SPEECH CODING BASED ON ADAPTIVE MEL-CEPSTRAL ANALYSIS

Keiichi Tokuda[1] *Hidetoshi Matsumura*[2] *Takao Kobayashi*[2] *Satoshi Imai*[2]

[1]Department of Electrical and Electronic Engineering, Tokyo Institute of Technology, Tokyo, 152 JAPAN
[2]Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama, 227 JAPAN

ABSTRACT

In this paper, we propose an ADPCM coder which uses a backward adaptive predictor based on the adaptive mel-cepstral analysis. The spectrum represented by the mel-cepstral coefficients has frequency resolution similar to that of the human ear which has high resolution at low frequencies. In the coder, since the transfer functions of noise shaping and postfiltering are also defined through the mel-cepstral coefficients, the effects of nose shaping and postfiltering should fit with characteristics of the human auditory sensation. We incorporate a pitch predictor into the ADPCM coder, and evaluate the speech quality based on objective and subjective performance tests. It is shown that the coder at 16 kb/s can produce a high quality speech comparable with that of the CCITT G.721 ADPCM coder at 32kb/s with no algorithmic delay.

1. INTRODUCTION

Many speech coding systems have used the AR spectral representation for short-term prediction. However, in some cases, spectral zeros are important and a more general model is required. Although many techniques have been proposed for simultaneous determination of both poles and zeros, they are not always successful for a variety of reasons (e.g., stability or convergence). On the other hand, the cepstral coefficients [1] can represent spectral poles and zeros with equal weights. Furthermore, the mel-cepstrum [2] is defined as frequency-transformed cepstrum so that the spectrum represented by 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 the mel-cepstral representation can be used for efficient spectral modeling in speech coders instead of the AR modeling. Although the cepstrum has been utilized for short-term adaptive predictor in the homomorphic vocoder [3], and the frequency-transformed AR model was also applied to an ADPCM system [4], the mel-cepstrum has not been applied to speech coding.

To demonstrate the effectiveness of the mel-cepstral representation in speech coding, we propose an ADPCM coder which uses a short-term adaptive predictor based on mel-cepstral representation of speech spectrum. In the coder, the mel-cepstral coefficients are updated by the algorithm for adaptive mel-cepstral analysis [5]. Since the transfer functions of noise shaping and postfiltering are also defined through the mel-cepstral coefficients, the effects of noise shaping and postfiltering should fit with characteristics of the human auditory sensation.

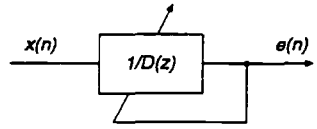

Fig. 1. Adaptive mel-cepstral analysis

The speech quality of the coder is evaluated by both objective and subjective performance tests. It is shown that a high quality speech corresponding to that of the CCITT G.721 ADPCM coder at 32kb/s can be produced by the proposed coder at 16 kb/s with no algorithmic delay.

2. ADAPTIVE MEL-CEPSTRAL ANALYSIS

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

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

where

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

For example, 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 $\alpha = 0.31$ is a good approximation to the mel frequency scale [6] based on subjective pitch evaluations.

In the mel-cepstral analysis [5], the gain factor of $D(z)$ is assumed to be unity so that the impulse response at time 0 equals unity. Then, under this condition, the coefficients $\tilde{c}(m)$ are determined to minimize

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

where $e(n)$ is the output of the inverse filter $1/D(z)$, as shown in Fig. 1. The adaptive mel-cepstral analysis [5] solves the minimization problem by an adaptive filter approach; the coefficients $\tilde{c}(m)$ are updated sample by sample based on an instantaneous estimate for the gradient of ε. It has been shown [5] that the adaptive algorithm has sufficiently fast convergence characteristics for speech analysis.

