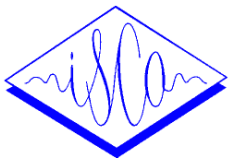

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

SPEECH CODING BASED ON ADAPTIVE MEL-CEPSTRAL ANALYSIS FOR NOISY CHANNELS

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 examine the robustness of an ADPCM coder based on adaptive mel-cepstral analysis. To improve the noisy channel performance, we use traditional techniques: the leakage factor and sign function. The subjective speech quality of the proposed 16kb/s coder and G.726 coder in terms of the opinion equivalent Q is measured and compared. It is shown that the proposed coder produces much higher quality speech than that of 16kb/s G.726 at BER(Bit Error Rate)=0 and BER=10^{-3}. Although the coder scores 4dB lower than 32kb/s G.726 at BER=0, the improvement of more than 5dB is achieved by the proposed coder over G.726 at BER=10^{-3}.

I. INTRODUCTION

In many speech coding systems, AR spectral representation has been used. Although spectral zeros are important in some cases, AR modeling cannot represent them. On the other hand, cepstral modeling [1] can represent spectral poles and zeros with equal weights. Furthermore, the spectrum represented by the mel-cepstral coefficients [2] has frequency resolution similar to that of the human ear which has high resolution at low frequencies [3]. 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 an ADPCM coder which uses a short-term adaptive predictor based on mel-cepstral representation of speech spectrum [4]. 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 effect of noise shaping and postfiltering should fit with the characteristics of the human auditory sensation. We have shown that this coder at 16kb/s can produce a high quality speech corresponding to that of CCITT G.726 ADPCM coder at 32kb/s [4]. However, we have not considered the robustness of the coder for noisy channels.

This paper examine the ADPCM coder based on adaptive mel-cepstral analysis for noisy channels. We modify the backward adaptive predictor using the traditional techniques [6]. Since it is not likely that the backward adaptive pith predictor can withstand bit errors, we do not use the pitch predictor. To compensate for the loss due to these modifications, we increase the order of mel-cepstral predictor [7].

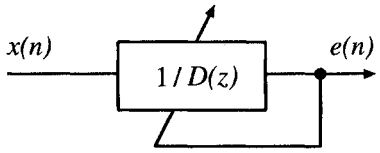

Fig. 1. Adaptive mel-cepstral analysis.

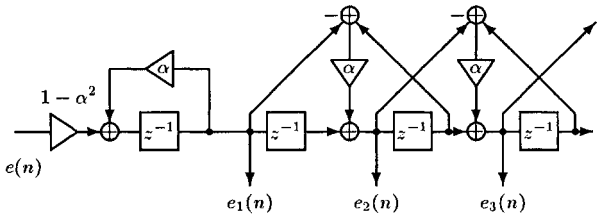

Fig. 2. Filter $\Phi_m(z)$.

The speech quality of the coder is evaluated by both objective and subjective performance tests. The subjective speech quality of the proposed 16kb/s coder and G.726 32kb/s coder in terms of the opinion equivalent Q is measured and compared. Although the coder scores 4dB lower than G.726 at BER=0, the improvement of more than 5dB is achieved by the proposed coder over G.726 at BER=10^{-3}. Particularly, it is shown that the coder achieves significant noise reduction by both noise shaping and postfiltering at not only clear channels but also noisy channels.

II. ADAPTIVE MEL-CEPSTRAL ANALYSIS

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

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

where

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

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 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, we determine the

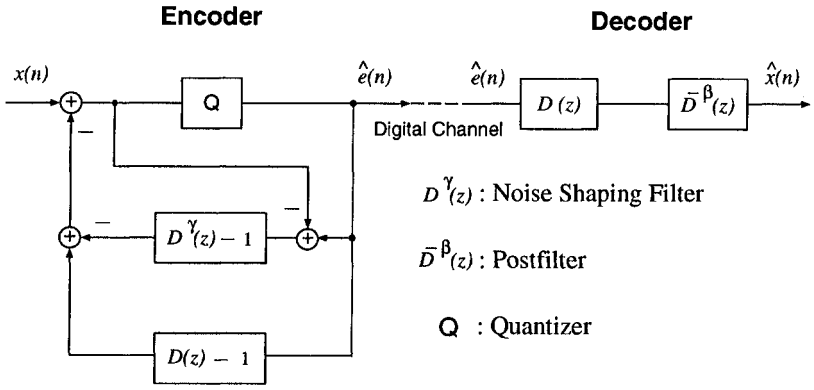

Fig. 3. Structure of the coder based on adaptive mel-cepstral analysis.

coefficients $\tilde{c}(m)$ 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

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

where

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

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

where the filter $\Phi_m(z)$ is shown in Fig. 2. It is noted that forcing the gain of $D(z)$ to be unity is equivalent to set $b(0) = 0$. The adaptive mel-cepatral analysis [5] solves the minimization problem by an adaptive filter approach; the coefficients $\boldsymbol{b} = [b(1), b(2), \cdots, b(M)]^T$ are updated sample by sample based on an instantaneous estimate for the gradient of ε. It has been shown that the adaptive algorithm has sufficiently fast convergence characteristics for speech analysis [5].

