PAPER

A low-bit-rate audio codec using mel-scaled linear predictive analysis

Yoshihisa Nakatoh[1,*] and Hiroshi Matsumoto[2]

[1]*Matsushita Electric Industrial Co., Ltd.,*
1006 Kadoma, Kadoma, 571–8501 Japan
[2]*Faculty of Engineering, Shinshu University,*
4–17–1 Wakasato, Nagano, 380–0922 Japan

(*Received 31 May 2006, Accepted for publication 8 November 2006*)

Abstract: In this paper, we propose a low-bit-rate audio codec using a new analysis method named mel-scaled linear predictive analysis (mel-LP analysis). In mel-LP analysis, a spectral envelope is estimated on a mel- or bark-frequency scale, so as to improve the spectral resolution in the low-frequency band. This analysis is accomplished with about a twofold increase in computation over standard LPC analysis. Our codec using mel-LP analysis consists of five key parts: time frequency transformation, flattening of MDCT coefficients using the mel-LP spectral envelope, power normalization, perceptual weighting estimation, and multistage VQ. In subjective experiments, we investigated the performance of our codec using the mel-LP analysis method, through 7-level paired comparison tests. The result shows that the codec using the mel-LP analysis method results in a good performance at a low bit rate, particularly at 16 kbps. In the cases of pop songs, piano music and male speech, sound quality was improved.

Keywords: Audio codec, Low bit rate, LPC analysis, Spectral resolution, VQ

PACS number: 43.60.Ek [doi:10.1250/ast.28.147]

1. INTRODUCTION

In the last few years, a significant reduction in bit rate has been demanded for wideband digital audio signal transmission and storage. Several audio-coding algorithms, such as MPEG1 [1], TwinVQ [2] and AAC [3] have been proposed. TwinVQ is a superior high quality audio codec based on vector quantization (VQ). In TwinVQ, the modified discrete cosine transform (MDCT) coefficients calculated from the input audio signal are flattened and divided using the LPC spectral envelope on the frequency domain, and the flattened MDCT coefficients are quantized using interleaved VQ. On the other hand, in a speech codec such as CELP [4], LPC analysis is used to flatten the spectrum of the input signal.

In this paper, we describe a low-bit-rate audio codec using a new analysis method named mel-LP analysis. Usually, speech signal production is modeled by an autoregressive process. However, in the low-frequency band, spectral resolution using standard LPC analysis is insufficient, because in LPC analysis the spectral resolution is equal at all frequency bands. Because many parameters are required to represent a spectral envelope well, the bit rate cannot be reduced. A linear prediction analysis on a mel- or bark-frequency scale proposed by Strube [5] is expected to be effective in an audio codec, because the spectral envelope obtained by Strube's method has frequency-resolution-like auditory characteristics. However, since Strube's method has a high computational cost, we proposed a mel-LP analysis method [6,7], and applied it to an audio codec [8]. In mel-LP analysis, a spectral envelope is estimated on the mel- or bark-frequency scale, as in Strube's method. Our method is computationally simple (about a twofold increase in computation compared with the standard LPC analysis), and the stability of the system is guaranteed.

In the following section, we describe our codec using the mel-LP analysis. The outstanding features of our codec are the flattening of the MDCT coefficients using the mel-LP spectral envelope on the frequency domain, and block-selective interleaved multistage VQ with perceptual weighting estimation. In subjective experiments, we investigated the performance of our proposed codec using mel-LP analysis, through paired comparison tests between mel-LP analysis and standard LPC analysis.

2. CODING SYSTEM

The block diagram of our codec is illustrated in Fig. 1.

