$D(z) = \exp \sum_{m=0}^{M} b(m)\Phi_m(z) \tag{4}$$

0-7803-2431-5/95 $4.00

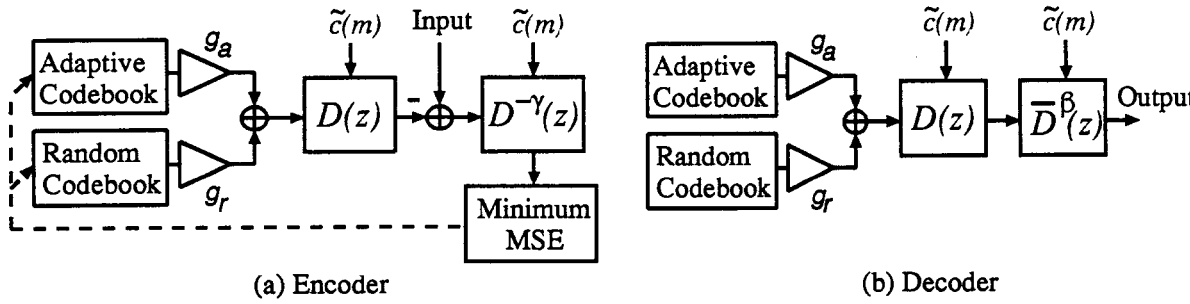


Fig. 2 CELP coder based on Mel-cepstral analysis.

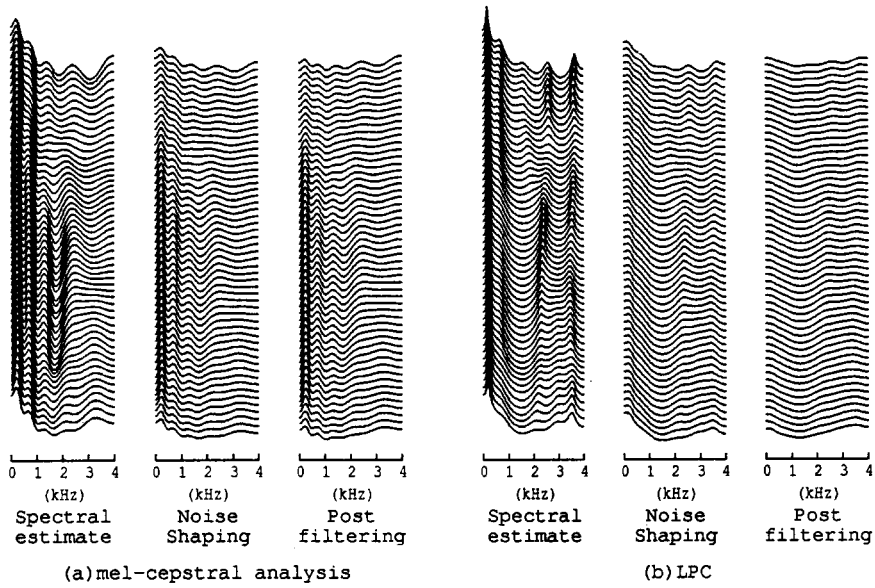


Fig. 3 Effects of noie shaping and postfiltering.

where

$$\tilde{c}(m) = \begin{cases} b(m), & m = M \\ b(m) + \alpha b(m+1), & 0 \le m < M \end{cases} \quad (5)$$

$$\Phi_m(z) = \begin{cases} 1, & m = 0 \\ \dfrac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}} \tilde{z}^{-(m-1)}, & m \ge 1. \end{cases} \quad (6)$$

It is noted that forcing the gain of $D(z)$ to be unity is equivalent to setting $b(0) = 0$. Since the transfer function $D(z)$ is theoretically minimum phase and the gain factor of $D(z)$ is normalized to unity, the impulse response of $1/D(z)$ at time 0 equals unity. As a result, the signal $e(n)$ can be viewed as the linear prediction error [5]. Therefore, instead of the linear prediction method, the mel-cepstral analysis can be used for the short-term prediction.