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

The adaptive algorithm [5] for updating the coefficient $\boldsymbol{b}$ from time n to $n+1$ is given by

$$\boldsymbol{b}^{(n+1)} = \boldsymbol{b}^{(n)} - \mu^{(n)}\hat{\nabla}\varepsilon_\tau^{(n)} \tag{7}$$

where

$$\mu^{(n)} = \frac{a}{M\varepsilon^{(n)}} \tag{8}$$

$$\varepsilon^{(n)} = \lambda\varepsilon^{(n-1)} + (1-\lambda)e^2(n) \tag{9}$$

$$\hat{\nabla}\varepsilon_\tau^{(n)} = \tau\hat{\nabla}\varepsilon_\tau^{(n-1)} - 2(1-\tau)e(n)\boldsymbol{e}_\Phi^{(n)} \tag{10}$$

$$\boldsymbol{e}_\Phi^{(n)} = [e_1(n), e_2(n), \cdots, e_M(n)]^T \tag{11}$$

and $0 < \lambda, a < 1$, $0 \le \tau < 1$. The m-th element of $\boldsymbol{e}_\Phi^{(n)}$, $e_m(n)$, is the output of the filter $\Phi_m(z)$ shown in Fig. 2 driven by $e(n)$.

Although the transfer functions $D(z)$ and $1/D(z)$ are not rational functions, MLSA filters [5],[9] can approximate $D(z)$ and $1/D(z)$ with sufficient accuracy and become minimum-phase IIR systems.

III. STRUCTURE OF THE CODER

Fig. 3 shows the structure of the proposed coder with noise shaping and postfiltering based on the mel-cepstral analysis. We choose a scalar adaptive quantizer used in CCITT G.726 at 16kb/s [6] for the quantizer Q. The coefficients $\boldsymbol{b}$ are updated in a backward adaptive fashion, that is, they are updated based on $\hat{e}(n)$ rather than $e(n)$. The z-transform of the decoded speech $\hat{x}(n)$ is

$$\hat{X}(z) = \{X(z) + D^\gamma(z)Q(z)\}\overline{D}^\beta(z) \tag{12}$$

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 $D^\gamma(z)$ shapes the noise spectrum, and $\overline{D}^\beta(z)$ is a postfilter. They have different forms from the conventional noise shaping and postfiltering based on the linear prediction method [10],[11]. The tunable parameters γ and β control the amount of noise shaping and postfiltering, respectively. To compensate for the global spectral tilt, we force $\tilde{c}(1)$ of $\overline{D}(z)$ to zero. By setting $\tilde{c}(1) = 0$, the transfer function $\overline{D}(z)$ is given by

$$\overline{D}(z) = \exp \sum_{m=0}^{M} \overline{b}(m)\Phi_m(z) \tag{13}$$

where

$$\overline{b}(m) = \begin{cases} b(m), & 2 \le m \le M \\ -\alpha b(2), & m = 1. \end{cases} \tag{14}$$

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 $b(m)$ and $\overline{b}(m)$ by γ and β, respectively. To avoid large gain excursions at the postfilter output, we add the output gain control [11] which scales the postfilter output signal so that it has roughly the same power as unfiltered speech.

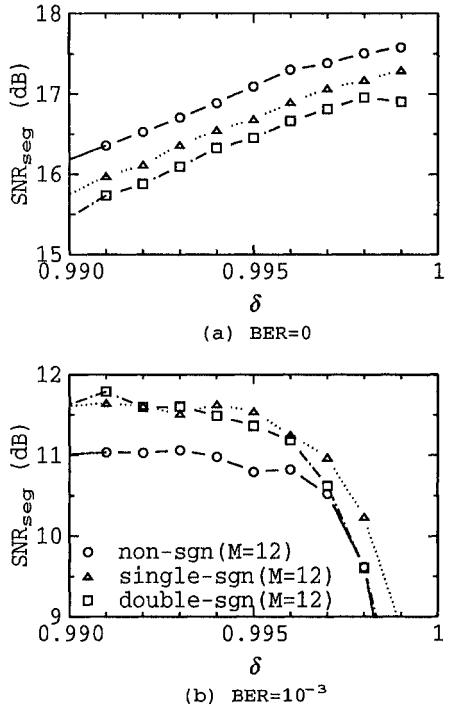

Fig. 4. Performance and robustness of the coder in terms of the leakage factor δ and sign function.