Since the transfer function $D(z)$ is theoretically minimum phase and the gain factor of $D(z)$ is normalized to unity,

0-7803-1775-0/94 $3.00

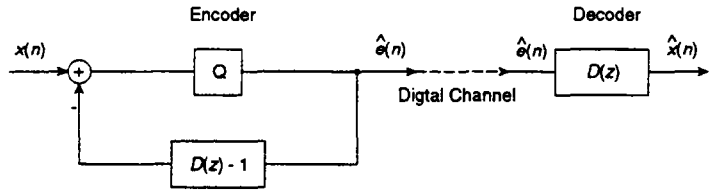

Fig. 2. Basic structure of the coder based on adaptive mel-cepstral analysis

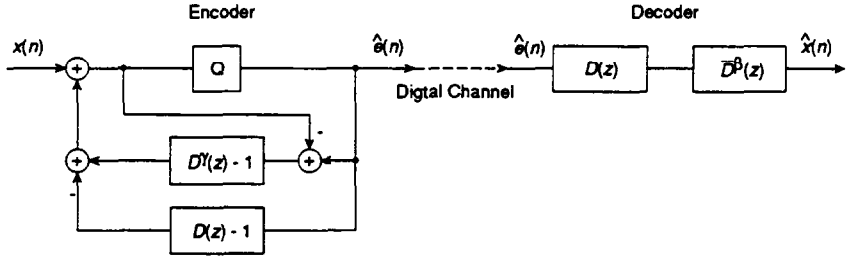

Fig. 3. Structure of the coder based on adaptive mel-cepstral analysis

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. Therefore, instead of the linear prediction method, the adaptive mel-cepstral analysis can be used for the short-term adaptive prediction.

Although the transfer functions $D(z)$ and $1/D(z)$ are not rational functions, MLSA filters [2], [5] can approximate $D(z)$ and $1/D(z)$ with sufficient accuracy and become minimum-phase IIR systems.

3. STRUCTURE OF THE CODER

3.1. Basic Structure

Fig. 2 shows the basic structure of the proposed coder. The coefficients $\tilde{c}(m)$ are updated in a backward adaptive fashion, that is, they are updated based on $\hat{e}(n)$ rather than $e(n)$. We choose a scalar adaptive quantizer proposed in [7] for Q. The z-transform of the decoded speech $\hat{x}(n)$ is

$$\hat{X}(z) = X(z) + Q(z) \tag{4}$$

where $X(z)$ and $Q(z)$ are the z-transforms of $x(n)$ and $q(n)$, respectively, and $q(n)$ is the quantization noise generated by the quantizer Q. The transfer function $D(z)$ is realized using the MLSA filter. It is noted that $D(z) - 1$ has no delay-free paths. The specific structures of $D(z)$ and $D(z) - 1$ used in this paper can be seen in [8].

3.2. Noise Shaping and Postfiltering

Fig. 3 shows the structure of the proposed coder with noise shaping and postfiltering based on the mel-cepstral analysis. The z-transform of the decoded speech $\hat{x}(n)$ is

$$\hat{X}(z) = \{X(z) + D^{\gamma}(z)Q(z)\}\,\overline{D}^{\beta}(z). \tag{5}$$

The transfer function $D^{\gamma}(z)$ shapes the noise spectrum, and $\overline{D}^{\beta}(z)$ is the postfilter. The transfer function $\overline{D}(z)$ is the same as $D(z)$ except that $c_{\gamma}(1)$ is forced to be zero to compensate for the global spectral tilt. The tunable parameters γ and β control the amount of noise shaping and postfiltering, respectively. The case where $\gamma = 0$ and $\beta = 0$ corresponds to the basic structure shown in Fig. 2. Fig. 4(a) shows an example of noise shaping and postfiltering for $\gamma = 0.3$ and $\beta = 0.2$. For comparison, characteristics of conventional noise shaping and postfiltering [9], [10] are shown in Fig. 4(b). They are defined by

$$\hat{X}(z) = \left\{X(z) + \frac{B(z/0.85)}{B(z)}Q(z)\right\}\frac{B(z/0.5)}{B(z/0.8)}(1 - 0.5\,z^{-1}) \tag{6}$$

