AN ADAPTIVE ALGORITHM FOR MEL-CEPSTRAL ANALYSIS OF SPEECH

Toshiaki Fukada†, *Keiichi Tokuda*††, *Takao Kobayashi*†††*and Satoshi Imai*†††

†Information Systems Research Center, Canon Inc., Kawasaki, 211 JAPAN (toshiaki@cis.canon.co.jp)
††Department of Electrical and Electronic Engineering, Tokyo Institute of Technology, Tokyo, 152 JAPAN
†††Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama, 227 JAPAN

ABSTRACT

This paper describes a mel-cepstral analysis method and its adaptive algorithm. In the proposed method, we apply the criterion used in the unbiased estimation of log spectrum to the spectral model represented by the mel-cepstral coefficients. To solve the non-linear minimization problem involved in the method, we give an iterative algorithm whose convergence is guaranteed. Furthermore, we derive an adaptive algorithm for the mel-cepstral analysis by introducing an instantaneous estimate for gradient of the criterion. The adaptive mel-cepstral analysis system is implemented with an IIR adaptive filter which has an exponential transfer function, and whose stability is guaranteed. We also present examples of speech analysis and results of an isolated word recognition experiment.

1. INTRODUCTION

The spectrum represented by the mel-cepstral coefficients have frequency resolution similar to that of the human ear which has high resolution at low frequencies[1]. As a result, mel-cepstral coefficients are useful for speech synthesis and recognition. For obtaining mel-cepstral coefficients, several methods have been proposed. For example, the mel-cepstral coefficients are obtained from the LPC coefficients by using the technique of spectral resampling. No strict method, however, is proposed in which the spectral model is represented by mel-cepstral coefficients and a spectral criterion is minimized.

In this paper, we propose a mel-cepstral analysis method and its adaptive algorithm. In the mel-cepstral analysis method, the model spectrum is represented by the M-th order mel-cepstral coefficients and the criterion used in the unbiased estimation of log spectrum[2] is minimized with respect to the mel-cepstral coefficients. The minimization problem is solved efficiently by an iterative technique using the FFT, recursion formulas, and a fast algorithm that requires $O(M^2)$ arithmetic operations. We can show that the convergence is quadratic and typically a few iterations are sufficient to obtain the solution.

Furthermore, we present an adaptive algorithm for the mel-cepstral analysis. To derive the adaptive algorithm, we introduce an instantaneous estimate for the gradient of the criterion in a similar manner of the LMS algorithm[3]. The adaptive analysis system is implemented with an IIR adaptive filter which has the structure of the MLSA filter [4] and whose stability is guaranteed. The adaptive analysis system requires $O(M)$ operations per sample to implement the M-th order mel-cepstral analysis. We show examples of analysis for synthetic and speech signal. To evaluate the proposed methods, an isolated word recognition experiment was carried out.

2. SPECTRAL ESTIMATION BASED ON MEL-CEPSTRAL REPRESENTATION

2.1 Spectral Model and Criterion

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

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

where

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

The phase characteristic of the all-pass transfer function $\tilde{z}^{-1} = e^{-j\tilde{\omega}}$ is given by

$$\tilde{\omega} = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha}. \tag{3}$$

For example, for a sampling frequency of 10kHz, $\tilde{\omega}$ is a good approximation to the mel scale based on subjective pitch evaluations when $\alpha = 0.35$[4].

To obtain an unbiased estimate, we use the following criterion[2] and minimize it with respect to $\{\tilde{c}(m)\}_{m=0}^{M}$.

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} \{\exp R(\omega) - R(\omega) - 1\}\, d\omega \tag{4}$$

where

$$R(\omega) = \log I_N(\omega) - \log \left| H(e^{j\omega}) \right|^2 \tag{5}$$

and $I_N(\omega)$ is the modified periodogram of a weakly stationary process $x(n)$ with a time window of length N. To take the gain factor K outside from $H(z)$, we rewrite (1) as

$$H(z) = \exp \sum_{m=0}^{M} b(m)\, \Phi_m(z) = K \cdot D(z) \tag{6}$$

where

$$K = \exp b(0) \tag{7}$$

$$D(z) = \exp \sum_{m=1}^{M} b(m)\, \Phi_m(z) \tag{8}$$

and

$$\tilde{c}(m) = \begin{cases} b(m), & m = M \\ b(m) + \alpha b(m+1), & 0 \le m < M \end{cases} \tag{9}$$

$$\Phi_m(z) = \begin{cases} 1, & m = 0 \\ \dfrac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}}\, \tilde{z}^{-(m-1)}, & m \ge 1. \end{cases} \tag{10}$$