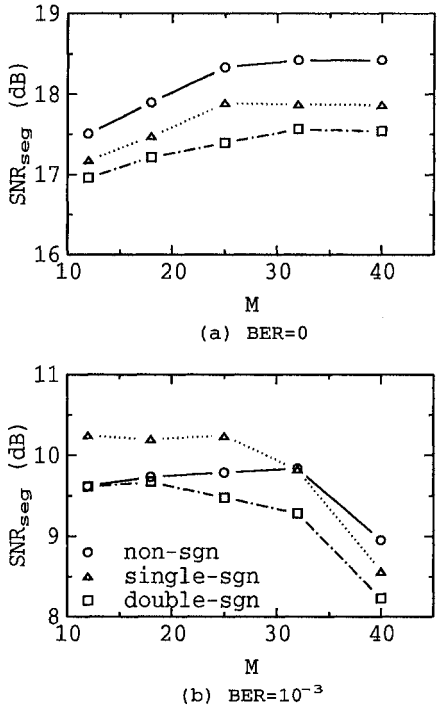

Fig. 5. Performance and robustness of the coder in terms of the order M ($\delta = 0.998$).

IV. PREDICTOR DESIGN FOR NOISY CHANNELS

We modify the adaptive algorithm in a similar manner of the G.726 ADPCM coder [6]. Incorporating a leakage factor δ in (7), we have

$$\boldsymbol{b}^{(n+1)} = \delta \boldsymbol{b}^{(n)} - \mu^{(n)} \hat{\nabla} \varepsilon_\tau^{(n)} \tag{15}$$

where $0 < \delta < 1$. The leakage factor δ allows the coder to forget past value of the coefficients. Without the leakage factor ($\delta = 1$), it is impossible for the receiver to recover from only one transmission error. To reduce sensitivity of channel errors, we examine

$$\hat{\nabla} \varepsilon_\tau^{(n)} = \tau \hat{\nabla} \varepsilon_\tau^{(n-1)} - 2(1-\tau)\,\mathrm{sgn}[e(n)] \boldsymbol{e}_\Phi^{(n)} \tag{16}$$

$$\hat{\nabla} \varepsilon_\tau^{(n)} = \tau \hat{\nabla} \varepsilon_\tau^{(n-1)} - 2(1-\tau)\,\mathrm{sgn}[e(n)]\,\mathrm{sgn}[\boldsymbol{e}_\Phi^{(n)}] \tag{17}$$

where sgn[ ] is the sign function defined by

$$\mathrm{sgn}[x] = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0. \end{cases} \tag{18}$$

Fig. 4 shows the performance and robustness of the coder in terms of the leakage factor δ and sign function at BER=0 and BER=10^{-3}. In Fig. 4, 'non-sgn', 'single-sgn' and 'double-sgn' mean the equation (10), (16) and (17), respectively. As δ decreases, the time constant for recovery from the channel errors decreases, that is, the robustness at noisy channels increases. Similarly, using the sign function improve the robustness. However, techniques which are used to improve the noisy channel performance usually result in decreased performance at a noise-free channels. Such performance/robustness tradeoffs are often encountered in adaptive system design.

We have proposed a 16kb/s ADPCM coder with pitch predictor [4], which can produce a high quality speech corresponding to that of the G.726 at 32kb/s. Since backward-adaptive pitch predictor is very sensitive to channel errors, we do not use pitch predictor. To compensate for the loss due to these modifications, we increase the order of the short-term predictor in the same manner as the LD-CELP coder [7]. Fig. 5 shows the performance and robustness of the coder in terms of the order M. In Fig. 5(a), the performance at noise-free channels saturates around order 32. In Fig. 5(b), there is a knee in the performance curves around order 32. Therefore, we use the order of $M = 32$.

Informal listening tests were performed in order to determine δ and the use of the sign function. In this test, it has been shown that the postfilter can improve the performance of the coder at both noise-free and noisy channels. To increase postfilter's performance, the coefficients $\boldsymbol{b}$ need to be updated as exactly as possible. Therefore, the leakage factor should be chosen to be very close to 1 and the sign function is not necessary. As a result, We let $\delta = 0.998$ and eliminate the sign function.

