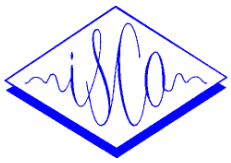

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

GENERALIZED CEPSTRAL ANALYSIS OF SPEECH — UNIFIED APPROACH TO LPC AND CEPSTRAL METHOD

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

†Department of Electrical and Electronic Engineering, Tokyo Institute of Technology, Tokyo, 152 Japan
††Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology, Yokohama, 227 Japan

ABSTRACT

This paper describes a spectral estimation method based on the generalized cepstral representation. The model spectrum represented by the generalized cepstrum varies from the all-pole spectrum to that represented by the cepstrum according to the value of the parameter γ in the range of $[-1, 0]$. We apply the criterion used in the unbiased estimation of log spectrum to the spectral model. As a result, the proposed method includes linear prediction and the unbiased estimation of log spectrum as the special cases. To solve the non-linear minimization problem involved in the method, we give an iterative algorithm whose convergence is guaranteed. The stability of the model solution is guaranteed. We also show some results of natural speech analysis to discuss the optimal value of γ in the sense of minimizing the prediction error.

I. INTRODUCTION

Linear prediction [1] is a generally accepted method for obtaining all-pole representations of speech, in which a small number of coefficients are necessary to represent the formants with narrow bandwidths. However, in some cases such as nasalization studies, spectral zeros are important and a more general modeling procedure 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 spectrum obtained by the conventional cepstral method [2] represents poles and zeros with equal weights. However, the cepstral method tends to overestimate the bandwidths of the formants, when the number of the cepstral coefficients is small.

The generalized cepstral coefficients [3] are identical with the cepstral and AR coefficients in the special cases where the parameter $\gamma = 0$ and -1, respectively. Using the generalized cepstral representation, we can vary the model spectrum continuously from the all-pole spectrum to that represented by the cepstrum according to the value of γ in the range of $[-1, 0]$. Furthermore, the spectral model has the advantages of both the all-pole and cepstral representations with an appropriate choice of γ.

Unfortunately, the spectral estimates obtained by the conventional method using the generalized cepstrum [3] has the bias caused by linear smoothing of the generalized logarithmic spectra. To overcome this problem, we apply the criterion used in the unbiased estimation of log spectrum [4],[5] to the spectral model based on the generalized cepstral representation. When $\gamma = 0$, the proposed method is identical with the unbiased estimation of log spectrum, in which the estimator of the log spectrum is represented by the cepstrum. Furthermore, since the criterion has the same form as that for linear prediction, the proposed method is identical with linear prediction when $\gamma = -1$.

The method involves a non-linear minimization problem except when $\gamma = -1$. We give a computationally efficient iterative algorithm in which a set of linear equations is solved using a fast algorithm that requires $O(M^2)$ arithmetic operations at each iteration. It can be shown that the convergence is quadratic and the stability of the model solution is guaranteed.

Finally, we show some results for synthetic signals and natural speech to discuss the optimal value of γ in the sense of minimizing the prediction error.

II. SPECTRAL MODEL AND CRITERION

The generalized logarithmic function is defined as

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

and, further, the generalized cepstrum $c_\gamma(m)$ is defined as the inverse Fourier transform of the generalized logarithm of a spectrum $X(e^{j\omega})$:

$$c_\gamma(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} s_\gamma \left(X(e^{j\omega}) \right) e^{j\omega m} \, d\omega \tag{2}$$

$$s_\gamma \left(X(e^{j\omega}) \right) = \sum_{m=-\infty}^{\infty} c_\gamma(m) \, e^{-j\omega m}. \tag{3}$$

In this paper, we assume that a speech spectrum $H(e^{j\omega})$ can be modeled by the $M+1$ generalized cepstral coefficients as follows:

$$H(z) = s_\gamma^{-1} \left(\sum_{m=0}^{M} c_\gamma(m) \, z^{-m} \right) = \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c_\gamma(m) \, z^{-m} \right)^{1/\gamma}, & 0 < |\gamma| \le 1 \\ \exp \sum_{m=0}^{M} c_\gamma(m) \, z^{-m}, & \gamma = 0. \end{cases} \tag{4}$$