[*] e-mail: nakatoh.yoshihisa@jp.panasonic.com

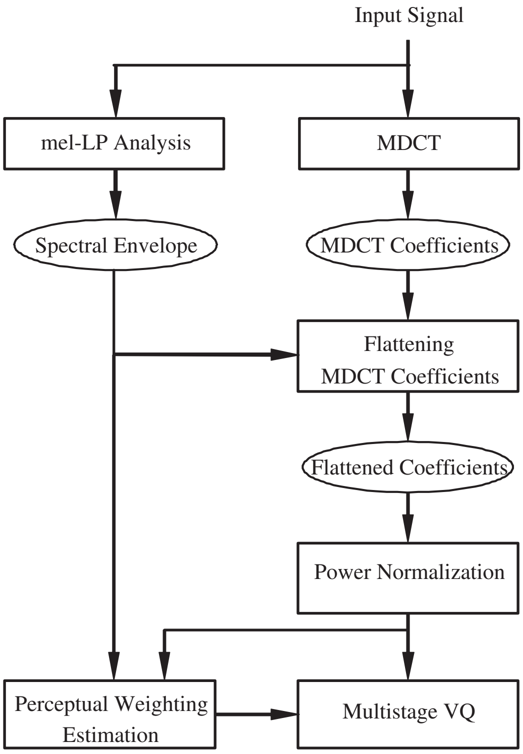


Fig. 1 Block diagram of proposed codec.

The encoder consists of the following five key parts:

1) time frequency transformation (MDCT),
2) flattening of MDCT coefficients by the mel-LP spectral envelope,
3) power normalization,
4) perceptual weighting estimation, and
5) multistage vector quantization (VQ).

First, the input signal is transformed into the MDCT coefficients in the MDCT block. The MDCT coefficients are flattened using the spectral envelope obtained by mel-LP analysis. The flattened MDCT coefficients are normalized by dividing by the spectral element with maximum power in the analysis frame. These power-normalized MDCT coefficients are quantized using the block-selective interleaved multistage VQ. In the second and third VQ stages, block selection is executed adaptively to find the optimal-frequency region to minimize the quantization error.

2.1. mel-LP Analysis

The basic idea of all-pole modeling on the warped frequency scale was proposed by Strube [5], whose method is expected to be effective in speech and audio codecs, because the spectral envelope obtained by Strube's method has frequency-resolution-like auditory characteristics. However, this method has rarely been used in coding applications due to its relatively high computational load compared with standard LPC analysis. In this paper, we propose a simple and efficient time-domain technique (mel-LP analysis) to directly estimate warped predictors from input speech, and we apply this analysis method to an audio codec. In this study, we use a warped inverse filter on the linear frequency axis, unlike in Strube's method,

$$A_w(z) = \tilde{A}_w(\tilde{z}) = \sum_{n=0}^{p} \tilde{a}_{w,n}\tilde{z}^{-n}, \tag{1}$$

where

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}}. \tag{2}$$

$A_w(z)$ is an inverse filter on the linear-frequency axis and $\tilde{A}_w(\tilde{z})$ is a warped inverse filter on the mel-frequency axis. $\{\tilde{a}_{w,k}\}$ are warped predictors and p is the analysis order. $\tilde{z}^{-1}$ is the first-order all-pass filter and α is the mel-scale factor determined by the phase characteristic of the all-pass filter. For a windowed input signal segment $x[0], \ldots, x[N-1]$, the error energy $\tilde{\sigma}_w{}^2$ is given by Eq. (3), and the warped predictors are estimated so as to minimize the error power over an infinite time interval using the following equation:

$$\tilde{\sigma}_w{}^2 = \sum_{n=0}^{\infty} \left\{ \sum_{k=0}^{p} \tilde{a}_{w,k} \cdot y_k[n] \right\}^2, \tag{3}$$

where $y_k[n]$ is the output sequence of the kth order all-pass filter. It should be noted that the estimated predictors $\{\tilde{a}_{w,k}\}$ are different from the predictors $\{\tilde{a}_k\}$ defined in Strube's method. The warped predictors are obtained to solve the following normal equation:

$$\sum_{j=1}^{p} \phi(i,j)\tilde{a}_{w,j} = -\phi(i,0) \quad (i = 1, \ldots, p), \tag{4}$$