Although the transfer functions $D(z)$ and $1/D(z)$ are not rational functions, MLSA filters [3],[6] can approximate $D(z)$ and $1/D(z)$ with sufficient accuracy and become minimum-phase IIR systems.

3. STRUCTURE OF CODER

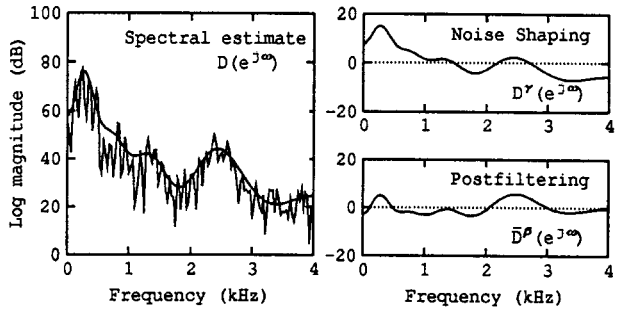
Fig. 2 shows the structure of the proposed coder. The synthesis filter $D(z)$ is realized using the MLSA filter. The transfer function $D^{-\gamma}(z)$ and $\overline{D}^{\beta}(z)$ are the perceptual weighting filter and postfilter, respectively. The transfer function $\overline{D}(z)$ is the same as $D(z)$ except that $\tilde{c}(1)$ is forced to be zero to compensate for the global spectral tilt. By setting $\tilde{c}(1) = 0$, the transfer function $\overline{D}(z)$ is written by

$$\overline{D}(z) = \exp \sum_{m=0}^{M} \overline{b}(m) \Phi_m(z) \quad (7)$$

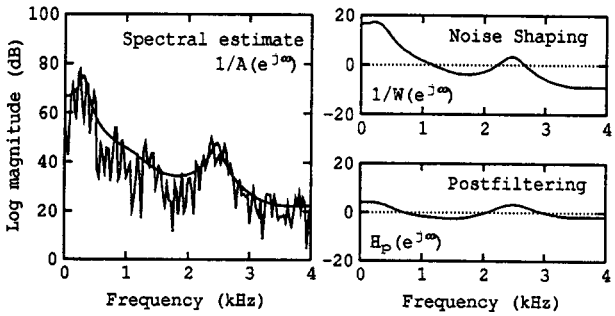
where

$$\overline{b}(m) = \begin{cases} b(m), & 2 \le m \le M \\ -\alpha b(2), & m = 1. \end{cases} \quad (8)$$

The tunable parameters γ and β control the amount of perceptual weighting and postfiltering, respectively. We can realize the perceptual weighting filter $D^{-\gamma}(z)$



(a) mel-cepstral analysis



(b) LPC

Fig. 4 Effects of noie shaping and postfiltering (for a frame extracted from Fig. 3).

and the postfilter $\overline{D}^{\beta}(z)$ in the same manner as $D(z)$ and $\overline{D}(z)$, by multiplying $\tilde{c}(m)$ by $-\gamma$ and β, respectively. To avoid large gain excursions at the postfilter output, we add the output gain control [7] which scale the postfilter output signal so that it has roughly the same power as unfiltered speech.

Fig. 3(a) and Fig. 4(a) show an example of noise shaping and postfiltering for $\gamma = \beta = 0.4$. For comparison, the effects of the conventional noise shaping and postfiltering [8],[9] are shown in Fig. 3(b) and Fig. 4(b). In Fig. 3(b) and Fig. 4(b), perceptual weighting filter $W(z)$ and postfilter $H_p(z)$ are defined by

