

PAPER

Speech Analysis Based on AR Model Driven by t -Distribution Process

Junibakti SANUBARI†, *Nonmember*, Keiichi TOKUDA† and Mahoki ONODA†, *Members*

SUMMARY In this paper, a new M -estimation technique for the linear prediction analysis of speech is proposed. Since in the conventional linear prediction (CLP) method the obtained estimates are very much affected by the large amplitude residual parts, in the proposed method we use a loss function which assigns large weighting factor for small amplitude residuals and small weighting factor for large amplitude residuals which is for instance caused by the pitch excitations. The loss function is based on the assumption that the residual signal has an independent and identical t -distribution $t(\alpha)$ with α degrees of freedom. The efficiency of this new estimator depends on α . When $\alpha = \infty$, we get the CLP method. When the proposed method with small α is applied to the problems of estimating the formant frequencies and bandwidths of the synthetic speech by finding the roots of the prediction polynomial, we can achieve a more accurate and a smaller standard deviation (SD) estimate than that with large α . When the signal is very spiky, the proposed method can achieve more efficient and accurate estimates than that with robust linear prediction (RBLP) method. The loss function is modified in the similar manner as the autocorrelation method. The solution is calculated by the Newton-Raphson iteration technique. The simulation results show that only few iterations are needed to reach a stationary point, the stationary point is always a local minimum and the obtained prediction filter is always minimum phase. Preliminary experiments on the human speech data indicate that the obtained results are insensitive to the placement of the analysis window and a higher spectral resolution than the CLP and RBLP method can be achieved.

key words: AR modeling, t -distribution, M -estimate, speech analysis

1. Introduction

The formulation of the linear prediction (LP) model for speech analysis and synthesis is based on the linear model of speech production⁽¹⁾ as

$$S(z) = \frac{E(z)}{A(z)} \quad (1)$$

where

$$A(z) = 1 + \sum_{j=1}^p a_j z^{-j} \approx \frac{1}{G(z) \cdot V(z) \cdot L(z)} \quad (2)$$

The driving function is $E(z)$. The glottal shaping model, the lip radiation model and the all-pole vocal tract model are denoted by $G(z)$, $L(z)$ and $V(z)$ respectively. $S(z)$ is the speech signal. Equations (1)

and (2) can be equivalently expressed in the sampled data domain as

$$s_i + \sum_{j=1}^p a_j s_{i-j} = \varepsilon_i. \quad (3)$$

Equation (3) indicates that s_i is an autoregressive model of order p , AR(p). In the speech analysis, the innovation ε_i is often approximated either as an impulse train with period P for the voiced sounds, or as a random noise having a flat spectrum for the unvoiced sounds.

In the conventional LP(CLP) speech analysis, the predictor coefficients a_j , $1 \leq j \leq p$, are determined to minimize the sum of the squares of the prediction residuals. Therefore the result is least square fit. The same weighting function is assumed for all signal amplitude, so that the obtained estimate is very much affected by the strong signal parts and results in difficulties for the LP analysis of high-pitched voices.

In the CLP method, the structure of the source excitation is not taken into account. It is well known that for voiced speech, the source is of a quasi-periodic with spiky excitations which is not Gaussian process⁽²⁾. For these kinds of signals, the least square method is biased and inefficient⁽³⁾. If the source characteristic can be taken into account in the estimation of the prediction coefficients, we can get accurate and efficient estimates for the system parameters; formant frequency and bandwidth.

Several techniques have been introduced to reduce the error⁽⁴⁾⁻⁽⁷⁾. Miyoshi et al.⁽⁶⁾ and Yanagida et al.⁽⁷⁾ use the weighted least square method. Since the least square method is efficient only for exact Gaussian process, to get an efficient estimate the weighting function has to be chosen in such a way so that the data which is used to calculate the optimal predictor coefficient is Gaussian. Because of that the selection of the weighting function is critical and very difficult. On the other hand, Chin⁽⁴⁾ and Denöel et al.⁽⁵⁾ calculate the optimal coefficient by minimizing a loss function which is based on assuming that the residual signal has a certain probability distribution function. Since the exact distribution of the speech signal is unknown, the loss function should be carefully selected to get an efficient estimate for a wide range of distribution. Chin⁽⁴⁾ and Denöel et al.⁽⁵⁾ use the loss function based

Manuscript received October 25, 1991.

Manuscript revised February 17, 1992.

† The authors are with the Faculty of Engineering, Tokyo Institute of Technology, Tokyo 152 Japan.

on the heavy-tailed Huber's distribution function⁽³⁾ and the least absolute error, respectively. However still we can not get an efficient estimate for both Gaussian and very heavy-tailed processes. Recently, Murahara, Yoshida and the first author of this paper; used the similar approach as Chin⁽⁴⁾ and Denöel et al.⁽⁵⁾; suggested to select the predictor coefficient by assuming that the excitation has the t -distribution with three degrees of freedom; $\alpha=3$ ⁽⁸⁾. By using t -distribution with small degree of freedom, we can get an efficient estimate for wide range of distribution; both Gaussian and very heavy-tailed processes.

Extending the previous result⁽⁸⁾, in this paper, we report the more complete investigations about the other α , the efficiency and the accuracy of the estimator. The stability problem and the convergence, which have not been addressed in the previous paper⁽⁸⁾, and the relationship with the CLP method are also discussed in more detail. We modify the loss function in the similar manner as the autocorrelation method so that the proposed method can be seen as a generalization of the conventional autocorrelation method. The optimal solution is calculated iteratively by the the Newton-Raphson method. Simulation results show that we can always get a local minimum and a stable inverse system.

This paper is arranged as follows. The preliminary discussions are given in Sect. 2. In Sect. 3, the method of solving the optimization problem and the basic properties of the proposed method are given. The testing results are presented in Sect. 4. This paper is concluded in Sect. 5.

2. Preliminary Discussions

We consider a zero mean stationary time series generated by AR(p) model as is given in Eq.(3). The signal s_i is observed along a window; $1 \leq i \leq M$. The number of samples M is assumed to be large, $M \rightarrow \infty$. The signal outside the window is assumed to be zero. The residual signal ε_i can be expressed as a function of the linear prediction (LP) vector as

$$\varepsilon_i(\mathbf{a}) = s_i + \sum_{j=1}^p a_j \cdot s_{i-j},$$

where $\mathbf{a} = [a_1 \ a_2 \ \cdots \ a_p]^T$, (4)

and a_1, a_2, \dots, a_p are the linear prediction coefficients. The residual signal is assumed to have an independent and identical distribution (IID) $f(x)$. The logarithmic of the residual likelihood function is

$$L(\mathbf{a}|\boldsymbol{\varepsilon}) = \log \prod_{i=p+1}^M f(\varepsilon_i(\mathbf{a})) = \sum_{i=p+1}^M \log f(\varepsilon_i(\mathbf{a})),$$

where $\boldsymbol{\varepsilon} = [\varepsilon_{p+1} \ \varepsilon_{p+2} \ \cdots \ \varepsilon_M]^T$. (5)

The loss function is $\log f(\varepsilon_i(\mathbf{a}))$ and the influence function is defined as

$$\beta(x) = -\frac{\partial \log f(x)}{\partial x}. \quad (6)$$

The linear prediction coefficient vector \mathbf{a} is selected by maximizing the likelihood function in Eq. (5).