where the coefficient $\phi(i,j)$ is given by

$$\phi(i,j) = \sum_{n=0}^{\infty} y_i[n]y_j[n] \tag{5}$$

and

$$y_0[n] = x[n]. \tag{6}$$

In terms of Parseval's theorem, $\phi(i,j)$ can be proved to be equal to the autocorrelation function $\tilde{r}_w[i-j]$, of which its Fourier transform is equal to the warped and frequency-weighted power spectrum, $|\tilde{X}(e^{j\tilde{\lambda}}) \cdot \tilde{W}(e^{j\tilde{\lambda}})|^2$ as

$$\begin{aligned}\phi(i,j) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |\tilde{X}(e^{j\tilde{\lambda}}) \cdot \tilde{W}(e^{j\tilde{\lambda}})|^2 \cos(i-j)\tilde{\lambda}d\tilde{\lambda} \\ &= \tilde{r}_w[i-j]. \end{aligned} \tag{7}$$

$\tilde{X}(e^{j\tilde{\lambda}})$ is the spectral envelope in the mel- or bark-frequency domain. The weighting function $\tilde{W}(e^{j\tilde{\lambda}})$ is the frequency derivative of the phase transfer function of $\tilde{z}^{-1}$ and is given by

$$\tilde{W}(\tilde{z}) = \frac{\sqrt{1-\alpha^2}}{1+\alpha\cdot\tilde{z}^{-1}}. \tag{8}$$

Therefore, Eq. (4) becomes an autocorrelation equation, as in standard LPC analysis, and the estimated spectrum $\tilde{\sigma}_w/\tilde{A}_w(\tilde{z})$ represents the envelope of $\tilde{X}(e^{j\tilde{\lambda}})\cdot\tilde{W}(e^{j\tilde{\lambda}})$.

If necessary, the effect of the weighting function $\tilde{W}(\tilde{z})$ on the estimated spectrum can be completely compensated by filtering $\tilde{r}_w[i-j]$ using the second-order FIR filter, $[\tilde{W}(\tilde{z})\cdot\tilde{W}(\tilde{z}^{-1})]^{-1}$. The warped autocorrelation coefficients $\{\tilde{r}[i-j]\}$ are given by

$$\begin{aligned}\tilde{r}[i-j] = \beta_0\tilde{r}_w[i-j] + \beta_1\{\tilde{r}_w[i-j-1]\\ + \tilde{r}_w[i-j+1]\},\end{aligned} \tag{9}$$

where

$$\beta_0 = \frac{1+\alpha^2}{\sqrt{1-\alpha^2}} \quad \text{and} \quad \beta_1 = \frac{\alpha}{\sqrt{1-\alpha^2}}. \tag{10}$$

The resultant warped autocorrelation coefficients $\{\tilde{r}[i-j]\}$ lead to the same warped predictors $\{\tilde{a}_k\}$.

Furthermore, since $\phi(i,j)$ is a function of the difference $|i-j|$, $\phi(i,j)$ becomes equal to the sum of the following finite series without any approximation:

$$\phi(i,j) = \tilde{r}_w[i-j] = \sum_{n=0}^{N-1} x[n]\cdot y_{(i-j)}[n], \tag{11}$$

where the output sequence $y_k[n]$ is given by

$$\begin{aligned}y_k[n] = \alpha\cdot(y_k[n-1] - y_{(k-1)}[n]) + y_{(k-1)}[n-1]\\ (n = 0,\ldots,N-1, k = 1,\ldots,p).\end{aligned} \tag{12}$$

Therefore, in addition to requiring the computational load to obtain the autocorrelation coefficients in standard LPC analysis, computational load is required to generate the output signal of the multistage all-pass filter. Figure 2 is a block diagram showing the generation of the output signal of the all-pass filter. In the processing, $p*N$ multiplications and $2*p*N$ additions and subtractions are needed. Therefore, because of the cost of computing N points of $y_k[n]$, mel-LP analysis is accomplished with about a twofold increase in computation compared with standard LPC analysis. This computational load is much lower than those of both the "autocorrelation method" and the "covariance method" by Strube [5]. In the autocorrelation method, to calculate the autocorrelation coefficients of the warped signal in the mel-frequency domain, the resampling of the power spectra and bilinear conversion in the autocorrelation domain are required. In the covariance method, there is a high computational complexity of the covariance process. Therefore, the mel-LP analysis method that we propose can be analyzed with low computational complexity in the time domain.