where $B(z)$ is the prediction polynomial of z^{-1} obtained by the linear prediction method. It is seen that the estimated speech spectrum $D(e^{j\omega})$ has high resolution at low frequencies; accordingly, spectra of noise shaping $D^{\gamma}(e^{j\omega})$ and postfiltering $\overline{D}^{\beta}(e^{j\omega})$ also have high resolution at low frequencies. Thus, the effects of noise shaping and postfiltering should fit to characteristics of the human auditory sensation and can improve the perceptual performance of the coder. Informal listening tests indicate that the subjective performance when $\alpha = 0$, i.e., the cepstrum is used, is inferior to that with $\alpha = 0.31$. The result is expected since noise shaping and postfiltering are performed on linear frequency scale when $\alpha = 0$.

The transfer functions $D(z)$ and $\overline{D}(z)$ are realized using the MLSA filters. We can also realize $D^{\gamma}(z)$ and $\overline{D}^{\beta}(z)$ in the same manner as $D(z)$ and $\overline{D}(z)$, by multiplying $\tilde{c}(m)$ by γ and β, respectively. To avoid large gain excursions at the postfilter output, we add the output gain control [10] which scales the postfilter output signal so that it has roughly the same power as the unfiltered speech.

3.3. Structure with Pitch Predictor

Fig. 3 shows the structure of the coder with pitch predictor. The z-transform of the decoded speech $\hat{x}(n)$ is

$$\hat{X}(z) = \left\{X(z) + \frac{D^{\gamma}(z)}{A_n(z)}Q(z)\right\}A_p(z)\,\overline{D}^{\beta}(z). \tag{7}$$

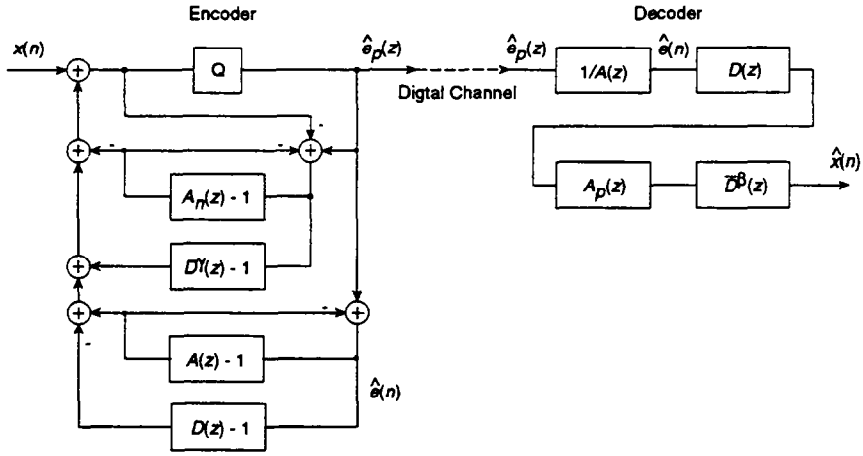

Fig. 5. Structure of the coder with pitch predictor.

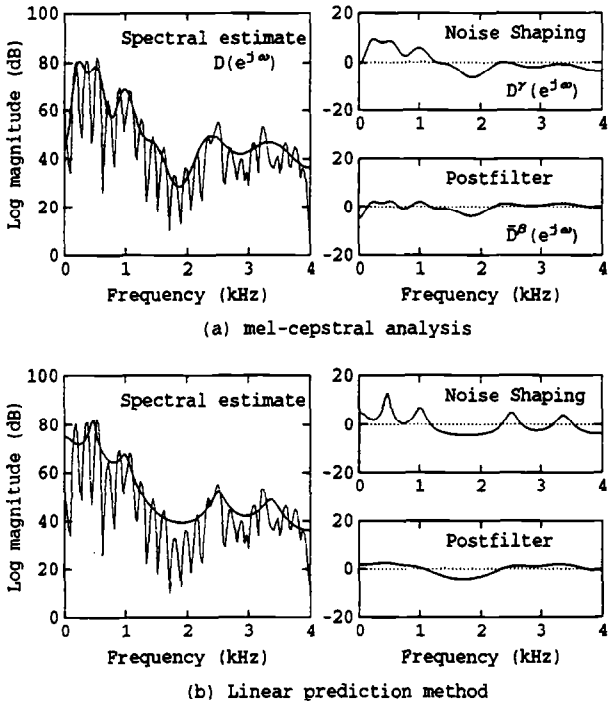

Fig. 4. Noise shaping and postfiltering.