The Gaussian distribution

$$f_G(x) = \frac{1}{\sqrt{2\pi}} \exp^{-x^2} \quad (7)$$

is used for $f(x)$ in the CLP or the Gaussian estimation⁽¹⁾. The Huber's probability density function (PDF)⁽⁴⁾

$$f_H(x) = \frac{1-\theta}{\sqrt{2\pi}} \exp^{-\rho_H(x)}, \quad (8)$$

$$\rho_H(x) = \begin{cases} \frac{x^2}{2} & |x| \leq c \\ c|x| - \frac{c^2}{2}, & |x| > c \end{cases} \quad (9)$$

is used as $f(x)$ in the Huber's M -estimation. For heavy-tailed distribution processes, the Huber's M -estimate is more efficient than the CLP method^{(3),(4)}. This is because the Huber distribution is heavy-tailed, so that the influence function $\beta(x)$ assigns less weight for the large residuals caused by the spiky excitation. The plot of the Huber's influence function $\beta_H(x)$ with $c=1.5$ and the Gaussian's influence function $\beta_G(x)$ are shown in Fig. 1. In this paper it is proposed to use the heavy-tailed t -distribution model to construct a M -estimate. The t -distribution⁽⁹⁾ with α degrees of freedom, $t(\alpha)$ is defined by

$$f_\alpha(x) = \frac{1}{\sqrt{\alpha\pi}} \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \frac{1}{\left(1+\frac{x^2}{\alpha}\right)^{(\alpha+1)/2}}. \quad (10)$$

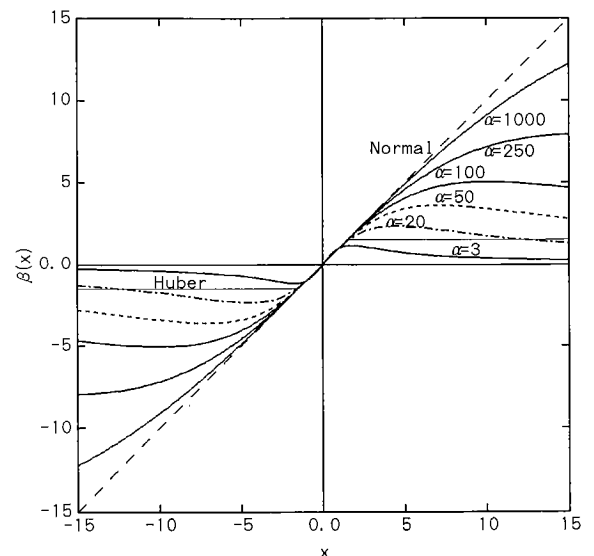


Fig. 1 The influence function $\beta(x)$ for various $f(x)$.

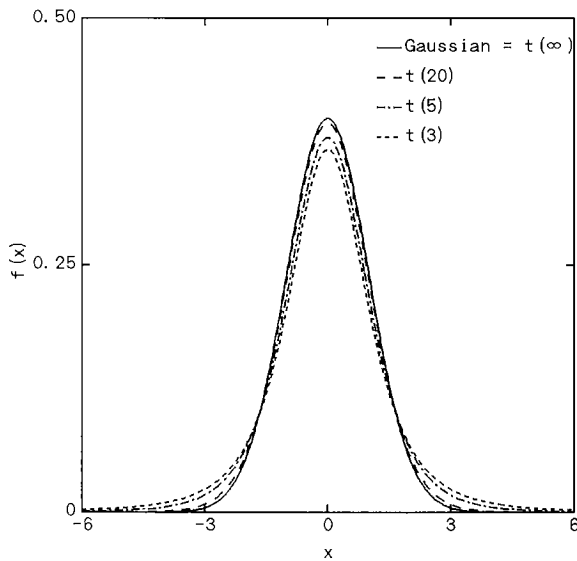


Fig. 2 The plot of the probability distribution $t(\alpha)$ for various α .

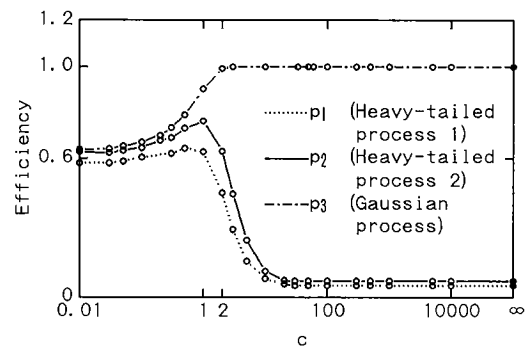
The PDF of the t -distribution with various α are shown in Fig. 2. The plot of t -distribution's influence function $\beta_t(x)$ with various α are shown in Fig. 1. Please note that $t(1)$ is the Cauchy distribution and $t(\infty)$ is the Gaussian distribution with unity standard deviation (SD) and zero mean $N(0, 1)$. For the estimation purpose, $f(x)$ has to have a finite second moment⁽⁹⁾. Since $f_\alpha(x)$ for $\alpha < 3$ has an infinite second moment⁽⁹⁾, here we use $\alpha \geq 3$.

The asymptotic efficiency⁽³⁾ of the estimator for heavy-tailed processes is used to evaluate the performance of the various estimation techniques. For the heavy-tailed process we use the contaminated Gaussian process

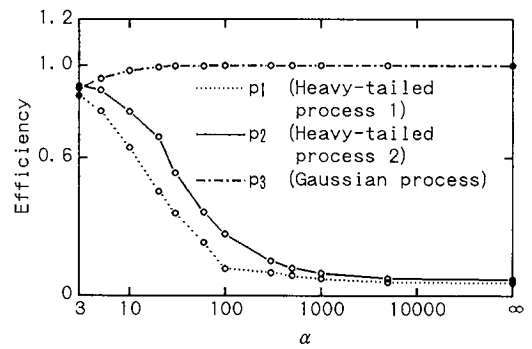
$$CND(\gamma, \chi) = (1 - \gamma)N(0, 1) + \gamma N(0, \chi). \quad (11)$$

Processes with large γ and χ have more probability on their tail. Various values of γ and χ were used to investigate the asymptotic efficiency of the estimators. The asymptotic efficiency of the Huber's M -estimate as a function of c for various processes are plotted in Fig. 3(a) and for the M -estimate with t -distribution as a function of α are depicted in Fig. 3(b).

