CELP CODING SYSTEM BASED ON MEL-GENERALIZED CEPSTRAL ANALYSIS

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

†Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama, 226 Japan
‡Dept. of Intelligence and Computer Science, Nagoya Institute of Technology, Nagoya, 466 Japan

ABSTRACT

This paper presents a CELP speech coding system based on mel-generalized cepstral analysis. In the mel-generalized cepstral analysis, we can vary the model spectrum continuously from AR to cepstral modeling by changing the value of a parameter γ and we can choose an appropriate model spectrum. Furthermore, the spectrum represented by mel-generalized cepstrum has frequency resolution similar to that of human ear. Since the perceptual weighting and postfiltering are carried out through the mel-generalized cepstrum, we expect the perceptual performance of the proposed coder to be improved. The subjective performance test indicates that the quality of the proposed CELP coder is about 2 dB higher than that of the conventional one.

1. INTRODUCTION

Code Excited Linear Prediction (CELP)[1] coding has received considerable attention for high quality speech coding at low bit rates. For improvement of quality, reduction of computational complexity, and increase of robustness to channel errors, most of the studies have been focused on excitation structure, e.g., ACELP[2], CS-CELP[3], VSELP[4], PSI-CELP[5] etc. While CELP coders can provide fairly good quality speech at around 8 kbits/s, its performance at below 4 kbits/s is yet unsatisfactory for many applications. For further improvement of speech quality, it is necessary to investigate not only excitation structure but also spectral analysis method.

In CELP coders, AR modeling has been used for short-term predictor. However, it cannot represent spectral zeros. On the other hand, the cepstral modeling can represent spectral poles and zeros with equal weights. Moreover, mel-cepstrum is defined as frequency-transformed cepstrum and the spectrum represented by the mel-cepstral coefficients has frequency resolution similar to that of human ear which has high resolution at low frequencies. From the above point of view, we have proposed speech coders based on mel-cepstral analysis and showed the effectiveness of mel-cepstral representation in speech coding[6] [7].

In this paper, we introduce the mel-generalized cepstral analysis[8] to CELP coding. In the mel-generalized cepstral analysis, we can vary the model spectrum continuously from AR to cepstral modeling by changing the value of γ and we can choose an appropriate model spectrum. Furthermore, the spectrum represented by the mel-generalized cepstrum also has high resolution at low frequencies. Consequently, we can represent speech spectrum more efficiently. In addition, the proposed CELP coder includes the following features: (a) spectral representation using mel-generalized cepstrum with ensuring the filter stability after quantization; (b) perceptual weighting and postfiltering carried out through mel-generalized cepstrum; (c) bandwidth expansion in $\tilde{z}$ plane; and (d) adaptive control method for compensating spectral tilt in postfilter. The proposed CELP coder is subjectively compared to the conventional one in the same framework and bit allocations.

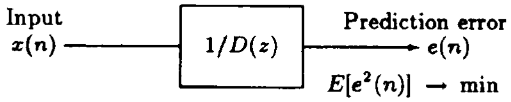

Figure 1: Interpretation of mel-generalized cepstral analysis as the least mean square of the linear prediction error.

2. SPECTRAL ANALYSIS

2.1. Spectral Model and Criterion

For a given order M, we assume that a speech spectrum $H(e^{j\omega})$ is modeled as follows:

$$H(z) = s_\gamma^{-1}\left(\tilde{\boldsymbol{z}}^T \boldsymbol{c}\right) \tag{1}$$

$$= \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c(m)\tilde{z}^{-m}\right)^{1/\gamma}, & -1 \le \gamma < 0 \\ \exp \sum_{m=0}^{M} c(m)\tilde{z}^{-m}, & \gamma = 0 \end{cases} \tag{2}$$

where the coefficients $\boldsymbol{c} = [c(0), \cdots, c(M)]^T$ are the mel-generalized cepstrum and $\tilde{\boldsymbol{z}} = [1, \tilde{z}^{-1}, \cdots, \tilde{z}^{-M}]^T$. The function $s_\gamma^{-1}(w)$ is the inverse of the generalized logarithmic function

$$s_\gamma(w) = \begin{cases} (w^\gamma - 1)/\gamma, & 0 < |\gamma| \le 1 \\ \log w, & \gamma = 0. \end{cases} \tag{3}$$

An all-pass system $\tilde{z}^{-1}$ is defined by

$$\tilde{z}^{-1} = (z^{-1} - \alpha)/(1 - \alpha z^{-1})\Big|_{z=e^{j\omega}} = e^{-j\tilde{\omega}} \tag{4}$$

where $|\alpha| < 1$. For a sampling frequency of 8 kHz, the phase characteristic $\tilde{\omega}$ of the system is a good approximation to the mel scale when $\alpha = 0.31$. From (2), it is noted that the spectral model becomes an all-pole model for $(\alpha, \gamma) = (0, -1)$ and cepstral representation for $(\alpha, \gamma) = (0, 0)$.