Taking the gain factor K outside, we have

$$H(z) = K \cdot D(z) \tag{5}$$

where

$$K = s_\gamma^{-1} \left(c_\gamma(0) \right) \tag{6}$$

$$D(z) = \begin{cases} \left(1 + \gamma \sum_{m=1}^{M} c'_\gamma(m) \, z^{-m} \right)^{1/\gamma}, & 0 < |\gamma| \le 1 \\ \exp \sum_{m=1}^{M} c'_\gamma(m) \, z^{-m}, & \gamma = 0 \end{cases} \tag{7}$$

and $c'_\gamma(m)$ is the normalized generalized cepstrum [3] given by

$$c'_\gamma(m) = c_\gamma(m)/(1 + \gamma c_\gamma(0)), \quad m = 1, 2, \ldots, M. \tag{8}$$

For convenience, we will call $c'_\gamma(m)$ the generalized cepstrum and it will be denoted simply by $c_\gamma(m)$ in the following discussion.

From (5)-(7), it is seen that for $\gamma = -1$ the model spectrum takes the form of the all-pole representation and that for $\gamma = 0$ the model spectrum is identical with the spectrum represented by the cepstrum.

To obtain an unbiased estimate, we use the criterion used in the unbiased estimation of log spectrum [4],[5]:

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} \{ \exp R(\omega) - R(\omega) - 1 \} \, d\omega \tag{9}$$

$$R(\omega) = \log I_N(\omega) - \log \left| H(e^{j\omega}) \right|^2 \tag{10}$$

where $I_N(\omega)$ is the modified periodogram of a weakly stationary process $x(n)$, and $\left| H(e^{j\omega}) \right|^2$ is the estimator of the power spectrum.

When $\gamma = 0$, the proposed method is identical with the unbiased estimation of log spectrum, because the estimator of the log spectrum is represented by the cepstrum. Furthermore, (9) has the same form as the spectral criterion for linear prediction. Therefore, when $\gamma = -1$, the proposed method is identical with linear prediction based on the autocorrelation method.

If we regard $H(z)$ as the transfer function of a speech synthesis filter, $H(z)$ must be a stable system. Assuming that $H(z)$ is a minimum phase system, we can show that minimizing E with respect to K and

$$\mathbf{c} = [c_\gamma(1), c_\gamma(2), \ldots, c_\gamma(M)]^T \tag{11}$$

is equivalent to first minimizing

$$\varepsilon = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{I_N(\omega)}{|\, D(e^{j\omega})\,|^2}\, d\omega \tag{12}$$

with respect to c and then minimizing E with respect to K [6]. The gain factor K is obtained by setting $\partial E/\partial K = 0$:

$$K = \sqrt{\varepsilon_{min}} \tag{13}$$

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

By setting $\nabla\varepsilon = \partial\varepsilon/\partial \mathbf{c}$, we obtain a set of equations

$$\nabla\varepsilon = -2\,\mathbf{r} = 0 \tag{14}$$

where

$$\mathbf{r} = [r(1), r(2), \ldots, r(M)]^T \tag{15}$$

$$r(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{I_N(\omega)}{|\, D^{1+\gamma}(e^{j\omega})\,|^2} D^\gamma(e^{j\omega})\, e^{j\omega k}\, d\omega. \tag{16}$$

It can be shown that (14) gives a global minimum, because ε is convex with respect to c when $-1 \le \gamma \le 0$ [6]. Furthermore, the stability of the model solution $H(z)$ given by (14) is guaranteed when $-1 \le \gamma \le 0$ [6].

III. SOLUTION FOR THE SET OF EQUATIONS

The set of equations (14) is non-linear except when $\gamma = -1$. We use here the the Newton-Raphson method to solve (14). For the i-th result $\mathbf{c}^{(i)}$, solving a set of linear equations

$$\mathbf{H}\,\Delta\mathbf{c}^{(i)} = -\nabla\varepsilon \Big|_{\ \mathbf{c} = \mathbf{c}^{(i)}}\ , \tag{17}$$

we have the values

$$\Delta\mathbf{c}^{(i)} = [\Delta c_\gamma^{(i)}(1), \Delta c_\gamma^{(i)}(2), \ldots, \Delta c_\gamma^{(i)}(M)]^T \tag{18}$$