$$W(z) = \frac{A(z/0.9)}{A(z/0.4)} \tag{9}$$

$$H_p(z) = \frac{A(z/0.5)}{A(z/0.8)}(1 - 0.5z^{-1}) \tag{10}$$

where $A(z)$ is the prediction polynomial obtained by the linear prediction method. It is noted that the perceptual weighting filter is the inverse filter of the noise shaping filter. From these figures, it is seen that the estimated speech spectrum $D(e^{j\omega})$ has high resolution at low frequencies; accordingly, spectra of noise shaping $D^{\gamma}(e^{j\omega})$ (perceptual weighting $D^{-\gamma}(e^{j\omega})$) and postfiltering $\overline{D}^{\beta}(e^{j\omega})$ also have high resolution at low frequencies. Consequently, we can expect that the effects of perceptual weighting and postfiltering fit with characteristics of the human auditory sensation and improve the perceptual performance of the coder.

Table 1: Experimental Conditions

Sampling Rate	8kHz
Subframe Period	2.5ms
Window	32ms Hamming
Analysis Order	10
Analysis Period	2.5ms

The cepstral representation has been utilized for the short-term predictor in the vector excited homomorphic vocoder [10]. Although the idea of the homomorphic vocoder is similar to ours, the proposed coder is quite different from the homomorphic vocoder. The essential differences are summarized as follows:

- The proposed coder is based on the mel-cepstral analysis, that is, frequency transformed cepstrum is used rather than the cepstrum.
- In the homomorphic vocoder, the synthesis filter is realized as a high-order FIR system, whereas the proposed coder is realized by the synthesis filter $D(z)$ using the MLSA filter, which is an IIR system.

4. PERFORMANCE EVALUATION

4.1. Experimental Conditions

To evaluate the performance of the short-term predictor only, we use a basic CELP structure without adaptive codebook. Linear prediction and mel-cepstral coefficients are updated once per every subframe and are not quantized. The random codebook gain is also not quantized. The random codebook contains 4096 codewords which overlap by a shift of 1. Other experimental conditions are provided in Table. 1.

4.2. Objective Evaluation

The objective speech quality was evaluated by the SNR and segmental SNR for 10 sentences(5 male and 5 female, 40 seconds speech). In the test, we let $\gamma = 0.4$ for the proposed coder and use the perceptual weighting filter $W(z)$ in (9) for the conventional CELP. The postfilter was not used for both coders. An improvement of 1.0 dB is achieved in SNR, and 0.8 dB in the segmental SNR.

4.3. Subjective Evaluation

The proposed CELP coder is subjectively evaluated by the equivalent Q value. In this test, we choose a parameter set $(\gamma, \beta) = (0.4, 0.4)$ for perceptual weighting and postfiltering of the proposed coder. For the conventional CELP, we use perceptual weighting filter

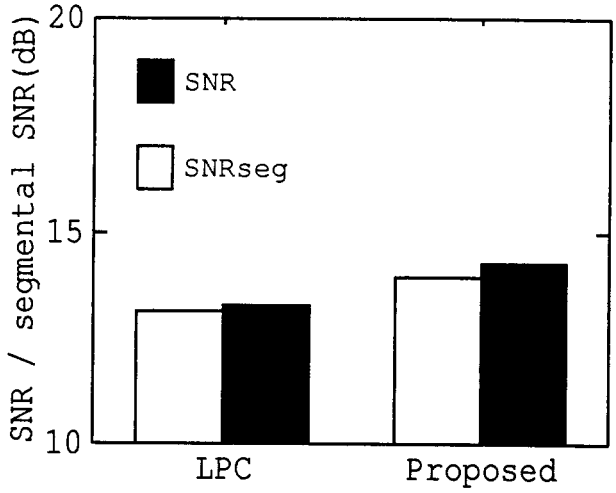


Fig. 5 Objective evaluation based on SNR and segmental SNR.