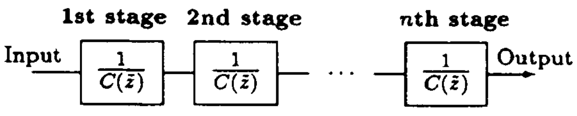

Figure 2: Synthesis filter $D(z)$ $(\gamma = -1/n)$.

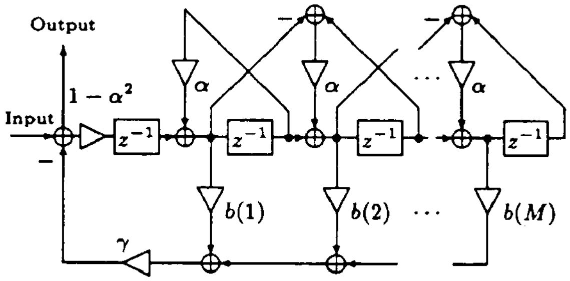

Figure 3: The structure of $1/C(\tilde{z})$, avoiding delay-free loop.

To determine the coefficients $\mathbf{c}$, we minimize the spectral criterion derived in the UELS[9]. Now, taking the gain factor K outside from $H(z)$, we have

$$H(z) = K \cdot D(z) = s_\gamma^{-1}(\boldsymbol{\alpha}^T \mathbf{c}) \cdot s_\gamma^{-1}(\tilde{\mathbf{z}}^T \mathbf{c}') \tag{5}$$

where $\boldsymbol{\alpha} = [1, (-\alpha), \cdots, (-\alpha)^M]^T$ and the coefficients $\mathbf{c}' = [c'(0), \cdots, c'(M)]^T$ are gain-normalized version of $\mathbf{c}$ given by

$$c'(m) = \begin{cases} \left(c(0) - \boldsymbol{\alpha}^T \mathbf{c}\right) / \left(1 + \gamma \boldsymbol{\alpha}^T \mathbf{c}\right), & m = 0 \\ c(m) / \left(1 + \gamma \boldsymbol{\alpha}^T \mathbf{c}\right), & 1 \le m \le M. \end{cases} \tag{6}$$

Since the gain of $D(z)$ is constrained to be unity, the output $e(n)$ of inverse filter $1/D(z)$ becomes the prediction error [10] as shown in Fig. 1. The minimization of the criterion leads to the minimization of the residual energy. We have given a computationally efficient iterative algorithm for solving the minimization problem and shown that a few iterations are sufficient to obtain the solution[8]. In addition, the stability of the model solution $H(z)$ is guaranteed[8].

2.2. Synthesis Filter

From (2) and (5), the synthesis filter $D(z)$ is not a rational function and, therefore, it cannot be realized directly. However, using MGLSA filter[11], $D(z)$ is approximated with sufficient accuracy and becomes minimum-phase IIR system. In particular, for $\gamma = -1/n$ and n is a natural number, we have

$$\begin{aligned} D(z) &= \{1/C(\tilde{z})\}^n && (7) \\ C(\tilde{z}) &= 1 + \gamma \tilde{\mathbf{z}}^T \mathbf{c}' = 1 + \gamma \sum_{m=0}^{M} c'(m) \tilde{z}^{-m}. && (8) \end{aligned}$$

In this case, the filter $D(z)$ is realized by the cascade of $1/C(\tilde{z})$ as shown in Fig. 2 without any approximation error.