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

$$\mathbf{c}^{(i+1)} = \mathbf{c}^{(i)} + \Delta\mathbf{c}^{(i)}. \tag{19}$$

By substituting (12) in (17) and using a matrix representation, we have

$$\{\mathbf{P} + (1+\gamma)\mathbf{Q}\}\,\Delta\mathbf{c}^{(i)} = \mathbf{r} \Big|_{\ \mathbf{c} = \mathbf{c}^{(i)}} \tag{20}$$

where

$$\mathbf{P} = \begin{bmatrix} p(0) & \cdots & p(M-1) \\ \vdots & \ddots & \vdots \\ p(M-1) & \cdots & p(0) \end{bmatrix} \tag{21}$$

$$\mathbf{Q} = \begin{bmatrix} q(2) & \cdots & q(M+1) \\ \vdots & \ddots & \vdots \\ q(M+1) & \cdots & q(2M) \end{bmatrix} \tag{22}$$

and

$$p(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{I_N(\omega)}{|\, D^{1+\gamma}(e^{j\omega})\,|^2}\, e^{j\omega k}\, d\omega \tag{23}$$

$$q(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{I_N(\omega)}{|\, D^{1+2\gamma}(e^{j\omega})\,|^2} D^{2\gamma}(e^{j\omega})\, e^{j\omega k}\, d\omega. \tag{24}$$

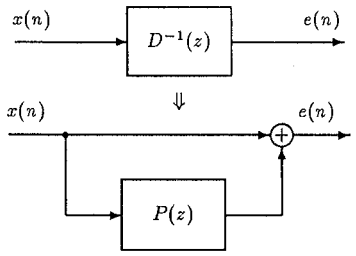

Fig. 1. Inverse filter $D^{-1}(z)$ and equivalent representation in term of linear predictor $P(z)$.

Since the coefficient matrix $\mathbf{P} + (1+\gamma)\mathbf{Q}$ is a symmetric Toeplitz plus Hankel matrix, the set of equations (20) can be solved using a fast recursive algorithm [7] which requires $O(M^2)$ arithmetic operations.

The convergence is quadratic, because when $-1 \le \gamma \le 0$ the Hessian matrix is positive definite, or identically, ε is convex with respect to c. When $\gamma = -1$, (14) can be solved directly, because (14) becomes a set of linear equations. In this case, (14) is equivalent to the normal equation in linear prediction. In other cases, we have found that typically a few iteration is sufficient to obtain the solution.

It is desirable that the initial guess $\mathbf{c}^{(0)}$ for c is close to the solution and can easily be obtained. As such an initial guess, we can use the generalized cepstrum corresponding to the smoothed generalized logarithmic spectrum, which is obtained by the following recursion formula [8]:

$$c_\gamma(m) = c(m) + \sum_{k=1}^{m-1} \frac{k}{m}\,\gamma\, c(k)\, c_\gamma(m-k), \quad m = 1, 2, \ldots, M \tag{25}$$

where $c(1), c(2), \ldots, c(M)$ are the cepstral coefficients obtained by the conventional cepstral method.

IV. INTERPRETATION AS LINEAR PREDICTION

Since the obtained transfer function $D(z)$ is minimum phase and the gain of $D(z)$ is unity, the inverse filter $D^{-1}(z)$ can be written as

$$D^{-1}(z) = 1 + P(z), \tag{26}$$

where

$$P(z) = \sum_{k=1}^{\infty} a(k)\, z^{-k} \tag{27}$$

and $a(k)$ is a stable sequence. As shown in Fig. 1, $P(z)$ is a linear predictor whose coefficients are $a(k)$, $k > 1$. Thus, the inverse filter output $e(n)$ shown in Fig. 1 is the prediction error given by

$$e(n) = x(n) - \sum_{k=1}^{\infty} x(n-k)\, a(k). \tag{28}$$

On the other hand, assuming the length of the time window N is sufficiently large, we can regard (12) as the mean square of $e(n)$, i.e.,

$$\varepsilon = E\left[e^2(n)\right]. \tag{29}$$

From (28) and (29), therefore, the proposed method can be interpreted as the minimization of the mean squared linear prediction error. It is noted that the linear predictor in the proposed method requires a linear combination of the previous all samples and the predictor coefficients are characterized by the Mth order generalized cepstral coefficients.

In the following section we will use the prediction gain defined by

$$G = \frac{E\left[x^2(n)\right]}{E\left[e^2(n)\right]} \tag{30}$$

to evaluate how well the model spectrum suit the signal to be analyzed.