The transfer function of the pitch prediction filter is given by

$$A(z) = 1 + \sum_{k=p-1}^{p+1} a(k)\, z^{-k}. \tag{8}$$

The pitch period p and the pitch predictor coefficients $a(k)$ are calculated from the correlation of $\hat{e}(n)$, which is obtained by using an exponential window and updated every sample. A stabilization technique [11] is applied to the coefficients $a(k)$ and they are forced to be zero at unvoiced samples detected by a simple voiced/unvoiced decision algorithm. The noise shaping and postfiltering are based on those proposed in [12], [13]: the transfer functions $A_n(z)$,

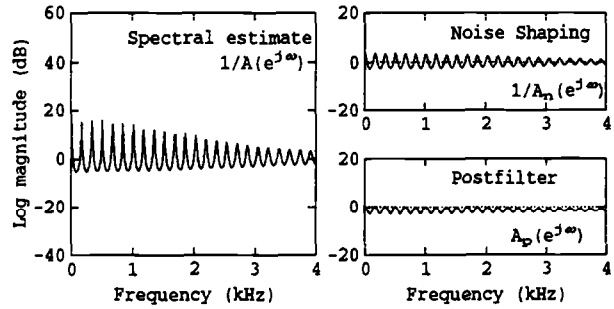

Fig. 6. Effect of noise shaping and postfiltering based on pitch predictor.

$A_p(z)$ are defined by

$$A_n(z) = 1 + \epsilon_n \sum_{k=p-1}^{p+1} a(k)\, z^{-k} \tag{9}$$

$$A_p(z) = \left(1 - \epsilon_p \sum_{k=p-1}^{p+1} a(k)\, z^{-k}\right) \Big/ \left(1 - \epsilon_p \sum_{k=p-1}^{p+1} a(k)\right), \tag{10}$$

respectively. The tunable parameters ϵ_n and ϵ_p control the amount of noise shaping and postfiltering, respectively. In the decoder, p, $a(k)$ are also calculated from the quantized residual samples $\hat{e}(n)$, i.e., the pitch predictor uses backward adaptation.

Fig. 6 illustrates effects of the noise shaping and postfiltering based on pitch prediction. Fig. 4 and Fig. 6 correspond to the same portion of the speech signal. Thus, the effects of noise shaping and postfiltering can be seen as the combination of Fig. 4(a) and Fig. 6.

4. PERFORMANCE ASSESSMENT

4.1. Objective Performance

The objective speech quality was evaluated by the segmental SNR. The proposed coder at 16kb/s was tested on 10 sentences uttered by five female and five male speakers,

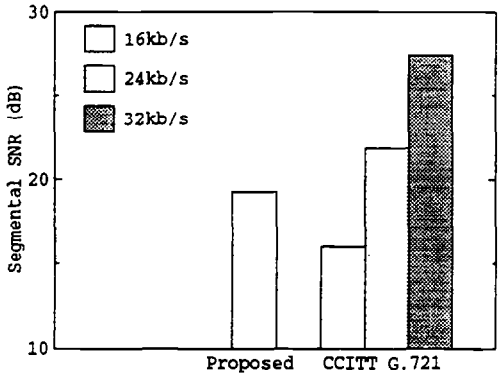

Fig. 7. Objective performance assessment based on segmental SNR.

about 40 seconds of total speech. In the test, we let (γ, β, ϵ_n, ϵ_p) = (0, 0, 0, 0). Fig. 7 shows the result. For comparison, the results of CCITT G.721 ADPCM coder at 16, 24, 32kb/s are also shown. From the figure, it is seen that the improvement of 3dB is achieved by the proposed coder at 16kb/s over the 16kb/s CCITT G.721 ADPCM coder.