To remove a delay-free loop from $D(z)$, we modify (8) as follows:

$$\begin{aligned} C(\tilde{z}) &= 1 + \gamma \tilde{\mathbf{z}}^T \boldsymbol{A} \boldsymbol{A}^{-1} \mathbf{c}' = 1 + \gamma \boldsymbol{\Phi}^T \boldsymbol{b} && (9) \\ &= 1 + \gamma \sum_{m=1}^{M} b(m) \Phi_m(z) && (10) \end{aligned}$$

where

$$\boldsymbol{A} = \begin{bmatrix} 1 & \alpha & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha \\ 0 & \cdots & 0 & 1 \end{bmatrix} \tag{11}$$

$$\begin{aligned} \boldsymbol{A}^T \tilde{\mathbf{z}} &= \left[1, \boldsymbol{\Phi}^T\right]^T = [1, \Phi_1(z), \cdots, \Phi_M(z)]^T && (12) \\ \Phi_m(z) &= \frac{(1-\alpha^2) z^{-1}}{1 - \alpha z^{-1}} \tilde{z}^{-(m-1)}. && (13) \end{aligned}$$

The coefficients $\boldsymbol{b}$ are given by

$$\boldsymbol{A}^{-1} \mathbf{c}' = \left[0, \boldsymbol{b}^T\right]^T = [0, b(1), b(2), \cdots, b(M)]^T. \tag{14}$$

where $\boldsymbol{A}^{-1}$ is the inverse matrix of $\boldsymbol{A}$. Instead of matrix multiplication $\boldsymbol{A}^{-1}\mathbf{c}'$, a recursive algorithm can be used:

$$b(m) = \begin{cases} c'(M), & m = M \\ c'(m) - \alpha b(m+1), & 0 \le m < M. \end{cases} \tag{15}$$

From (6), the first component of $\boldsymbol{A}^{-1}\mathbf{c}'$ is equal to zero, i.e., $\boldsymbol{\alpha}^T \mathbf{c}' = 0$. This indicates that the gain of $D(z)$ is unity. The structure of $1/C(\tilde{z})$ based on (10) is shown in Fig. 3.

2.3. Stability Problem

For $\gamma = 0$, i.e., cepstral model, the minimum-phase property of $D(z)$ is always preserved for any $\mathbf{c}'$[8]. However, for $-1 \le \gamma < 0$, it does not hold and, therefore, the filter may become unstable after quantization of $\mathbf{c}'$.

Assuming that $F(z) = F_1(z) + F_2(z)$ is a real polynomial and $F_1(z)$ and $F_2(z)$ are symmetric and antisymmetric polynomials, respectively. For $F(z)$ to be a stable polynomial, it is necessary and sufficient that[12] [13]: (a) The zeros of $F_1(z)$ and $F_2(z)$ are located on the unit circle; (b) They are simple; and (c) They separate each other. If $F(z)$ is the LPC polynomial, the angular positions of the zeros of $F_1(z)$ and $F_2(z)$ correspond to the LSP coefficients.

Since the all-pass system $\tilde{z}^{-1}$ maps the inside of the unit circle of the $\tilde{z}$ plane to the inside of the unit circle of the z plane if and only if $|\alpha| < 1$, the stability problem with respect to z plane is equivalent to that with respect to $\tilde{z}$ plane. Consequently, when the polynomial $C(\tilde{z})$ is decomposed into symmetrical and antisymmetrical polynomials, they have the roots on the unit circle of $\tilde{z}$ plane. The roots of interest are $e^{j\tilde{\omega}_i}$ for $i = 0, \cdots, M+1$. These parameters $\{\tilde{\omega}_i\}_{i=1,\cdots,M}$ are interpreted as the angular positions of the roots on warped frequency scale. (It is noted that $\tilde{\omega}_0 = 0$ and $\tilde{\omega}_{M+1} = \pi$ are fixed roots.) The minimum-phase property of $1/C(\tilde{z})$ is ensured after quantization if the parameters $\tilde{\omega}_i$ satisfy the above three conditions.

Fig. 4 shows the spectra estimated by mel-generalized cepstral analysis and fluctuation of the proposed spectral parameters for $(\alpha, \gamma, M) = (0.31, -1/3, 10)$. The LPC spectra with tenth order and fluctuation of the LSP parameters are also shown. Compared with LPC spectra, it is seen that the obtained spectra has high resolution at low frequencies. Although the distribution of each parameter is more limited than that of each LSP parameter, there exists a tradeoff between distribution range and spectral sensitivity. It is shown that the quantization and interpolation performance of the proposed parameters is slightly better than that of the LSP parameters[14].