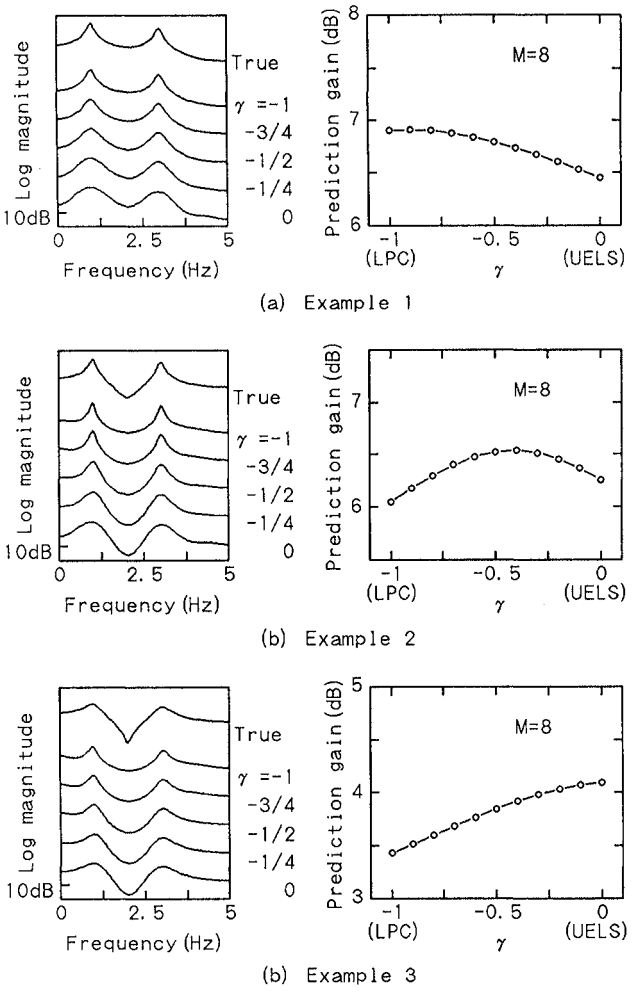

Fig. 2. Spectral estimates and prediction gain for synthetic signals.

V. EXPERIMENTAL RESULTS

Fig. 2 shows examples of spectral estimates for synthetic signals and the prediction gain G as a function of the parameter γ. In order to produce the signals to be analyzed, three digital filters were excited by a pulse train which has 75-point period. We arbitrarily set the sampling rate to 10kHz and chose the resonances and anti-resonances as shown in Table. I. The signals were windowed by a 25.6ms Blackman window and then analyzed using the generalized cepstral analysis with $M = 8$. The Fourier and inverse Fourier transforms involved in the algorithm were realized with 256-point FFTs. The spectra for $\gamma = -1$ and $\gamma = 0$ are identical with those obtained by conventional linear prediction (LPC) and the unbiased estimation of log spectrum (UELS), respectively.

Example 1 is a case where the true spectrum has only resonances. From Fig. 2(a), it is seen that a more accurate representation of the resonances is provided as γ approaches -1 and the prediction gain is maximized near $\gamma = -1$. On the other hand, the result of Example 3 (Fig. 2(c)), where the true spectrum has a narrow band anti-resonance and broad band resonances, shows that a valley corresponding to the anti-resonance appears more clearly as γ approaches 0. In this case, the prediction gain is maximized at $\gamma = 0$. From Example 2, where the true spectrum has narrow band resonances and a broad band anti-resonance as shown in Fig. 2(b), it appears that the obtained spectra

TABLE I Frequency response of the digital filters used for producing synthetic signals.
(Center frequency)/(Bandwidth) (Hz)

	resonance	anti-resonance	resonance
Example 1	1000/150	——	3000/150
Example 2	1000/150	2000/400	3000/150
Example 3	3000/400	2000/150	3000/400

with $-1 < \gamma < 0$ preserve both resonances and an anti-resonance, and the prediction gain is maximized at $\gamma = -0.4$.

Fig. 3 shows examples of spectral estimates for natural speech uttered by a male. The analysis conditions were the same as the previous examples except that $M = 15$. The speech segments are 25.6ms of /e/ and /N/ from the sentence "naNbudewa" sampled at 10kHz.