When $c = \infty$ or $\alpha = \infty$, we get the CLP method. This method is efficient for the Gaussian process only. For heavy-tailed processes, the CLP method is not efficient. The efficiency can be improved by using the Huber's M -estimate. By setting c around 1.0, we can get an efficient estimate for both Gaussian and heavy-tailed processes. The recommended value of c is between 1.0 and 2.0⁽⁴⁾. However from Fig. 3(b), we can see that by utilizing the M -estimate using t -distribution with small α , we can get even more efficient estimate than the Huber's M -estimate for both Gaussian and heavy-tailed processes. Therefore we



(a)



(b)

p_1 is a heavy-tailed process with $\gamma=0.4$ and $\chi=10$, p_2 is a heavy-tailed process with $\gamma=0.2$ and $\chi=10$ and p_3 is a pure Gaussian process.

Fig. 3 The plot of the efficiency for various processes. (a) Using Huber's M -estimate. (b) Using M -estimate with t -distribution.

proposed to use the M -estimate with t -distribution for analyzing a complex speech signal where the exact distribution is unknown.

3. The Solution Method and the Basic Properties

The algorithm for solving the optimization problem and the basic properties of the proposed method are given in the following two subsections.

3.1 The Solution Method

The solution is calculated by extending the summation range in Eq.(5) to cover the whole range of the possible non-zero residual ε_i , so that we get a similar approach with the autocorrelation method. By doing so, this method can be seen as a generalization of the conventional autocorrelation method. By using Eq. (10) and the above assumption, Eq.(5) can be rewritten as

$$L(a|\varepsilon) = K_a - e \tilde{L}(a), \quad (12)$$

where

$$\boldsymbol{\varepsilon} = [\varepsilon_1 \varepsilon_2 \cdots \varepsilon_{M+p}]^T$$

$$\hat{L}(\mathbf{a}) = \sum_{i=1}^{M+p} \log \left(1 + \frac{\left(\frac{\varepsilon_i(\mathbf{a})}{\hat{s}} \right)^2}{\alpha} \right) \quad (13)$$

$$K_\alpha = (M+p) \log \left(\frac{1}{\sqrt{\alpha\pi}} \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \right) \quad (14)$$

$$e = \frac{\alpha+1}{2}. \quad (15)$$

To get a scale-invariant estimate, ε_i is normalized with \hat{s} . As the mean and the variance of the residual signal are very sensitive to outliers, they are not a good choice for \hat{s} . The median m , defined so that $p(|\varepsilon_i| \leq m) \geq 0.5$ and $p(|\varepsilon_i| \geq m) \geq 0.5$, $1 \leq i \leq M+p$, is less sensitive because it is not determined by the exact shape of the PDF⁽³⁾. Therefore we choose

$$\hat{s} = \text{median } |\varepsilon_i|, \quad 1 \leq i \leq M+p. \quad (16)$$

Because K_α and e are constants, maximizing $L(\mathbf{a}|\boldsymbol{\varepsilon})$ in Eq.(12) is equivalent to minimizing $\hat{L}(\mathbf{a})$. Equation (12) is optimized by the Newton-Raphson procedure as follows:

$$\mathbf{G}\mathbf{a}^{k+1} = \mathbf{G}\mathbf{a}^k - \nabla. \quad (17)$$

where k is the iteration number.

The gradient vector ∇ is given in Eq.(18).

$$\nabla = \left[\frac{\partial \hat{L}(\mathbf{a})}{\partial a_1} \frac{\partial \hat{L}(\mathbf{a})}{\partial a_2} \cdots \frac{\partial \hat{L}(\mathbf{a})}{\partial a_p} \right]^T \quad (18)$$

where

$$\frac{\partial \hat{L}(\mathbf{a})}{\partial a_z} = \frac{2}{\alpha \hat{s}^2} \sum_{i=1}^{M+p} \varepsilon_i \cdot w_i \cdot s_{i-z}. \quad (19)$$

The positive definite matrix \mathbf{G} is given by

$$\mathbf{G} = \begin{bmatrix} G_{1,1} & G_{1,2} & \cdots & G_{1,p} \\ G_{2,1} & G_{2,2} & \cdots & G_{2,p} \\ \vdots & \vdots & & \vdots \\ G_{p,1} & G_{p,2} & \cdots & G_{p,p} \end{bmatrix} \quad (20)$$

where

$$G_{r,t} = G_{t,r} = \frac{2}{\alpha \hat{s}^2} \sum_{i=1}^{M+p} s_{i-r} \cdot s_{i-t} \cdot w_i, \quad (21)$$

$$w_i = \frac{1}{1 + \frac{(\varepsilon_i / \hat{s})^2}{\alpha}} \quad (22)$$

Since \mathbf{G} is a positive definite matrix (see Appendix A), the Cholesky decomposition is used in the calculation. The result from the CLP or the Levinson-Durbin method is used as the starting value. Two criterions

$$\sqrt{\sum_{z=1}^p \left(\frac{\partial \hat{L}(\mathbf{a}^k)}{\partial a_z} \right)^2} \leq 10^{-4} \quad (23)$$

and

$$|\hat{L}(\mathbf{a}^k) - \hat{L}(\mathbf{a}^{k-1})| \leq 10^{-4} \quad (24)$$

are used to terminate the iteration. Simulation results show that 10^{-4} is a suitable value for stopping the iteration. No further significant improvements can be obtained when a value lower than 10^{-4} is used. Only the number of iteration will increase.

The sequence of the calculation can be summarize as follows:

1. Calculate the initial \mathbf{a}^0 by the CLP method.
2. Calculate new \hat{s} based on Eq.(16).
3. Calculate new \mathbf{a} based on Eq.(17).
4. Repeat step 2 and 3 until either one or both of the stopping criterions in Eq.(23) and (24) are reached.

3.2 The Basic Properties

The basic properties of the proposed method are as follows:

1. In the iteration we use a positive definite matrix \mathbf{G} , instead of the exact Hessian matrix which is not always positive. The simulation results show that by using this method, usually only few iterations are needed to reach a stationary point.
2. The simulation results show that at the stationary point, the Hessian matrix is always positive definite. Thus we get a local minimum point.
3. The obtained inverse system in the simulations is always stable.
4. The conventional autocorrelation method is a special case of the proposed method; when $\alpha = \infty$. In this case the optimal solution can be obtained without any iteration.

The brief proof of the positive definite properties of matrix \mathbf{G} is given in Appendix A. In Appendix B and Appendix C, the positive definite property of the Hessian matrix at the stationary point and the stability of the inverse system are discussed, respectively.