Since $H(z)$ is a minimum phase system, we can show that the minimization of E with respect to $\{\tilde{c}(m)\}_{m=0}^{M}$ is equivalent to that of

$$\varepsilon = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{I_N(\omega)}{|D(e^{j\omega})|^2}\, d\omega \tag{11}$$

0-7803-0532-9/92 $3.00

with respect to

$$\mathbf{b} = [b(1), b(2), \ldots, b(M)]^T. \quad (12)$$

The gain factor K that minimizes E is obtained by setting $\partial E/\partial K = 0$:

$$K = \sqrt{\varepsilon_{min}} \quad (13)$$

where ε_{min} is the minimized value of ε.

2.2 Solution for the Minimization Problem

Since ε is convex with respect to $\mathbf{b}$ [5], the minimization problem of (11) can be solved by the Newton-Raphson method. For the i-th result $\mathbf{b}^{(i)}$, solving a set of linear equations

$$\mathbf{H}\,\Delta\mathbf{b}^{(i)} = -\nabla\varepsilon \Big|_{\mathbf{b}=\mathbf{b}^{(i)}}, \quad (14)$$

we have the values

$$\Delta\mathbf{b}^{(i)} = [\Delta b^{(i)}(1), \Delta b^{(i)}(2), \ldots, \Delta b^{(i)}(M)]^T \quad (15)$$

where $\mathbf{H}$ is the Hessian matrix $\mathbf{H} = \partial^2\varepsilon/\partial\mathbf{b}\,\partial\mathbf{b}^T$. Then the next result is obtained as follows:

$$\mathbf{b}^{(i+1)} = \mathbf{b}^{(i)} + \Delta\mathbf{b}^{(i)}. \quad (16)$$

The gradient $\nabla\varepsilon$ is given by

$$\nabla\varepsilon = -2\tilde{\mathbf{r}} = -2[\tilde{r}(1), \tilde{r}(2), \ldots, \tilde{r}(M)]^T \quad (17)$$

where

$$\tilde{r}(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{I_N(\omega)}{|D(e^{j\omega})|^2}\,\Phi_m^*(e^{j\omega})\,d\omega \quad (18)$$

and the Hessian matrix $\mathbf{H}$ is given by

$$\mathbf{H} = 2\{h(i,j)\}_{i,j=1}^{M} \quad (19)$$

where

$$h(i,j) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{I_N(\omega)}{|D(e^{j\omega})|^2}\left\{\Phi_i(e^{j\omega}) + \Phi_i^*(e^{j\omega})\right\}\Phi_j^*(e^{j\omega})\,d\omega. \quad (20)$$

Since the matrix $\mathbf{H}$ is a symmetric Toeplitz plus Hankel matrix, (14) can be solved using a fast recursive algorithm [6] which requires $O(M^2)$ arithmetic operations. Coefficients $h(i,j)$ and $\tilde{r}(m)$ can be calculated efficiently using the FFT and recursion formulas[5]. We can obtain an initial guess $\mathbf{b}^{(0)}$ from the FFT cepstrum using a recursion formula[5]. The convergence is quadratic because the Hessian matrix is positive definite. We have found that typically a few iterations are sufficient to obtain the solution.

3. ADAPTIVE MEL-CEPSTRAL ANALYSIS ALGORITHM

3.1 Derivation of Adaptive Algorithm

Replacing $\mathbf{H}$ in (14) with the unit matrix, we can derive the method of steepest descent from the Newton-Raphson method. That is, from the i-th result $\mathbf{b}^{(i)}$ the next result $\mathbf{b}^{(i+1)}$ is given by

$$\mathbf{b}^{(i+1)} = \mathbf{b}^{(i)} - \mu\,\nabla\varepsilon \Big|_{\mathbf{b}=\mathbf{b}^{(i)}} \quad (21)$$

where μ is the adaptation step size.