V. PERFORMANCE ASSESSMENT

We evaluated the objective speech quality by the segmental SNR. The proposed coder at 16kb/s was tested on 10 sentences uttered by five female and five male speakers, about 40 seconds of total speech. In the test, we let $(\gamma, \beta) = (0, 0)$, $(\delta, \alpha, \lambda, M, \tau, a)$=(0.998, 0.31, 0.98, 32, 0.95, 0.15). Fig. 6 shows the results. For comparison, the results

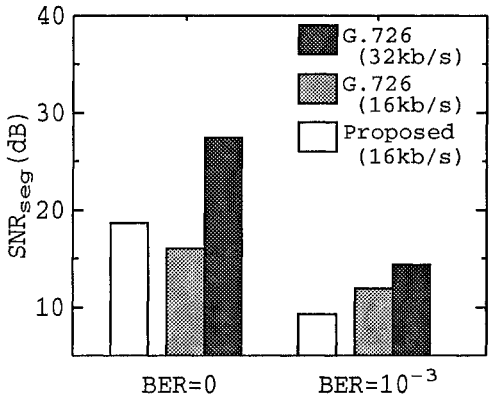

Fig. 6. Objective performance assessment based on segmental SNR.

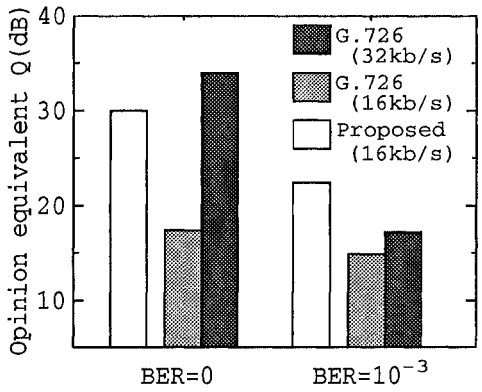

Fig. 7. Subjective performance assessment based on opinion equivalent Q.

of CCITT G.726 ADPCM coder at 16 and 32kb/s are also shown.

Fig. 7 shows the results of a speech quality assessment based on the opinion equivalent Q. The reference signal is the original speech. Test signals are decoded speech signal by the proposed coder at 16kb/s and by the CCITT G.726 ADPCM coder at 16 and 32kb/s, and MNR signals (Q=12, 18, 24, 30, 36, 42). We choose a parameter set (γ,β)=(0.2, 0.2) for noise shaping and postfiltering of the proposed coder based on informal listening tests. Other parameters are the same as that in objective assessment.

From Fig. 6, it is shown that the coder scores lower than 16kb/s G.726 at BER=10^{-3} because δ is very close to 1 and the sign function is not used. However, from Fig. 7, it is seen that the noise shaping and postfiltering improve the quality of decoded speech. The coder produces much higher quality speech than that of 16kb/s G.726 at noise-free and noisy channels. Although the coder scores 4dB lower than 32kb/s G.726 at BER=0, the improvement of more than 5dB is achieved by the proposed coder over G.726 at BER=10^{-3}. Particularly, the postfilter achieves significant noise reduction at not only noise-free but also noisy channels. Instead of a loss of the robustness at BER=10^{-3} , we might obtain the better performance at noise-free channels.

For noisy channels at bit error rates of 10^{-2}, since we have to use the more inaccurate coefficients $\boldsymbol{b}$ than that at BER=10^{-3}, the postfilter is not likely to work. By using a small value of the leakage factor and the sign function, we could improve the performance at a high bit error rate of 10^{-2} at the expense of the degradation of noise-free performance.

VI. CONCLUSION

We have examined an ADPCM coder based on adaptive mel-cepstral analysis for noisy channels. In the coder, 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 has shown that the coder at 16kb/s is quite robust compared with 32kb/s G.726 in the presence of channel errors(BER=10^{-3}). It has been also shown that the coder produces much higher quality speech than that of 16kb/s G.726.

To show effectiveness of mel-cepstral representation of speech spectrum in speech coding, we have chosen an ADPCM coder as an example. From the obtained results, it is surmised that the mel-cepstral representation can also be effective in CELP coder. To design CELP coder which uses mel-cepstral analysis will be presented at the next opportunity.

REFERENCES

[1] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Trans. on Audio and Electroacoustics*, AU-16, 2, pp.221–226, June. 1968.

[2] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. ICASSP-83*, 1983, pp.93–96.

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

[4] 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.

[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] CCITT Study Group XV, "Draft Recommendations G.726 and G.727," (Question 24/XV), August 1990.

[7] 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.

[8] 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.

[9] 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).

[10] 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.

[11] 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.