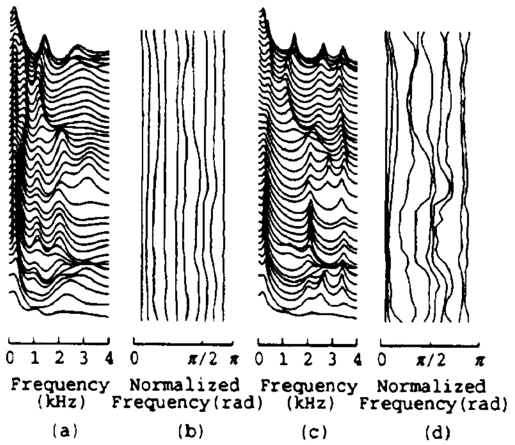

Figure 4: Spectral estimates and fluctuation of the spectral parameters. (a) mel-generalized cepstral analysis. (b) proposed parameter. (c) LPC analysis. (d) LSP.

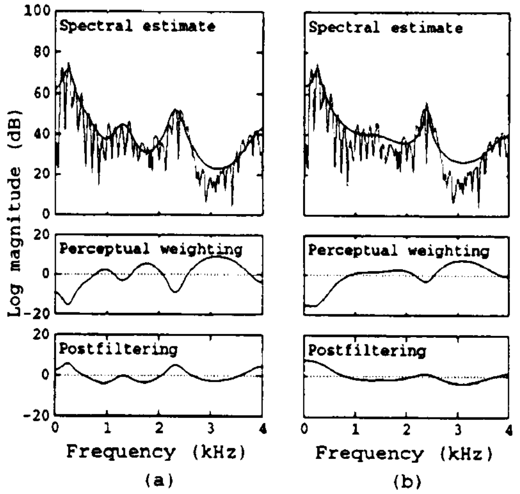

Figure 5: Example of perceptual weighting and postfiltering. (a) proposed coder. (b) conventional coder.

3. CODER STRUCTURE

In this section, we will describe the structure of the proposed CELP coder. Here, we let $(\alpha, \gamma, M) = (0.31, -1/3, 10)$ for spectral analysis. In this case, the synthesis filter is realized by three-stage cascade structure.

3.1. Perceptual Weighting Filter

The perceptual weighting filter $H_w(z)$ is defined by

$$H_w(z) = C(\tilde{z})C(\tilde{z}) \;/\; C(\tilde{z}/\beta_1) \tag{16}$$

where $\tilde{z}/\beta$ represents bandwidth expansion in $\tilde{z}$ plane. Since this bandwidth expansion operation replaces the coefficients $c'(m)$ by $\beta^m c'(m)$, the filter after bandwidth expansion can be written in the form

$$C'(\tilde{z}/\beta) = 1 + \gamma \tilde{\boldsymbol{z}}^T \boldsymbol{B} \boldsymbol{c}' \tag{17}$$

where matrix $\boldsymbol{B}$ is diagonal

$$\boldsymbol{B} = \text{diag}\left[1, \beta, \cdots, \beta^M\right]. \tag{18}$$

We modify (17) in a similar manner to (9):

$$\begin{aligned} C'(\tilde{z}/\beta) &= 1 + \gamma \tilde{\boldsymbol{z}}^T \boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{c}' && (19) \\ &= 1 + \gamma \left[1, \boldsymbol{\Phi}^T\right] \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{c}' && (20) \\ &= 1 + \gamma \left[1, \boldsymbol{\Phi}^T\right] \boldsymbol{b}'_\beta && (21) \end{aligned}$$

Again, the recursive algorithm can be used instead of the matrix multiplication $\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{c}'$:

$$b'_\beta(m) = \begin{cases} \beta^M c'(M), & m = M \\ \beta^m c'(m) - \alpha b'_\beta(m+1), & 0 \le m < M. \end{cases} \tag{22}$$

The gain of $C'(\tilde{z}/\beta)$ is not equal to zero, i.e., $b'_\beta(0) \neq 0$, due to bandwidth expansion in $\tilde{z}$ plane. Taking the gain K' outside from $C'(\tilde{z})$, we have

$$C'(\tilde{z}/\beta) = K' \cdot C(\tilde{z}/\beta) = (1 + \gamma b'_\beta(0))(1 + \gamma \boldsymbol{\Phi}^T \boldsymbol{b}_\beta). \tag{23}$$

The filter $C(\tilde{z}/\beta)$ is the gain-normalized version of $C'(\tilde{z}/\beta)$ and its coefficients $\boldsymbol{b}_\beta = [b_\beta(1), \cdots, b_\beta(M)]$ are given by

$$b_\beta(m) = b'_\beta(m)/(1 + \gamma b'_\beta(0)), \quad 1 \le m \le M. \tag{24}$$

From (23), replacing the coefficients $\boldsymbol{b}$ by $\boldsymbol{b}_\beta$ in Fig. 3, we can realize the filter $1/C(\tilde{z}/\beta)$ in the same manner as $1/C(\tilde{z})$.