Assuming that the time window length N is sufficiently large, we can interpret (11) as the mean square of $e(n)$:

$$\varepsilon = E\left[e^2(n)\right] \quad (22)$$

where $e(n)$ is the output of the inverse filter $1/D(z)$ driven by $x(n)$. According to the assumption, the gradient given in (17) becomes

$$\nabla\varepsilon = -2\cdot E\left[e(n)\,\mathbf{e}_\Phi^{(n)}\right] \quad (23)$$

where

$$\mathbf{e}_\Phi^{(n)} = [e_1(n), e_2(n), \ldots, e_M(n)]^T \quad (24)$$

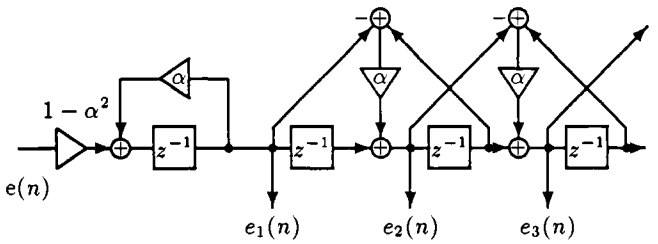


Fig. 1. Filter $\Phi_m(z)$.

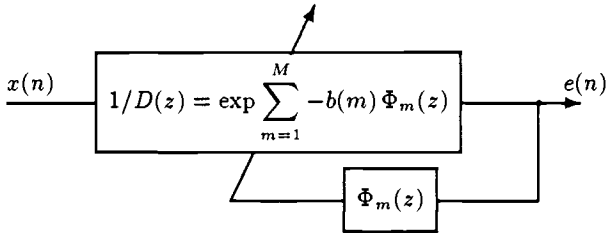


Fig. 2. Block diagram of the adaptive mel-cepstral analysis.

and $e_m(n)$ is the output of the filter $\Phi_m(z)$ (see Fig. 1).

To derive an adaptive algorithm, we introduce an instantaneous estimate in a similar manner of the LMS algorithm[3]

$$\hat{\nabla}\varepsilon^{(n)} = -2\,e(n)\,\mathbf{e}_\Phi^{(n)}. \quad (25)$$

In this paper, to suppress fluctuation of $\mathbf{b}$, we estimate $\nabla\varepsilon$ using an exponential window as follows:

$$\begin{aligned}\bar{\nabla}\varepsilon^{(n)} &= -2(1-\tau)\sum_{i=-\infty}^{n}\tau^{n-i}\,e(i)\,\mathbf{e}_\Phi^{(i)} \\ &= \tau\bar{\nabla}\varepsilon^{(n-1)} - 2(1-\tau)e(n)\mathbf{e}_\Phi^{(n)}, \quad 0\le\tau<1. \end{aligned} \quad (26)$$

With this estimate of the gradient, we can specify an adaptive algorithm based on the method of steepest descent: the coefficients vector $\mathbf{b}^{(n)}$ at time n is updated as

$$\mathbf{b}^{(n+1)} = \mathbf{b}^{(n)} - \mu^{(n)}\,\bar{\nabla}\varepsilon^{(n)}. \quad (27)$$

When the gain of the signal $x(n)$ is time-varying, μ is normalized as follows:

$$\mu^{(n)} = \frac{a}{M\,\varepsilon^{(n)}}, \qquad 0<a<1 \quad (28)$$

where $\varepsilon^{(n)}$ is an estimate of ε at time n

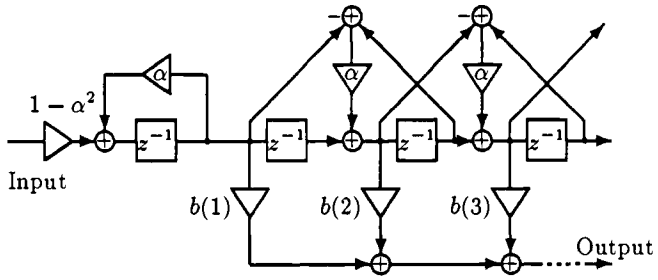
$$\begin{aligned}\varepsilon^{(n)} &= (1-\lambda)\sum_{i=-\infty}^{n}\lambda^{n-i}\,e^2(i) \\ &= \lambda\varepsilon^{(n-1)} + (1-\lambda)\,e^2(n), \qquad 0\le\lambda<1. \end{aligned} \quad (29)$$

Using (13), we can get an estimate of K at time n from $\varepsilon^{(n)}$. The block diagram of the adaptive mel-cepstral analysis is depicted in Fig. 2. In the next section, we will discuss a realization method of the exponential transfer function $1/D(z)$.