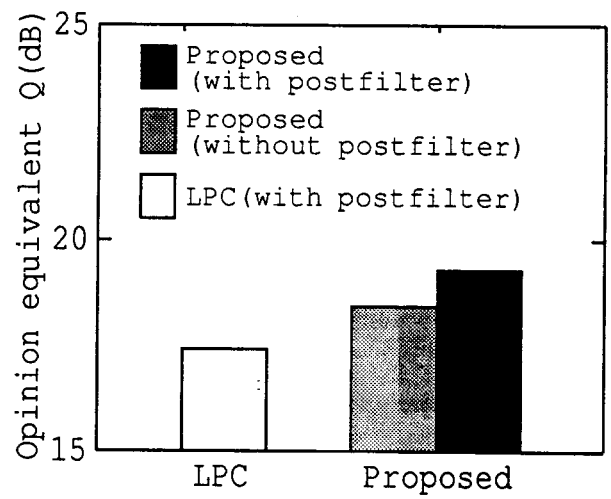


Fig. 6 Subjective evaluation based on opinion equivalent Q.

$W(z)$ and postfilter $H_p(z)$ defined by (9) and (10), respectively.

The evaluated speech data consisted of 6 sentences by 3 male and 3 female speakers. The reference signal is the original speech. Test signals are decoded speech signals by the proposed coder (with and without postfilter) and by a conventional CELP (with postfilter), and MNR signals (Q=12, 15, 18, 21, 24, 27, 30).

Fig. 6 shows the result of a speech quality evaluation based on the equivalent Q value. From Fig. 6, it is shown that the improvement of 1.0dB is achieved by the proposed coder without postfilter over the conventional CELP with postfilter. Furthermore, the improvement of more than 1.8dB is achieved by the proposed coder with postfilter over the conventional CELP. The perceptual weighting and postfiltering carried out through the mel-cepstral coefficients are effective in the CELP type speech coder.

5. CONCLUSIONS

We have proposed a CELP speech coder based on mel-cepstral analysis, in which the perceptual weighting and postfiltering are carried out through the mel-cepstral coefficients. The subjective performance test shows that the proposed coder achieves the improvement of 1.8dB over a conventional CELP. From the result, it is shown that perceptual weighting and postfiltering carried out through the mel-cepstral coefficients are effective in a low bit rate speech coding such as CELP.

Incorporation of an adaptive codebook, quantization of the mel-cepstral coefficients and codebook gain are future problems.

REFERENCES

[1] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP) : high quality speech at very low bit rates", IEEE Proc. *ICASSP-85*. pp.937–940 (1985).

[2] K. Tokuda, H. Mastumura, T. Kobayashi and S. Imai, "Speech coding based on adaptive mel-cepstral analysis," in *Proc. ICASSP-94*, 1994, pp.I-197–I-199.

[3] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech, " in *Proc. ICASSP-92*, 1992, pp.I-137–I-140.

[4] G. Fant, *Speech sound and features*. Cambridge: MIT Press, 1973.

[5] K. Tokuda, T. Kobayashi and S. Imai, "Generalized cepstral analysis of speech —unified approach to LPC and cepstral method," in *Proc. ICSLP-90*, 1990, pp.37–40.

[6] S. Imai, K. Sumita, C. Furuichi, "Mel log spectral approximation filter for speech synthesis," *Trans. IECE*, vol. J66-A, pp.122–129, Feb. 1983 (in Japanese).

[7] R. Fenichel, "Proposed federal standard 1016 (second draft)", National Communications Systems, Office of Technology and Standards, Washington, DC 20305-2010, 13 November 1989.

[8] J. H. Chen, R. V. Cox, Y. C. Lin, N. Jayant and M. J. Melchner, "A low-delay CELP coder for the CCITT 16kb/s speech coding standard," *IEEE Journal of Selected Areas in Communications*, vol. 10, pp.830-849, June 1992.

[9] J. H. Chen and A. Gersho, "Real-time vector APC speech coding at 4800bps with adaptive postfilter," in *Proc. ICASSP-87*, 1987, pp.2185–2188.

[10] J. H. Chung and R. W. Schafer, "A 4.8Kbps homomorphic vocoder using analysis-by-synthesis excitation analysis," in *Proc. ICASSP-89*, 1989, pp.144–147.