4. The Simulation Results

The proposed method has been implemented on a workstation and tested on both synthetic speech signals and human speech signals. All calculations were done with a double-precision floating point arithmetic. The overall results are reported in the following two subsections.

4.1 Testing Results on the Synthetic Speech

The output signal from a 10th-order all pole system was used as the synthetic speech signal to test

the performance of the proposed method. The sampling rate of 10 KHz was used. The system has 800 Hz and 100 Hz as the first formant frequency f_1 and bandwidth B_1 , respectively. The second formant frequency f_2 is 1270 Hz and the bandwidth B_2 is 120 Hz. The third formant frequency f_3 and bandwidth B_3 are 2022 Hz and 382 Hz, respectively. The fourth formant frequency f_4 and the fifth f_5 are 3304 Hz and 4415 Hz respectively while the fourth formant bandwidth B_4 and fifth formant bandwidth B_5 are 649 Hz and 450 Hz.

A pulse train with 400 Hz pitch frequency was used as the first input of the system. We used 10th-

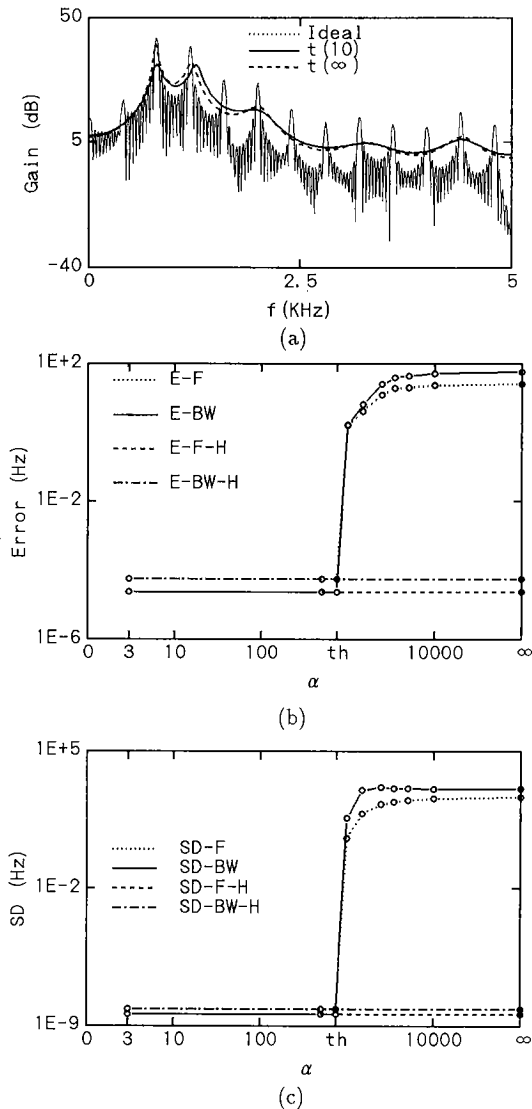


Fig. 4 Estimation results of the 400 Hz pulse excitation and rectangular window. (a) The ideal and estimated spectra. (b) The averages of the errors of the formant frequencies (E-F) and bandwidths (E-BW) from the proposed method, E-F-H and E-BW-H are from RBLP. (c) The averages of the SD of the formant frequencies (SD-F) and bandwidths (SD-BW) from the proposed method, SD-F-H and SD-BW-H are from RBLP.

order predictor and various values of α . The analysis was performed every 1 ms using a 25.6 ms rectangular window. We analyzed 800 frames. Figure 4(a) shows the ideal spectrum of the system, the periodogram and the estimated spectrum using $\alpha=10$ and $\alpha=\infty$ for a frame where the CLP method gives the largest spectrum error. The periodogram is plotted with a solid and thin line. Although the ideal spectrum of the system is plotted with a dotted line, it coincides with the estimated spectrum using $\alpha=10$. Therefore we only see one solid line. From those 800 frames, the average and the standard deviation (SD) of the estimated five formant frequency and bandwidth were calculated. The absolute difference between the averages and the true values, called the absolute average errors, were calculated. The average of the five formant frequencies and bandwidths absolute error as a function of α are shown in Fig. 4(b); marked as E-F and E-BW, respectively. The average of the SD of the five formant frequencies and bandwidths as a function of α are depicted in Fig. 4(c); marked as SD-F and SD-BW, respectively. These figures show that when the CLP method ($\alpha=\infty$) is used, the error is large. This is because we analyzed a high-pitched signal which means that it has more probability on its tail. This is consistent with the previous result in Fig. 3. The accuracy and the SD of the estimation can be improved by using a small α . When small α is used we can get a better estimate; the average error is undetectable and the SD is low. Figure 4(b) and Fig. 4(c) show that there is a threshold value, th . When $\alpha < th$, the estimate is less biased and efficient. Because of the space limitation, the results for the other pitch frequencies are omitted, but they show that when the pitch frequency is small, th is high and vice versa.

The second synthetic speech signal is generated by applying a random binary sequence as the excitation of the system to simulate a more complex spiky signals. The random binary sequence was generated by using a non-linear operation

$$rb_i = \begin{cases} 1 & \text{if } fl_i \geq 0.75 \\ -1 & \text{if } fl_i \leq -0.75 \\ 0 & \text{otherwise} \end{cases} \quad (25)$$

where rb_i is the random binary sequence and fl_i is the random signal which has a flat distribution between -1 and 1 . We carried out the analysis in the same way as above.

The estimated and the ideal spectrum are plotted in Fig. 5(a). The averages of the absolute error of the five formants and the averages of the SD are depicted in Fig. 5(b) and Fig. 5(c), respectively. Again we see that when small α is used we can get a precise and an efficient estimate for the spectrum and the formants.