From Fig. 3(a), for a vowel, it is clear that the resonances are represented accurately and the prediction error is maximized when γ is close to -1. For a nasal case, it is seen from Fig. 3(b) that a more accurate representation of the anti-resonances is provided as γ approaches 0 at the expense of increasing the bandwidths of the resonances. The prediction gain takes the maximum value near $\gamma = -1/3$. Consistent with this result, the estimated spectrum with $\gamma = -1/3$ preserves both the resonances and anti-resonance with appropriate bandwidths.

Fig. 4 shows the estimated spectra for the sentence "naNbudewa". The values of γ which maximize the prediction gain are also shown. From the figure, every phoneme has its own optimal value of γ, i.e., at the portions corresponding to vowels, the prediction gain takes the maximum value near $\gamma = -1$, whereas at the portions corresponding to nasals, the prediction gain takes the maximum value near $\gamma = 0$. The parameter γ, however, is generally chosen to have a constant value. Thus, it is reasonable to choose the value of γ in such a way that the average of the prediction gain is maximized. Fig. 5 shows the average of the prediction gain for one minute speech (about 6000 frames). From the figure, the optimal value of γ, in the sense that the prediction gain is maximized, is -0.4 for $M = 15$. It is noted that the optimal value of γ depends on M. However, when $M = 25$, we have observed that the analysis-synthesis system based on the generalized cepstral representation [9] with $\gamma = -1/2$ or $-1/3$ generates higher quality speech than that with $\gamma = -1$ or 0.

VI. CONCLUSION

In this paper, we have described a new spectral estimation method based on the generalized cepstral representation, which is considered a generalization of linear prediction as well as the cepstral method. When $\gamma = -1$ and 0, the proposed method is identical with linear prediction and the unbiased estimation of log spectrum, respectively, and has the advantages of the both methods with an appropriate choice of γ. To discuss the optimal value of γ, we showed the prediction error for synthetic signals and natural speech. The method can be used for efficient spectral modeling of speech which includes nasals instead of the pole-zero modeling.

REFERENCES

[1] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Heidelberg: Springer-Verlag, 1986.

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

[3] T. Kobayashi and S. Imai, "Spectral analysis using generalized cepstrum," *IEEE Trans. Acoust., Speech, Signal processing*, vol. ASSP-32, pp.1087–1089, Oct. 1984.

[4] S. Imai and C. Furuichi, "Unbiased estimation of log spectrum," *Trans. IECE*, vol. J70-A, pp.461–480, Mar. 1987.