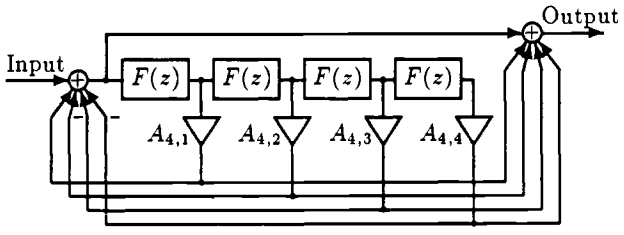
The mel-cepstral coefficients $\{\tilde{c}(m)\}_{m=0}^{M}$ can be obtained from K and $\mathbf{b}$ using (7) and (9). Note that the above algorithm is equivalent to the adaptive cepstral analysis algorithm[7] when $\alpha = 0$.

3.2 Realization of Exponential Transfer Function

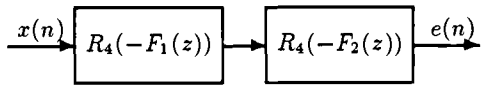
Although the transfer function $1/D(z)$ is not a rational function, the MLSA filter[4] can approximate $1/D(z)$ in Fig. 2 with sufficient accuracy. The complex exponential



(a) Basic filter $F(z)$



(b) $R_L(-F(z)) \simeq 1/D(z) \quad L=4$



(c) Two-stage cascade structure
$R_4(-F_1(z)) \cdot R_4(-F_2(z)) \simeq 1/D(z)$

Fig. 3. Implementation of $1/D(z)$.

TABLE I Coefficients of $R_4(w)$.

l	$A_{4,l}$
1	4.999273×10^{-1}
2	1.067005×10^{-1}
3	1.170221×10^{-2}
4	5.656279×10^{-4}

function $\exp w$ is approximated by a rational function

$$\exp w \simeq R_L(w) = \frac{1+\sum_{l=1}^{L} A_{L,l}\, w^l}{1+\sum_{l=1}^{L} A_{L,l}\,(-w)^l}. \tag{30}$$

Thus $1/D(z)$ is approximated as follows:

$$R_L(-F(z)) \simeq \exp(-F(z)) = 1/D(z) \tag{31}$$

where $F(z)$ is defined by

$$F(z) = \sum_{m=1}^{M} b(m)\Phi_m(z). \tag{32}$$

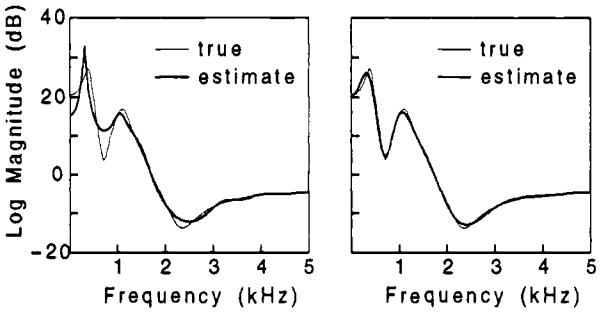
The filter structure of $F(z)$ is shown in Fig. 3(a). Figure 3(b) shows the block diagram of the MLSA filter $R_L(-F(z))$ for the case of $L=4$.

When we use the coefficients $A_{4,l}$ shown in Table I, $R_4(-F(z))$ is stable and becomes a minimum phase system under the condition

$$\left| F(e^{j\omega}) \right| \leq 6.2. \tag{33}$$

Furthermore, we can show that the approximation error $\left| \log 1/D(e^{j\omega}) - \log R_4(-F(e^{j\omega})) \right|$ does not exceed 0.24dB[8] under the condition

$$\left| F(e^{j\omega}) \right| \leq 4.5. \tag{34}$$



(a) LPC analysis (b) Mel-cepstral analysis

Fig. 4. Spectral estimates for synthetic signal($M = 12$).

When $F(z)$ is expressed as

$$F(z) = F_1(z) + F_2(z) \tag{35}$$

the exponential transfer function is approximated in a cascade form

$$\begin{aligned} 1/D(z) &= \exp(-F(z)) = \exp(-F_1(z)) \cdot \exp(-F_2(z)) \\ &\simeq R_L(-F_1(z)) \cdot R_L(-F_2(z)) \end{aligned} \tag{36}$$

as shown in Fig. 3(c). If

$$\max_{\omega} \left| F_1(e^{j\omega}) \right|, \max_{\omega} \left| F_2(e^{j\omega}) \right| < \max_{\omega} \left| F(e^{j\omega}) \right|, \tag{37}$$

it is expected that $R_L(-F_1(e^{j\omega})) \cdot R_L(-F_2(e^{j\omega}))$ approximates $1/D(e^{j\omega})$ more accurately than $R_L(-F(e^{j\omega}))$.