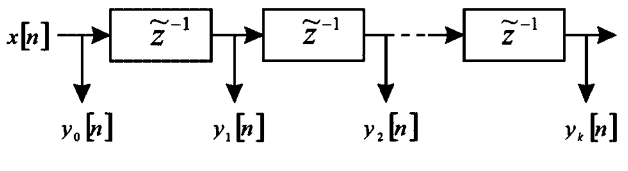


Fig. 2 Generation of output signal in all-pass filter.

2.2. Flattening of MDCT Coefficients

The input signal is transformed into MDCT coefficients by adaptive-block-size MDCT [9] and the MDCT coefficients are flattened using the spectral envelope obtained by mel-LP analysis. However, the MDCT coefficients are represented in the linear-frequency domain. On the other hand, the spectral envelope obtained by mel-LP analysis is represented in the mel- or bark-frequency domain. Therefore, the spectral envelope $\tilde{H}(e^{j\tilde{\lambda}})$ in the mel- or bark-frequency domain has to be transformed into the spectral envelope $H(e^{j\lambda})$ in the linear-frequency domain. The spectral envelope $\tilde{H}(e^{j\tilde{\lambda}})$ in the mel- or bark-frequency domain is given by

$$\tilde{H}(e^{j\tilde{\lambda}}) = \frac{\tilde{\sigma}_w{}^2}{|\tilde{A}_w(e^{j\tilde{\lambda}})\cdot\tilde{W}(e^{j\tilde{\lambda}})|^2}. \tag{13}$$

The spectral envelope $H(i)$ $(i = 1,\cdots,M)$, which is a descrete representation of $H(e^{j\lambda})$, is obtained by resampling the spectral envelope $\tilde{H}(e^{j\tilde{\lambda}})$ in the linear-frequency domain, where M is the number of MDCT components. The flattened MDCT coefficients $\hat{S}(i)$ are obtained by dividing the MDCT coefficients $S(i)$ by the spectral envelope $H(i)$ using the following equation:

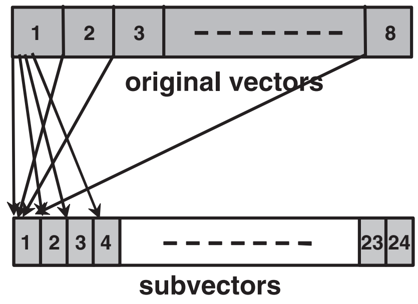
$$\hat{S}(i) = \frac{S(i)}{H(i)} \quad (i = 1,\cdots,M). \tag{14}$$

In addition, the above process, which flattens the MDCT coefficients using the spectral envelope in the frequency domain, is equivalent to inverse filtering in the time domain.

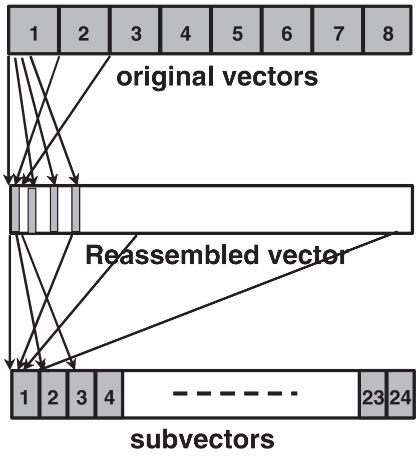
Next, the flattened MDCT coefficients are normalized by dividing by the spectral element with maximum power in the analysis frame. The power-normalized MDCT coefficients are then interleaved. Figure 3 shows interleaving in the cases of a long block size (1,024) and a short block size (128). When we apply interleaving to the window of a long block size, the MDCT coefficients are interleaved [2] and split into subvectors composed of 24 elements. When the window of a short block size is interleaved, the MDCT coefficients of 8 short blocks are reassembled in order from low to high frequency. The reassembled vector coefficients are interleaved into the multiple subvectors.