4.2. Subjective Performance

Fig. 8 shows the result of a speech quality assessment based on the opinion equivalent Q. The reference signal is the original speech. Test signals are decoded speech signals by the proposed coder at 16kb/s and by the CCITT G.721 ADPCM coder at 16, 24, 32kb/s, and MNR signals ($Q = 6, 12, 18, 24, 30, 36, 42$dB). We chose a parameter set (γ, β, ϵ_n, ϵ_p) = (0.3, 0.2, 0.3, 0.1) for noise shaping and postfiltering of the proposed coder by informal listening.

From Fig. 8, it is seen that the noise shaping and postfiltering improve the equivalent Q by nearly 12dB in the proposed coder at 16kb/s. Particularly, the postfilter achieves significant noise reduction. The coder at 16kb/s without noise shaping and postfiltering achieves improvement of 5dB over the CCITT G.721 ADPCM coder at 16kb/s. Further, the improvement of more than 17dB is achieved by the proposed 16kb/s coder with noise shaping and postfiltering over the 16kb/s CCITT G.721 ADPCM coder, and the quality corresponds to that of the CCITT G.721 at 32kb/s.

5. CONCLUSIONS

We have proposed an ADPCM coder in which adaptive prediction, noise shaping, and postfiltering are carried out through the mel-cepstral coefficients. Although the coder uses a scalar quantizer rather than a vector quantizer and has no algorithmic delay, the subjective performance test shows that the coder at 16kb/s produces a high quality speech corresponding to that of the CCITT G.721 ADPCM coder at 32kb/s. In this paper, we have not considered channel errors. Coder design for noisy channels is a future problem.

To show effectiveness of mel-cepstral representation of speech spectrum in speech coding, we chose an ADPCM coder as an example. From the result, it is surmised that the mel-cepstral representation can also be effective in other types of coder, e.g., CELP coders.

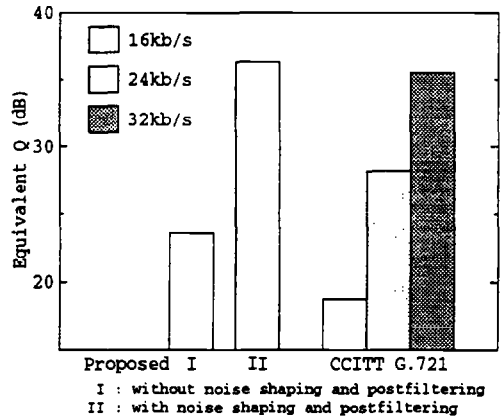

Fig. 8. Subjective performance assessment based on opinion equivalent Q.

REFERENCES

[1] A. V. Oppenheim and R. W. Schafer, *Descrete-Time Signal Processing*, Englewood Cliffs, N.J.: Prentice-Hall, 1989.

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

[3] 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.

[4] E. Krüger and H. W. Strube, "Linear prediction on a warped frequency scale," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp.1529-1531, Sep. 1988.

[5] 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.

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

[7] N. S. Jayant, "Adaptive quantization with a one-word memory," *Bell Syst. Tech. J.*, vol. 52, pp.1119–1144, Sep. 1973.

[8] K. Tokuda, H. Matsumura, T. Kobayashi and S. Imai, "Speech coding system based on adaptive mel-cepstral analysis and its evaluation," in *Proc. IEICE Technical Report*, SP93-62, Aug. 1993, pp.39–46 (in Japanese).

[9] B. A. Atal, M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp.247-254, June 1979.

[10] 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.

[11] R. P. Ramachandran and P. Kabal, "Stability and performance analysis of pitch filters in speech coders," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp.937-946, July 1987.

[12] I. A. Gerson, M. A. Jasiuk, "Techniques for improving the performance of CELP-type speech coders," *IEEE Journal of Selected Areas in Communications*, vol. 10, pp.858-865, June 1992.

[13] 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.