In addition, we have also applied the 25.6 msec Hamming window in the analysis. The estimated

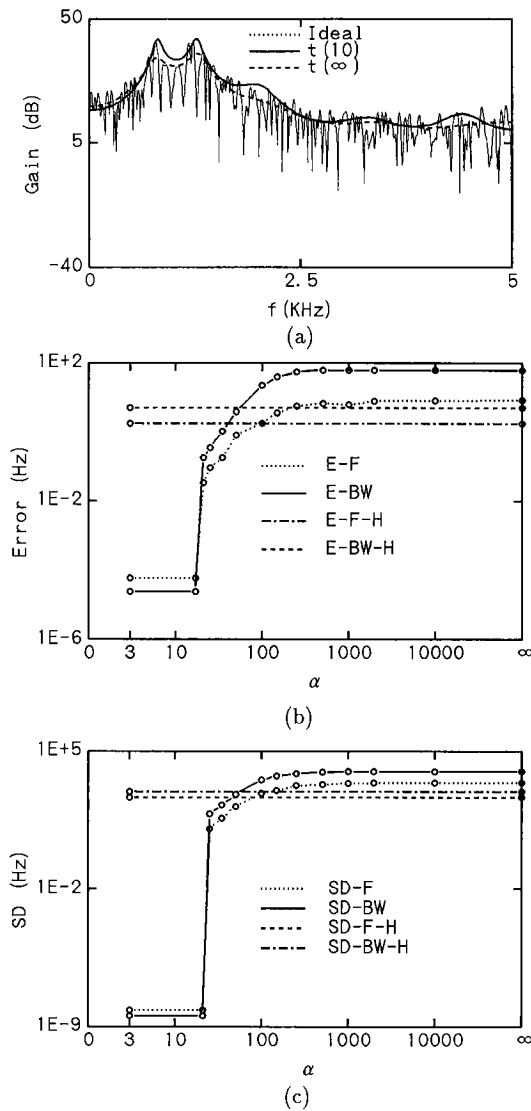


Fig. 5 Estimation results of the random binary excitation and rectangular window. (a) The ideal and estimated spectra. (b) The averages of the errors of the formant frequencies (E-F) and bandwidths (E-BW) from the proposed method, E-F-H and E-BW-H are from RBLP. (c) The averages of the SD of the formant frequencies (SD-F) and bandwidths (SD-BW) from the proposed method, SD-F-H and SD-BW-H are from RBLP.

spectrum and the ideal spectrum, the average of the absolute error of the five formants and the average of the SD as a function of α for the 400 Hz pulse train and the random binary sequence inputs are shown in Fig. 6 and Fig. 7. From those figures we can see that the window improves the accuracy and the efficiency of the CLP method. But when small α is used, we can get a more better results. The non-uniform weighting of the window affects the result of the proposed method. The improvement, however, caused by the proposed method is less significant than when a uniform weight-

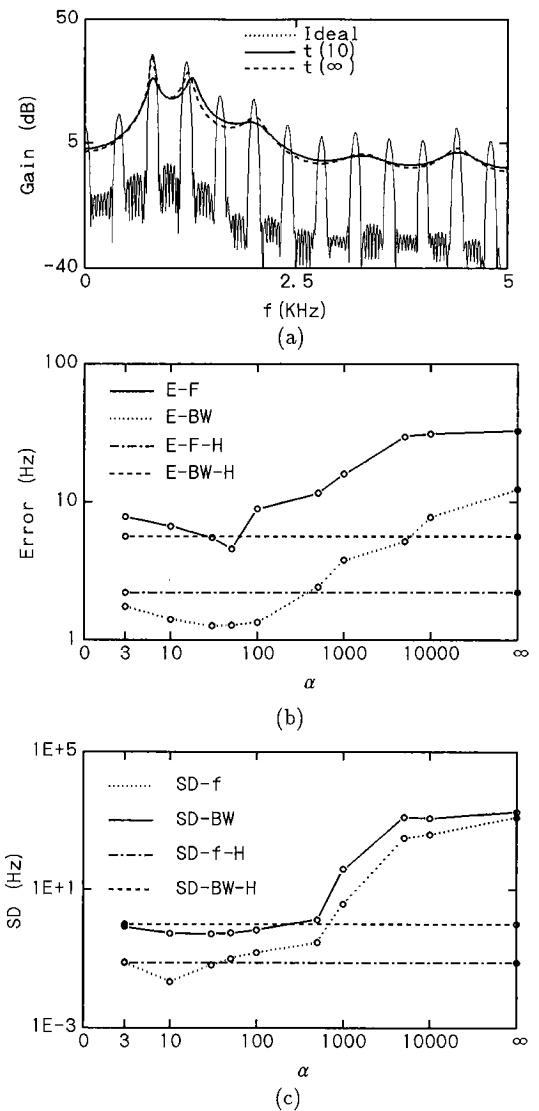


Fig. 6 Estimation result of the 400 Hz pulse excitation and Hamming window. (a) The ideal and estimated spectra. (b) The averages of the errors of the formant frequencies (E-F) and bandwidths (E-BW) from the proposed method, E-F-H and E-BW-H are from RBLP. (c) The averages of SD of formant frequencies (SD-F) and bandwidths (SD-BW) from the proposed method, SD-F-H and SD-BW-H are from RBLP.

ing window such as the rectangular window is used.

From all those simulations, it is clear that the accuracy and the efficiency of the proposed linear prediction method depends on α . Furthermore, all simulations show that indeed only few iterations are sufficient to reach a local minimum and the obtained inverse systems are always minimum phase.

To compare the performance of the proposed method with the recently proposed robust linear prediction (RBLP)⁽⁴⁾, a program with Eq.(9) as the loss function and Eq.(16) as the robust scale estimate has

been developed. The optimal solution is calculated by using the Newton-Raphson. It was tried to analyze both synthetic signals. Since when c lower than 1.0 is used we get unstable inverse systems in some frames, we set $c=1.0$ as the lowest value. The average of the absolute error of the five formants and the average of the SD are marked as E-F-H, E-BW-H, SD-F-H and SD-BW-H, respectively. The obtained values by using the rectangular and Hamming analysis window for both synthetic signal are plotted in Fig. 4, 5, 6, and 7. When higher c is used, the average of the absolute error and the average of the SD will be higher too. The

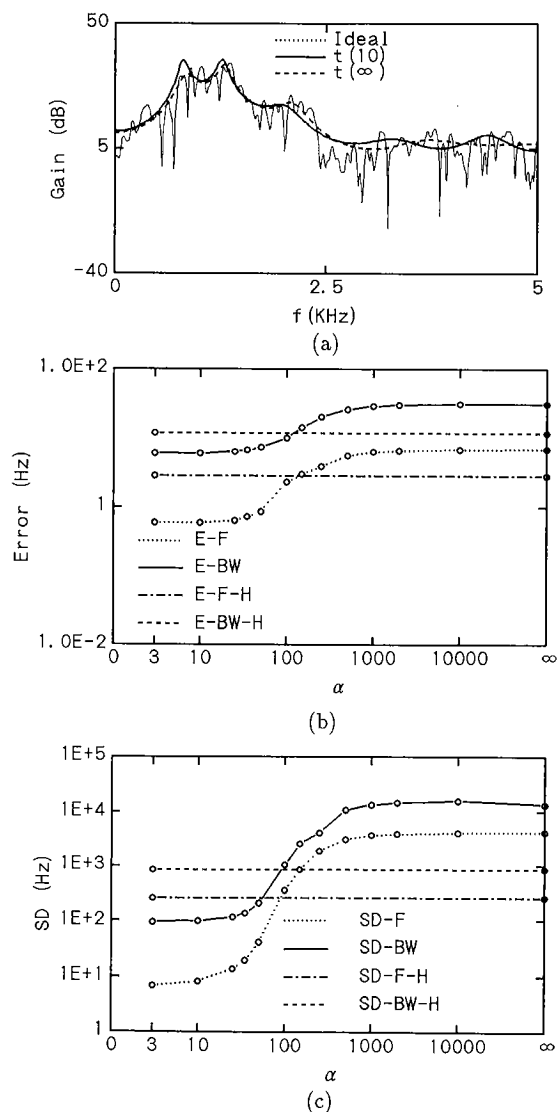


Fig. 7 Estimation results of the random binary excitation and Hamming window. (a) The ideal and estimated spectra. (b) The averages of the errors of the formant frequencies (E-F) and bandwidths (E-BW) from the proposed method, E-F-H and E-BW-H are from RBLP. (c) The averages of the SD of the formant frequencies (SD-F) and bandwidths (SD-BW) from the proposed method, SD-F-H and SD-BW-H are from RBLP.