2.3. Multistage Vector Quantization

The interleaved vectors are quantized by block-selective multistage (3 stages) vector quantization. Figure 4 shows the block diagram of multistage VQ with perceptual



(a) long block size (1,024 points).



(b) short block size (128∗8 points).

Fig. 3 Interleaving of MDCT coefficients.

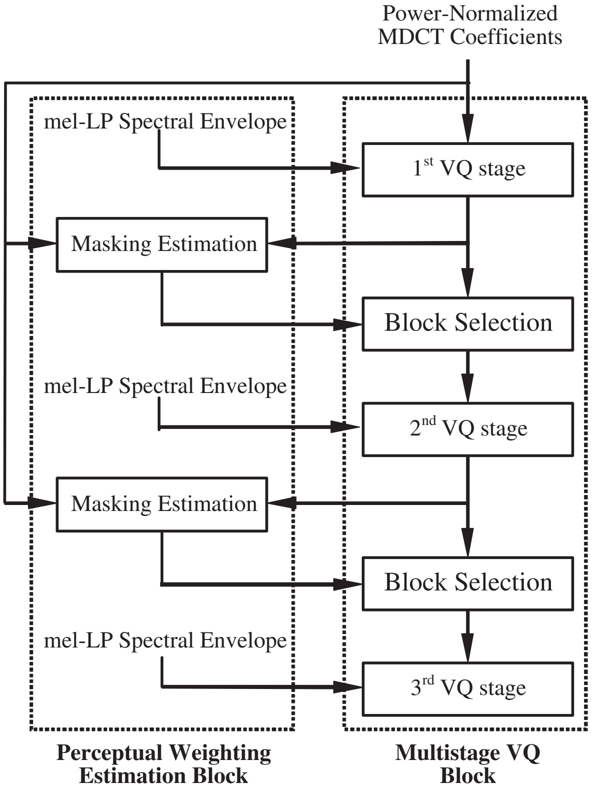


Fig. 4 Block diagram of multistage VQ.

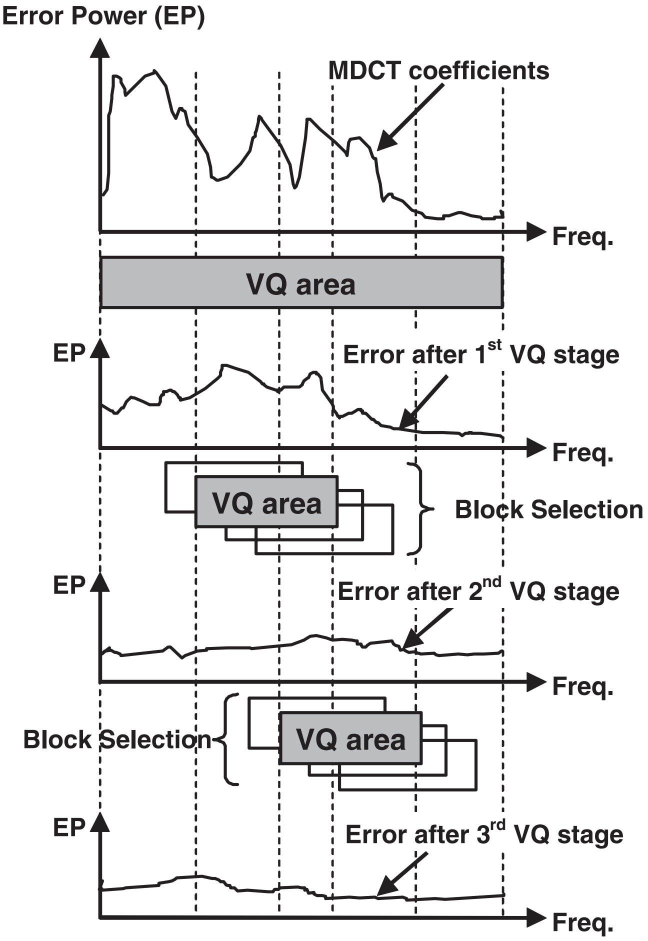


Fig. 5 Block selection in multistage VQ.