In the following experiments, we let

$$F_1(z) = b(1)\,\Phi_1(z) \tag{38}$$

$$F_2(z) = \sum_{m=2}^{M} b(m)\,\Phi_m(z). \tag{39}$$

Since we empirically found that

$$\max_{\omega} \left| F_1(e^{j\omega}) \right|, \max_{\omega} \left| F_2(e^{j\omega}) \right| < 4.5 \tag{40}$$

for speech sounds, $R_L(-F_1(z)) \cdot R_L(-F_2(z))$ approximates the exponential transfer function $1/D(z)$ with sufficient accuracy and becomes a stable system.

To implement the M-th order adaptive mel-cepstral analysis, the analysis system requires $O(M)$ operations per sample. Thus it can be implemented with one currently available DSP.

4. EXPERIMENTAL RESULTS

In the following experiments, α was set to 0.35 and M was set to 12 except that it was set to 15 in **4.3**. In the adaptive algorithm, a, λ, and τ were set to 0.12, 0.98, and 0.92, respectively.

4.1 Analysis of Synthetic Signal

Figure 4 shows the spectral estimates for synthetic signal compared with the LPC analysis. The signal was generated by driving an ARMA filter by a pulse train with unit variance. It is seen that mel-cepstral analysis can estimate resonances and anti-resonances which the LPC analysis can not estimate accurately.

Figure 5(a) shows the convergence characteristics of the adaptive mel-cepstral analysis for the same synthetic signal. The spectral estimates obtained from the mel-cepstral coefficients at the 400th and the 800th iterations are given in Figs. 5(b) and (c), respectively. From Fig. 5, it is seen that the adaptive mel-cepstral analysis has fast and stable convergence characteristics.

4.2 Analysis of Natural Speech

Figure 6 shows the result of a natural speech analysis. The signal shown in Fig. 6(a) is a natural Japanese speech

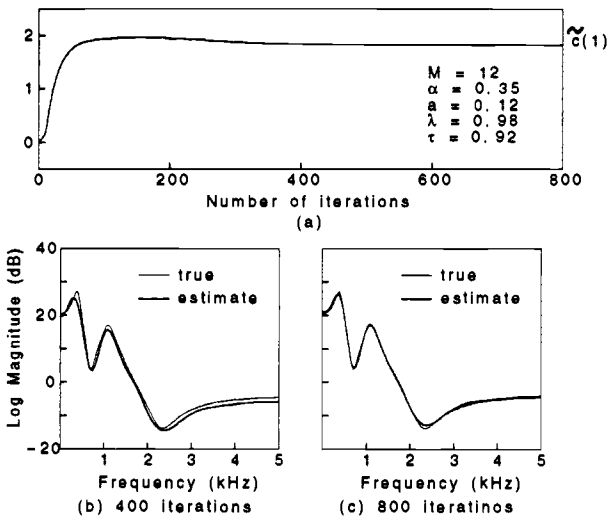


Fig. 5. Convergence characteristics of the adaptive mel-cepstral analysis algorithm.

/naNbudewa/ sampled at 10kHz. Figure 6(b) shows the mel-cepstral coefficient $\tilde{c}(1)$ versus iteration number. Log magnitude spectra shown in Figs. 6(c) and (d) are obtained from the mel-cepstral coefficients $\{\tilde{c}(m)\}_{m=0}^{M}$ at intervals of 5ms. From Fig. 6, it is seen that the proposed analysis methods have high resolution at low frequencies in spite of a small order of analysis($M = 12$), and the adaptive algorithm has sufficiently fast convergence characteristics for speech analysis.

4.3 Application to Speech Recognition

To evaluate the performance of the mel-cepstral analysis method(MC) and the adaptive mel-cepstral analysis method(AMC), an isolated word recognition experiment was carried out. We selected a vocabulary of highly confusing 20 Japanese city names uttered twice by five male speakers. The recognition was accomplished using the DTW(Dynamic Time Warping) and the Euclidean mel-cepstral distance calculated from the mel-cepstral coefficients at intervals of 10ms. For comparison, we used the mel-cepstral coefficients($M = 15$) calculated from the LPC coefficients($M = 12$) using the recursion formula(LPC). Table II shows the results for the recognition rates. We achieved higher recognition rates using MC and AMC than that using LPC.