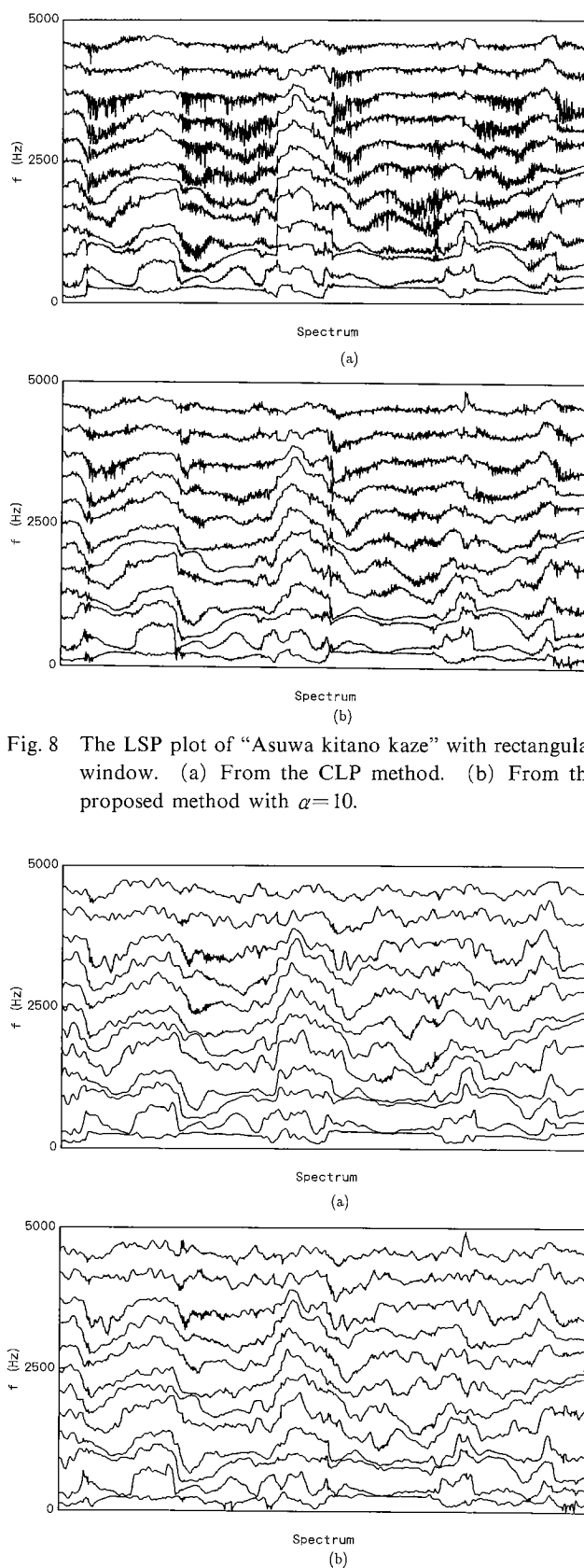


Fig. 8 The LSP plot of "Asuwa kitano kaze" with rectangular window. (a) From the CLP method. (b) From the proposed method with $\alpha=10$.

Fig. 9 The LSP plot of "Asuwa kitano kaze" with Hamming window. (a) From the CLP method. (b) From the proposed method with $\alpha=10$.

results show that when the signal is very spiky; the random binary case; the proposed method is more superior than the RBLP. When the signal is not very spiky; the 400 Hz impulse case; the performance of the RBLP is comparable with the proposed using small degree of freedom α .

4.2 Testing Results on the Human Speech

The proposed algorithm has been applied to analyze the human speech; the Japanese sentence "Asuwa kitano kaze" spoken by a female speaker with pitch frequency around 230 Hz. The sampling rate of 10 KHz and 12th-order predictor were used. We per-

formed the analysis every 1 ms on a 25.6 ms rectangular window and Hamming window. We analyzed 1200 frames. Since the true formant frequencies and bandwidths are unknown, the average error can not be calculated. We can only analyze the fluctuation of the obtained results caused by the analysis window's placement. As an illustration, we use the line spectrum pairs (LSP)⁽¹¹⁾. The obtained LSP from the CLP method and the proposed method with $\alpha=10$ both using a rectangular window are depicted in Fig. 8(a) and Fig. 8(b) respectively, while the LSP from the CLP method and the proposed method with $\alpha=10$ both using a Hamming window are shown in Fig. 9(a) and Fig. 9(b). Here, we can get a minimum phase system from

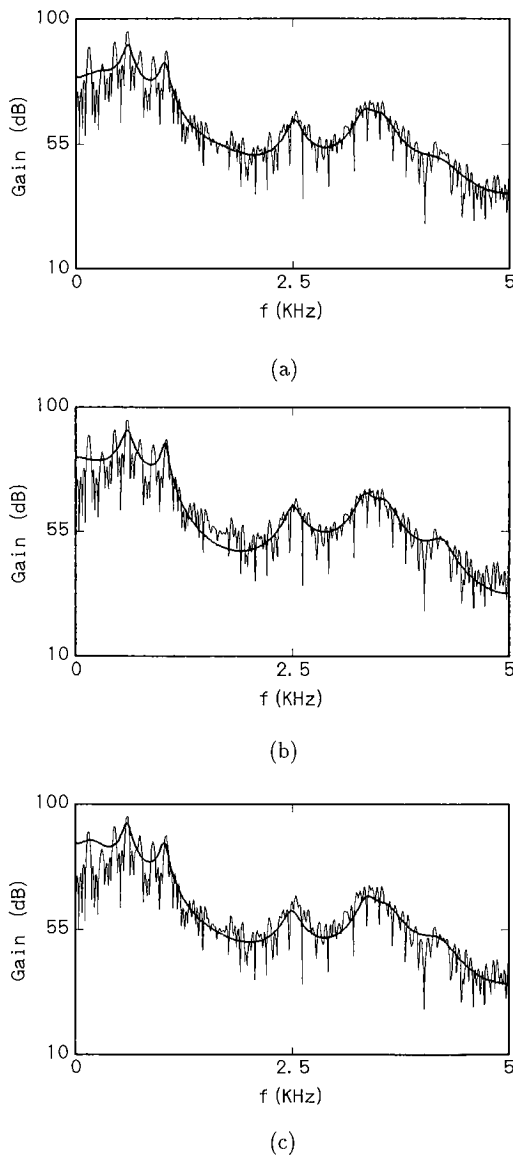


Fig. 10 The spectrum of the Japanese vowel /a/ with rectangular window. (a) From the CLP method. (b) From the the RBLP method. (c) From the proposed method with $\alpha=10$.