weighting estimation. In the first VQ stage, the interleaved vector is vector-quantized on the basis of weighted distance using the mel-LP spectral envelope [2]. Similarly, vector quantization based on the weighted distance is performed at the other VQ stage. Figure 5 shows block selection in multistage VQ. The horizontal axis represents frequency and the vertical axis represents error power. Block selection is applied before processing in the second and third VQ stages. Block selection is executed adaptively to find the optimal frequency area (VQ area) by minimizing the sum of quantization errors at a constant band width. The sum is calculated by weighting the masking curve. In the perceptual weighting estimation, the masking curve is determined by weighting the original masking curve, calculated from the input signal, by the correlation coefficients between the quantization error power and the power-normalized MDCT coefficients. The method of determining the masking curve is similar to the psycho-acoustic model of MPEG1.

3. EXPERIMENTS

3.1. Experimental Conditions

In experiments, we used 5 kinds of audio samples, pop

Table 1 Specification of the mel-LP or LPC analysis.

Analysis window length	2,048 points
Analysis order p	10
Mel scale factor α	0.65

Table 2 Bit allocations at each bit rate.

Coding parameter	bits/frame (1,024 points)			
LSP	32			
# of window	4			
Power	32			
Shift point	$3 * 2$			
1st VQ	$(4+4)*32$	$(6+4)*40$	$(8+5)*40$	$(8+7)*40$
2nd VQ	$(3+0)*8$	$(5+0)*8$	$(6+3)*8$	$(8+7)*8$
3rd VQ	$(2+0)*8$	$(5+0)*8$	$(6+3)*8$	$(8+7)*8$
Total bits (Bit rate)	378 (16 kbps)	554 (24 kbps)	738 (32 kbps)	914 (40 kbps)

song, harpsichord, piano, triangle and male speech. The specifications of mel-LP or standard LPC analysis are shown in Table 1. The analysis order p was determined on the basis of restrictions on the bit stream when we proposed our codec to the MPEG4 standard. The mel scale factor α was calculated using the phase characteristic of the all-pass filter given by Eq. (2). The value of α was determined as 0.65 which can approximate the mel-frequency scale well at 44.1 kHz [10]. We used 4 bit rates, 16, 24, 32 and 40 kbps. The sampling frequency was 44.1 kHz. At 16 kbps, the decoded sounds had a frequency bandwidth of 16 kHz and at other bit rates, the decoded sounds had a frequency bandwidth of 20.7 kHz. The MDCT window length was long (2,048 points) or short (256 points). The number of window is 4. The bit allocations in each bit rate are shown in Table 2. In the bit stream, the mel-LP (or LPC) coefficients are converted to mel-LSP (or LSP) parameters. The bits for a shift point represent the position of the selected band in the block.

3.2. Subjective Experiments

In subjective experiments, we investigated the performance of our codec using the mel-LP analysis method, through 7-level paired comparison tests. In this experiment, we first produced a pair of decoded sounds flattened using the mel-LP spectral envelope or the LPC envelope in our codec. We presented it to 8 listeners (including acoustic specialists) in both orders: (1) LPC then mel-LP and (2) mel-LP then LPC. Next, all listeners rated the comparative sound quality on 7-level scale after listening to a pair of sounds. Table 3 shows the conditions in the subjective experiments. The test question is "How good is the former as compared with the latter regarding sound quality?"

Table 3 Conditions in subjective evaluations.

Quality evaluation of 7-level scale	+3: much better +2: better +1: slightly better 0: the same −1: slightly worse −2: worse −3: much worse
Listeners	8 persons (including acoustic specialists)
Presentation order	(1) LPC then mel-LP (2) mel-LP then LPC

Table 4 Preference scores for mel-LP analysis in comparison with standard LPC analysis (average of all listeners).