5. CONCLUSION

In this paper, we have presented the mel-cepstral analysis method and its adaptive algorithm. The mel-cepstral analysis is efficient for the estimation of spectra which have resonances and anti-resonances at low frequencies. The adaptive mel-cepstral analysis system is implemented with an IIR adaptive filter which has an exponential transfer function and whose stability is guaranteed. It is shown that the adaptive algorithm requires $O(M)$ operations per sample to obtain the M-th order mel-cepstral coefficients and has fast convergence properties. The effectiveness of the proposed methods was also shown by the some experimental results. Potential application of the adaptive analysis to speech coding[9] is currently investigated.

ACKNOWLEDGMENT

Acknowledgment is made to the Director of Information Systems Research Center of Canon Incorporated for permission to publish this paper.

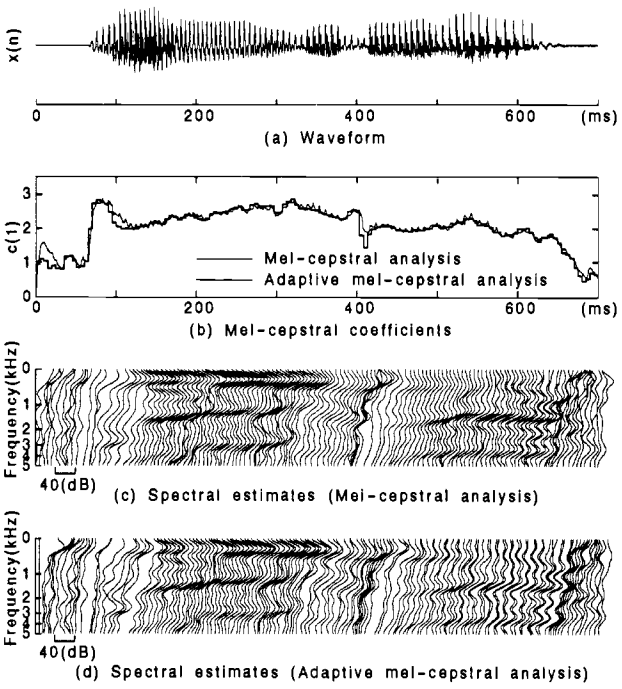


Fig. 6. An example of analysis for natural speech.

TABLE II Recognition rates for several analysis methods.

Analysis	intraspeaker	interspeaker	total
LPC	96 %	90.0 %	91.2 %
MC	99 %	91.8 %	93.2 %
AMC	98 %	91.0 %	92.4 %

REFERENCES

[1] G. Fant : "Speech sound and features," MIT Press, Cambridge (1973).

[2] S. Imai and C. Furuichi : "Unbiased estimator of log spectrum and its application to speech signal processing," in *Proc. 1989 EURASIP*, Sep. 1988, pp.203–206.

[3] B. Widrow and S. D. Stearns : *Adaptive Signal Processing*, Englewood Cliffs, N.J.: Prentice-Hall, 1985.

[4] S. Imai, K. Sumita and C. Furuichi : "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Trans. IECE*, vol. J66-A, pp.122–129, Feb. 1983.

[5] K. Tokuda, T. Kobayashi, T. Fukada, H. Saito and S. Imai : "Spectral estimation of speech based on mel-cepstral representation," *Trans. IEICE*, vol. J74-A, pp.1240–1248, Aug. 1991.

[6] I. Gohberg and I. Koltracht : "Efficient algorithm for Toeplitz plus Hankel matrices," *Integral Equations and Operator Theory*, vol. 12, pp.136–142, 1989.

[7] K. Tokuda, T. Kobayashi, Shoji Shiomoto and S. Imai : "Adaptive filtering based on cepstral representation — Adaptive cepstral analysis of speech," in *Proc. ICASSP 90*, Apr. 1990, pp.377–380.

[8] T. Kobayashi and S. Imai : "Complex Chebyshev approximation for IIR digital filters using an iterative WLS technique," in *Proc. ICASSP 90*, Apr. 1990, pp.1321–1324.

[9] T.Fukada, K. Tokuda, T. Kobayashi and S. Imai : "A study of speech coding based on the adaptive mel-cepstral analysis," *IEICE Technical Report*, SP91-73, Oct. 1991, pp.53–60.

[10] K. Tokuda, T. Kobayashi, T. Fukada and S. Imai : "Adaptive mel-cepstral analysis of speech," *Trans. IEICE*, vol. J74-A, pp.1249–1256, Aug. 1991.