3.2. Postfilter

The postfilter $H_p(z)$ has the transfer function

$$H_p(z) = \{C(\tilde{z}/\beta_2) \;/\; C(\tilde{z})\}\,(1 - \mu z^{-1})^p. \tag{25}$$

The postfilter consists of two parts: formant postfilter $C(\tilde{z}/\beta_2)/C(\tilde{z})$ and spectral tilt compensation filter $(1 - \mu z^{-1})^p$. The parameter μ controls the global spectral tilt in postfilter. For example, in conventional postfilter, the parameter μ was chosen to be 0.5, or $0.15k_1$ where k_1 is the first reflection coefficient and $p = 1$. Here, we propose a new adaptive control method for compensating global spectral tilt. This method is realized by setting the first mel-cepstral coefficient of $H_p(z)$ to be zero. The first mel-cepstral coefficient is calculated using the recursive formulas[15]. Under such constraint, μ is given by

$$\mu = \frac{-\gamma(1-\beta_2)c'(1)}{-\alpha\gamma(1-\beta_2) + p(1-\alpha^2)}. \tag{26}$$

To restrict absolute of μ to within unity, the tilt compensation filter has p-stage cascade structure.

3.3. Frequency Responses of Perceptual Weighting and Postfiltering

Fig. 5 shows the frequency responses of perceptual weighting filter and postfilter in proposed coder whose tunable parameters are $\beta_1 = 0.7, \beta_2 = 0.0, p = 2$. For comparison, those of conventional coder are also shown in Fig. 5. Their transfer functions are defined by

$$\begin{aligned} H_w(z) &= A(z/0.9) \;/\; A(z/0.4) && (27) \\ H_p(z) &= \{A(z/0.5) \;/\; A(z/0.8)\}\,(1 - 0.15k_1 z^{-1}) && (28) \end{aligned}$$

where $A(z)$ is LPC polynomial. From these figures, the spectral envelope obtained by mel-generalized cepstral analysis has high resolution at low frequencies; accordingly the spectra of perceptual weighting and postfiltering retain fine structure at low frequencies.

Table 1: Speech analysis conditions.

Sampling Frequency	8 kHz
Frame Length	20 ms
Subframe Length	10 ms
Order of Analysis	10
Window	32ms Hamming

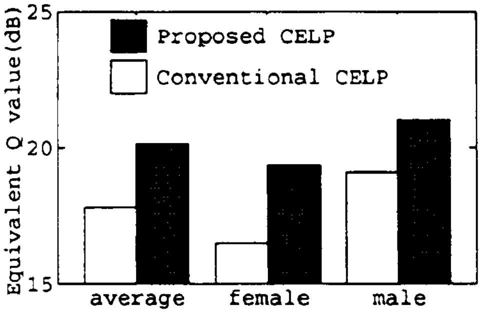

Figure 6: Opinion equivalent Q values.

4. PERFORMANCE

The performance of the proposed CELP coder is assessed by comparing it to the conventional one. In the experiment, the same framework is used except for the parts concerned with the spectral analysis and filter. We chose the parameter set $(\beta_1, \beta_2, p) = (0.7, 0.0, 2)$ for the perceptual weighting filter and postfilter of the proposed coder and used (27) and (28) for those of the conventional coder. The quantized spectral parameters are used for the coefficients of perceptual weighting filter. A harmonic weighting filter[16] is incorporated into perceptual weighting filter. The speech analysis conditions is provided in Table 1. Both coders operate at 4 kbits/s.

4.1. Framework of CELP System

The spectral parameters are first computed and coded with 24 bits/frame by the two-stage vector quantizer. Each stage has 4096 codevectors in its codebook. These codebooks are designed by the LBG algorithm. In the conventional coder, LSP parameters are quantized with the weighted Euclidean distance measure. This weighting function will have large weights when two LSP parameters are close to each other[17]. The proposed parameter described in 2.3 are quantized with the Euclidean distance measure.