Audio Sample	bit rate			
	16 kbps	24 kbps	32 kbps	40 kbps
Pop Song	1.00 ± 0.3	0.36 ± 0.4	0.36 ± 0.5	0.21 ± 0.4
Harpsichord	0.29 ± 0.3	0.14 ± 0.4	-0.29 ± 0.5	0.29 ± 0.4
Piano	0.79 ± 0.4	0.21 ± 0.4	0.57 ± 0.5	0.57 ± 0.4
Triangle	0.36 ± 0.4	0.43 ± 0.4	-0.14 ± 0.5	0.29 ± 0.4
Male speech	0.79 ± 0.3	0.36 ± 0.4	0.07 ± 0.4	0.14 ± 0.3

Table 4 shows the preference score for the mel-LP analysis method in comparison with the standard LPC analysis method. 95% confidence level is 0.3 to 0.5. The result shows that the codec using mel-LP analysis has a good performance, at low bit rates, particularly 16 kbps. In 16 kbps, the scores for pop song, piano and male speech are 1.00, 0.79 and 0.79, respectively. On the other hand, the score for triangle is 0.43 at 24 kbps. The scores of piano at 32 kbps and 40 kbps are both 0.57. There are differences in the preference score among the five audio materials. In the cases of pop song, piano and male speech, sound quality has been significantly improved. On the other hand, for triangle and harpsichord, the effectiveness is slightly less.

Furthermore, in the preliminary experiment, we evaluated the difference in the sound quality between our basic codec (when LPC analysis is used) and other codecs (MPEG1-Layer3, TwinVQ). The performance of our basic codec exceeded that of MPEG1-Layer3 below at 64 kbps. On the other hand, at equivalent performance was obtained in comparison with TwinVQ. In particular, it is reported by another research that block-selective interleaved multistage VQ shows a good performance [11].

4. DISCUSSION

Using the mel-LP method, for the cases of pop song, piano and male speech, sound quality was improved. On

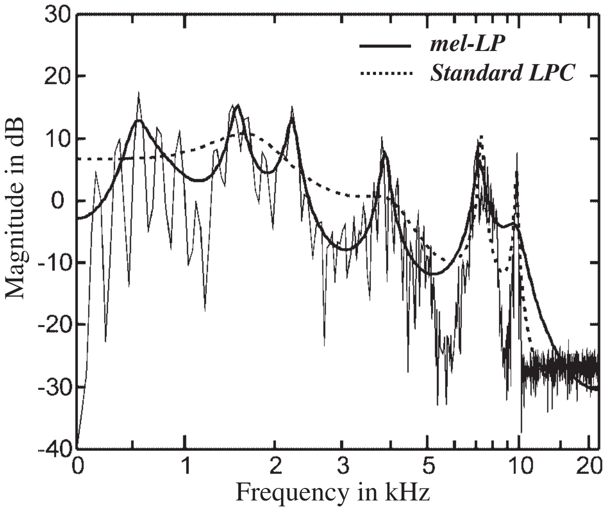


Fig. 6 Comparison of the mel-LP spectral envelope with that of the standard LPC for male speech at analysis order of $p = 16$.

the other hand, for triangle and harpsichord, the method was less effective. We observed the difference in the spectrum of each evaluation data. Consequently, in the former, a spectral peak with a high power existed in the low-frequency band. On the other hand, in the latter, a spectral peak with high power existed over a wide frequency range irrespective of the frequency band. The mel-LP analysis is excellent for the expression of a spectral peak in a low-frequency band, which is an important auditory effect. Therefore, for sound sources in which peaks exist in the low-frequency band, it is effective. On the other hand, for sound sources in which peaks exist over a wide frequency range, its effectiveness is slightly less.