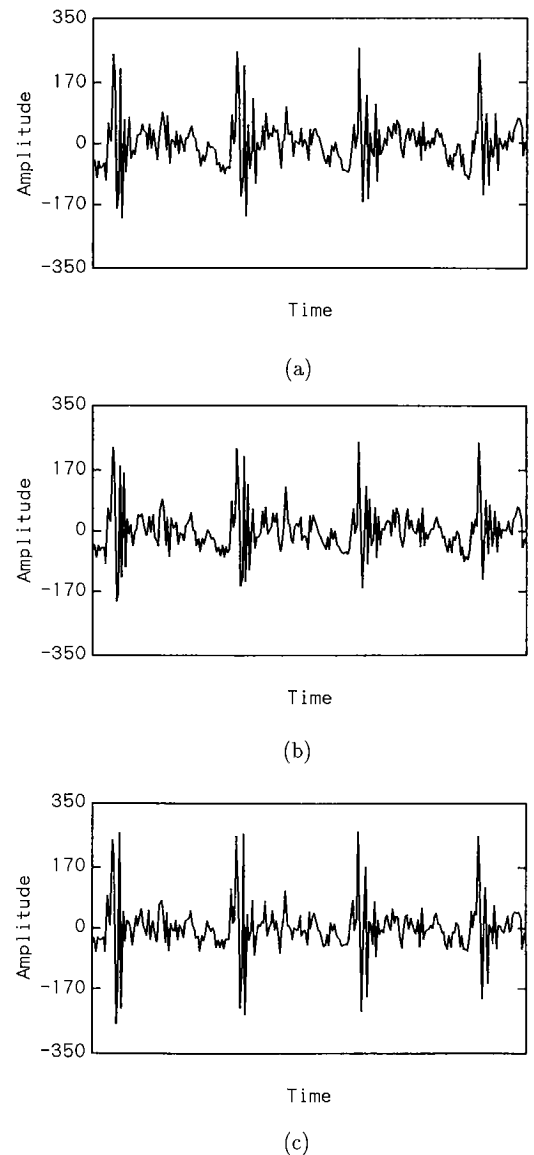


Fig. 11 The residual signal of the Japanese vowel /a/ with rectangular window. (a) From the CLP method. (b) From the RBLP method. (c) From the proposed method with $\alpha=10$.

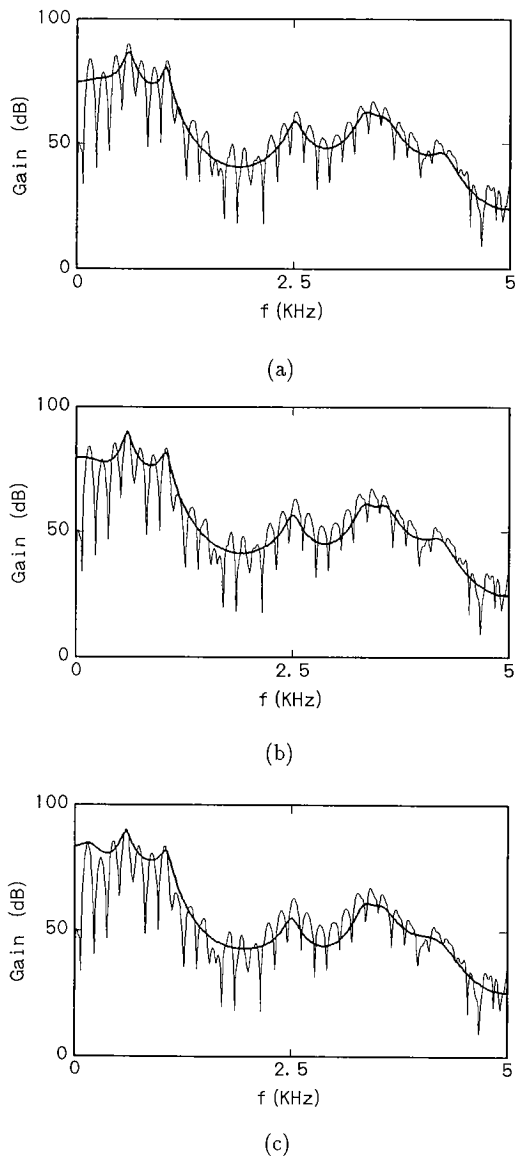


Fig. 12 The spectrum of the Japanese vowel /a/ with Hamming window. (a) From the CLP method. (b) From the RBLP. (c) From the proposed method with $\alpha=10$.

the proposed method. Those figures show that the obtained LSP from the proposed method exhibit smaller local variations due to the positioning of the window than that from the CLP method, so that by using the proposed method it is expected that a higher coding efficiency can be achieved⁽¹²⁾.

The proposed method has been also applied to analyze a vowel /a/ uttered by a Japanese male which has a pitch frequency around 150 Hz. The obtained spectrum from the CLP, the RBLP with $c=1.0$ and the proposed method with $\alpha=10$; all using 17th-order predictor and 256 points rectangular window; are shown in Figs. 10(a), 10(b) and 10(c), respectively. The proposed method shows its superiority by successfully recognized the first formant which can not be