The adaptive, random and gain codebooks are coded with 8, 11 and 9 bits/subframe, respectively. These codebooks are searched by closed-loop analysis sequentially. The adaptive codebook represents pitch lags from 20 to 147 samples with fractional delays. The shift-rotation and polarity techniques are applied to the random codebook. Each codevector is used eight times by shifting 2 samples in a rotational order. The gains of adaptive and random codebooks are vector quantized. The random and gain codebooks are alternatively trained 10 times by the generalized Lloyd algorithm[18].

4.2. Subjective Evaluation

The proposed CELP coder was evaluated by the opinion equivalent Q value and compared with the conventional one. The test material included twenty Japanese sentences uttered by six female and six male. Six listeners took part in the test. Fig. 6 shows subjective quality scores. The average scores are 20.1 dB and 17.8 dB for the proposed and conventional coders, respectively. It is shown that, on the average, the proposed coder achieves the improvement of about 2 dB over the conventional one. In particular, it seems that the perceptual weighting and postfiltering carried out through mel-generalized cepstrum make a large contribution to the improvement of the perceptual performance.

5. CONCLUSIONS

We have proposed CELP coding system based on mel-generalized cepstral analysis. By subjective evaluation, the performance of the proposed coder is about 2 dB higher than that of the conventional coder. This result indicates that the mel-generalized cepstral representation is effective in a low bit rate speech coding such as CELP. The coder design for IRS speech and noisy channels is a future problem.

ACKNOWLEDGMENT

This work is supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

REFERENCES

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

[2] R. Salami, et al. "8 kbit/s ACELP coding of speech with 10 ms speech frame: a candidate for CCITT standardization," *Proc. ICASSP'94*, pp.II-97–II-100, 1994.

[3] A. Kataoka, et al. "An 8-kbit/s speech coder based on conjugate structure CELP," *Proc. ICASSP'93*, pp.II-592–II-595, 1993.

[4] I. A. Gerson, et al. "Vector Sum Excited Linear Prediction (VSELP) speech coding at 8kbps," *Proc. ICASSP'90*, pp.461–464, 1990.

[5] S. Miki, et al. "Pitch synchronous Innovation CELP (PSI-CELP)," *Proc. EUROSPEECH'93*, pp.261–264, 1993.

[6] K. Tokuda, et al. "Speech coding based on adaptive mel-cepstral analysis," *Proc. ICASSP'94*, pp.197–200, Apr. 1994.

[7] K. Koishida, et al. "CELP coding based on mel-cepstral analysis," *Proc. ICASSP'95*, pp.33–36, 1995.

[8] K. Tokuda, et al. "Spectral estimation of speech by mel-generalized cepstral analysis," *Trans. IEICE*, vol. J75-A, pp.1124–1134, July 1992 (in Japanese). Translation: *Electronics and Communications in Japan (Part 3)*, vol. 76, no. 2, pp.30–43, July 1993.

[9] S. Imai, et al. "Unbiased estimator of log spectrum and its application to speech signal processing," *Proc. EURASIP'88*, pp.203–206, 1988.

[10] K. Tokuda, et al. "Generalized cepstral analysis of speech —unified approach to LPC and cepstral method," *Proc. ICSLP'90*, pp.37–40, 1990.

[11] T. Kobayashi et al. "Mel-generalized log spectral approximation filter," *Trans. IECE*, vol. J68-A, pp.610–611, Feb. 1985 (in Japanese).

[12] H. W. Schussler, "A stability theorem for discrete systems," *IEEE Trans. Acoust., Speech & Signal Processing*, vol. ASSP-24, pp.87–89, Feb. 1976.

[13] Y. Bistritz, "A discrete stability equation theorem and method of stable model reduction," *Systems & Contr. Lett.*, vol. 1, pp.373–381, May 1982.

[14] K. Koishida, et al. "Spectral representation of speech using mel-generalized cepstrum and its properties," *IEICE Technical Report*, SP95-49, pp.1–8, 1995 (in Japanese).

[15] K. Tokuda, et al. "Recursion formula for calculation of mel generalized cepstrum coefficients," *Trans. IEICE*, vol. J71-A, pp.128–131 Jan. 1988 (in Japanese).

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

[17] N. Phamdo, et al. "Combined source-channel coding of LSP parameters using multi-stage vector quantization," *IEICE Technical Report*, SP90-52, pp.63–70, 1990.

[18] T. Moriya, et al. "Training method of the excitation codebook for CELP," *Proc. EUROSPEECH'93*, pp.1155–1158, 1993.