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

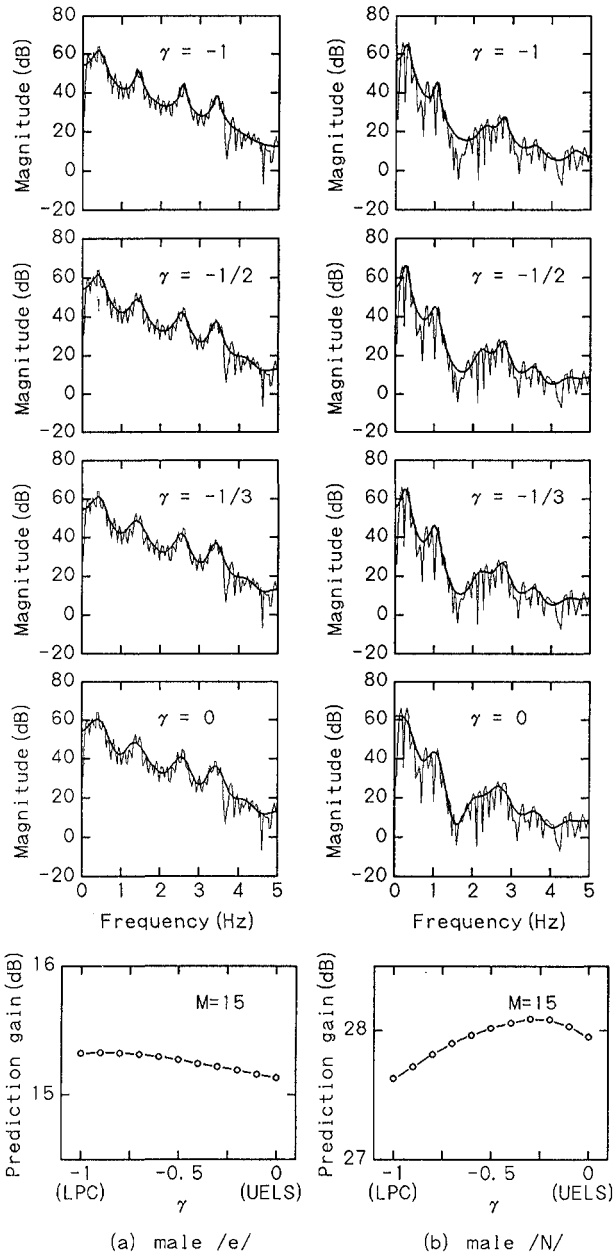

Fig. 3. Spectral estimates and prediction gain for natural speech.

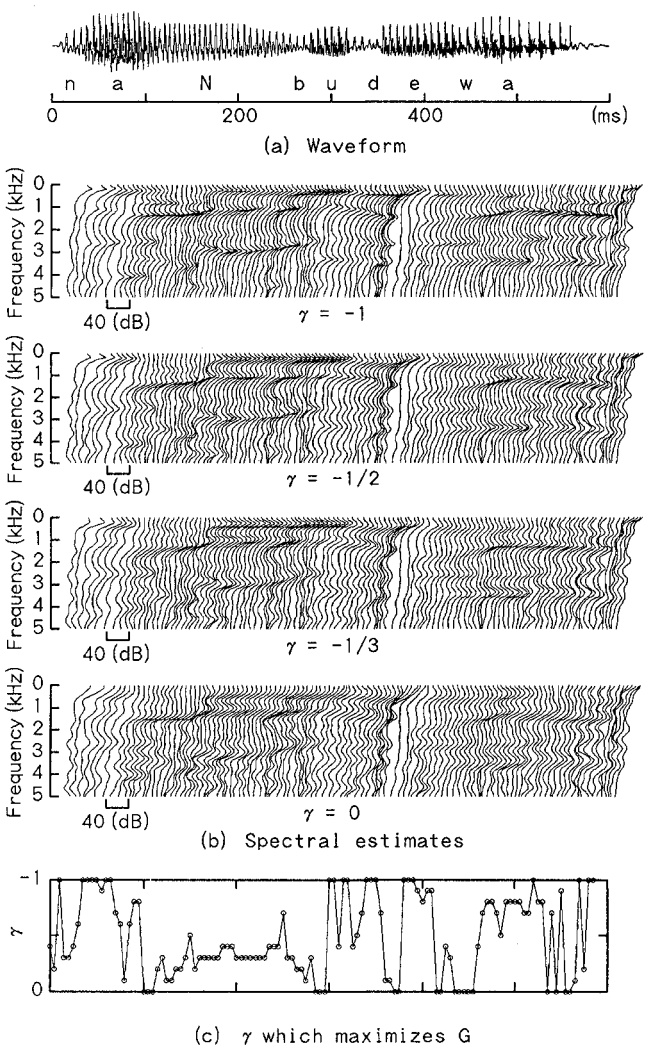

Fig. 4. Spectral estimates of natural speech and plot of γ which maximizes the prediction gain G in (30).

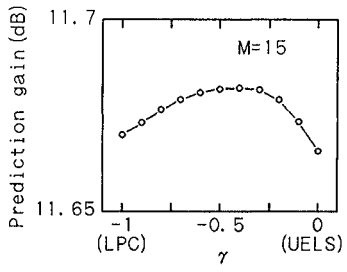

Fig. 5. Average of the prediction gain for natural speech.

[6] K. Tokuda, T. Kobayashi, R. Yamamoto and S. Imai, "Spectral estimation of speech based on generalized cepstral representation," *Trans. IEICE*, vol. J72-A, pp.457–465, Mar. 1989.

[7] G. A. Merchant and T. W. Parks, "Efficient solution of a Toeplitz-plus-Hankel coefficient matrix system of equations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp.40–44, Feb. 1982.

[8] K. Tokuda, T. Kobayashi and S. Imai, "Recursion formula for calculation of mel generalized cepstrum coefficients," *Trans. IEICE*, vol. J71-A, pp.128–131 Jan. 1988.

[9] K. Tokuda, T. Kobayashi and S. Imai : "Speech synthesis based on generalized cepstral representation", in *Proc. 1988 IEICE Spring National Convention Record*, Mar. 1988, A-34, 1-34.