We compared the proposed mel-LP spectral envelope with that of the standard LPC to investigate the performance of mel-LP analysis. Figure 6 shows the spectral envelopes of male speech at the analysis order of $p = 16$. In this figure, the solid line is the mel-LP spectral envelope and the dotted line is that of the standard LPC. The FFT spectrum is also illustrated in this figure. The horizontal axis represents the mel-frequency scale to clarify the detail of the spectrum at the low-frequency band. It is clear that the spectrum obtained by mel-LP analysis is much superior to the spectrum obtained by LPC analysis. The difference is particularly marked at the low frequency band under 5 kHz. When the analysis order is 10 ($p = 10$), the tendency is even more marked.

Although we did not show this as an experiment result, when the analysis order was about 20, the sound quality obtained by mel-LP analysis was superior to that obtained by LPC analysis. However, for sound quality, the difference between mel-LP analysis and LPC analysis became small as the analysis order became large. The same tendency was reported by Strube [5].

5. CONCLUSION

We proposed a low-bit-rate audio codec using mel-LP analysis. The outstanding features of our codec are the mel-LP spectral envelope and the block-selective interleaved multistage VQ with perceptual weighting estimation. Mel-LP analysis is a method of estimating the spectral envelope on a mel- or bark-frequency scale, and in this paper, we proposed a simple and efficient time-domain technique to directly estimate warped predictors from an input speech. In subjective experiments, we investigated the performance of our codec using mel-LP analysis, through 7-level paired comparison tests. The result shows that the codec using mel-LP analysis results in a good performance at low bit rates, particularly at 16 kbps. In the cases of pop song, piano and male Speech, sound quality was improved. On the other hand, for triangle and harpsichord, the effectiveness was slightly less. In the future, we aim to improve our coding algorithm and apply it to wideband speech coding.

REFERENCES

[1] K. Brandenburg and G. Stoll, "The ISO/MPEG-1 audio codec: A generic standard for coding of high quality digital audio," *J. Audio Eng. Soc.*, **42**, 780–792 (1994).

[2] N. Iwakami, T. Moriya and S. Miki, "High-quality audio coding at less than 64 kbit/s by using TwinVQ," *Proc. ICASSP 95*, Vol. 5, pp. 3095–3098 (1995).

[3] K. Brandenburg and M. Bosi, "Overview of MPEG audio: Current and future standards for low-bit rate audio coding," *J. Audio Eng. Soc.*, **45**, 4–21 (1997).

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

[5] H. W. Strube, "Linear prediction on a warped frequency scale," *J. Acoust. Soc. Am.*, **68**, 1071–1076 (1980).

[6] H. Matsumoto, Y. Nakatoh and Y. Furuhata, "An efficient Mel-LPC analysis method for speech recognition," *Proc. ICSLP 98*, pp. 1051–1054 (1998).

[7] S. Nakagawa, M. Okada and T. Kawahara, *Spoken Language Systems* (Ohmsha, Ltd., Tokyo, 2005), Chap. 7.

[8] Y. Nakatoh, T. Norimatsu, A. H. Low and H. Matsumoto, "Low bit rate coding for speech and audio using mel linear predictive coding (Mel-LPC) analysis," *Proc. ICSLP 98*, Vol. 6, pp. 2591–2594 (1998).

[9] M. Iwadare, A. Sugiyama, F. Hazu, A. Hirano and T. Nishitai, "A 128 kb/s Hi-Fi audio codec based on adaptive transform coding with adaptive block size MDCT," *IEEE J. Sel. Areas Commun.*, **10**, 138–144 (1992).

[10] S. Imai, T. Kitamura and H. Takeya, "A direct approximation technique of log magnitude response for digital filters," *IEEE Trans. Acoust. Speech and Signal Process.*, **ASSP-25**, 127–133 (1977).

[11] M. Tsushima, T. Ishikawa, T. Norimatsu, N. Tanaka and K. Yoshida, "Proposal of bitrate scalability for TwinVQ based core and comparison with NTT scalable coder," ISO/IEC JTC1/SC29/WG11 MPEG97/M2058 (1997).