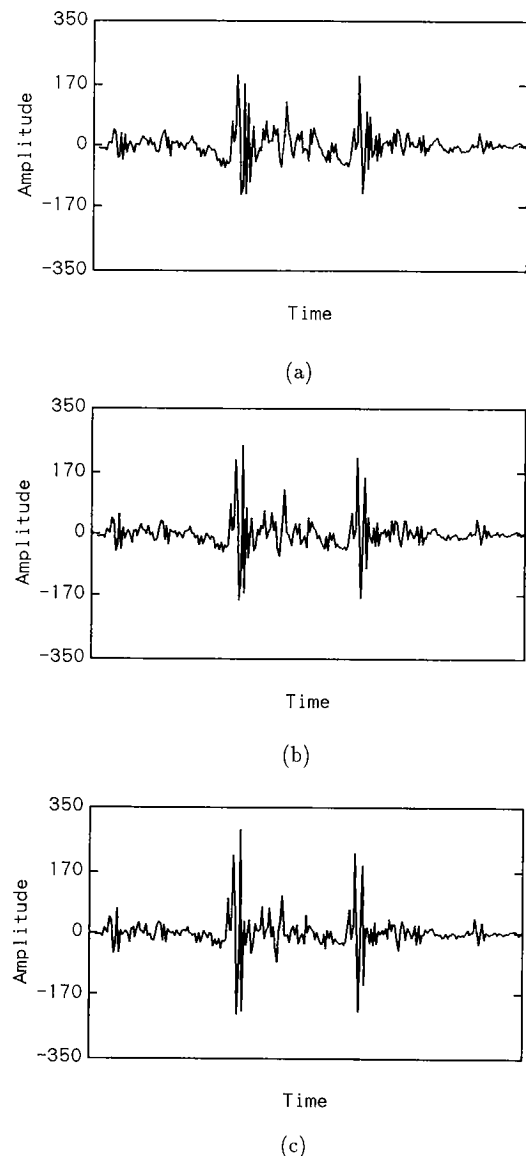


Fig. 13 The residual signal of the Japanese vowel /a/ with Hamming window. (a) From the CLP method. (b) From the RBLP method. (c) From the proposed method with $\alpha=10$.

recovered by the CLP method and the RBLP method. The resulted residual signal from the proposed method; shown in Fig. 11(c); is more spiky than the residual signal from the CLP and RBLP method which are shown in Figs. 11(a) and 11(b), respectively. This result indicates the possibility of using the residual signal for pitch recognition. When Hamming window is used, the obtained spectral and residuals from the CLP, RBLP and the proposed method; depicted in Figs. 12 and 13 respectively; behave in the same way. We used 17th-order predictor and 256 points Hamming window.

5. Conclusions

In this paper, a new linear prediction method by assuming that the residual signal has an IID t -distribution has been proposed. Theoretically, the new method is very efficient for many kinds of distributions. The run test shows that when we use a small α , this method produces a less biased and smaller SD estimate for formant frequencies and bandwidths. By using small α , a better separation between the source excitation and the vocal tract system can be achieved. When the signal is very spiky, by applying small α we can get more accurate and efficient estimates than that with RBLP method, while when the signal is not very spiky the result from the proposed method and the RBLP are comparable. The CLP method is a special case of the proposed method; when $\alpha = \infty$. For the human speech, the obtained result from the proposed method with small α is less sensitive to the placement of the window than CLP. Also from the proposed method we can get more higher spectral resolution and more spiky residual than RBLP and CLP. Further study still has to be carried out for real application in the human speech analysis. Also to reduce the calculation burden caused by the iteration and to determine an appropriate degree of freedom α .

Acknowledgment

The first author would like to deeply thank all members of the Onoda, Kuneida and Kaneko laboratory who have provided a pleasant working and studying environment. Thanks are also due to Professor Y. Yoshida and Mr. Y. Murahara, Faculty of Science, Sophia University, Tokyo, Japan who have firstly pointed out the problems treated in this paper.

References

- (1) Markel J.D. and Gray, Jr A.H., Linear Prediction of Speech, New York, Springer Verlag (1976).
- (2) Gabor G. and Györfi Z.: "On higher order distribution of speech signal", IEEE Trans. Acoustic Speech Signal Processing, **ASSP-36**, pp. 602-603 (April 1988).
- (3) Huber P.J.: "Robust Statistics", John Wiley and Sons Inc. (1981).
- (4) Lee C.H.: "On robust linear prediction" IEEE Trans. Acoustic Speech and Signal Processing, **ASSP-36**, pp. 642-650 (May 1988).
- (5) Denöel E. and Solvay J.: "Linear prediction of speech with a least absolute error criterion", IEEE Trans. Acoustic Speech Signal Processing, **ASSP-33**, pp. 1397-1403 (Dec. 1985).
- (6) Miyoshi Y., Yamato K., Yanagida M. and Kakusho O.: "Analysis of Speech Signals of short pitch period by sample-selective linear prediction", ICASSP-86, Tokyo, Japan, pp. 1245-1249 (1986).
- (7) Yanagida M. and Kakusho O.: "A weighted Linear Prediction Analysis of Speech Signals by Using the Given's Reduction", IASTED "Applied Signal Processing", M. H. Hamza Ed., pp. 129-132 (1985).
- (8) Murahara Y., Yoshida Y. and Sanubari J.: "Speech Analysis based on AR model Driven by non-Gaussian Process", Proc. of 1988 Japanese National Conference of Acoustic.
- (9) Cramer H.: "Mathematical method of statistic", Princeton University Press, chapter 32 and 33 (1946).
- (10) Saito S. and Nakata K.: "Fundamental of Speech Signal Processing", Academic Press (1985).
- (11) Itakura F.: "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signal", J. Acoustic Soc. Am., **57**, Supplement no. 1, Spring 1975, pp. S. 35 (1975).
- (12) Soong F.K. and Juang B.H.: "Line Spectrum Pair (LSP) and Speech Data Compression", ICASSP-84, San Diego, CA, pp. 1.10.1-1.10.4 (1984).

Appendix A: The Proof of the Positive Definite Property of the Matrix G

From Eq.(21) and Eq.(22) we can write

$$\sum_{i=1}^{M+p} \left\{ \left(\sum_{j=0}^p b_{p-j} s_{i-p+j} \right)^2 w_i \right\} \geq 0. \quad (A \cdot 1)$$

Equation (A·1) can be manipulated to become

$$\sum_{r=0}^p b_r^2 G_{r,r} + 2 \sum_{r=0}^{p-1} \sum_{t=r+1}^p b_r b_t G_{r,t} \geq 0 \quad (A \cdot 2)$$

or

$$\sum_{r=0}^p b_r \sum_{t=0}^p b_t G_{r,t} \geq 0 \quad (A \cdot 3)$$

for any real b_r , $0 \leq r \leq p$, which proves that matrix G is a positive definite matrix.

Appendix B: The Positive Definite Property of the Hessian Matrix at the Stationary Point

The Hessian can be calculated from Eq.(12).

$$\frac{\partial^2 \hat{L}(\mathbf{a})}{\partial \mathbf{a} \partial \mathbf{a}^T} = \mathbf{H}^1 - \mathbf{H}^2$$

$$= \begin{bmatrix} H_{1,1}^1 & H_{1,2}^1 & \cdots & H_{1,p}^1 \\ H_{2,1}^1 & H_{2,2}^1 & \cdots & H_{2,p}^1 \\ \vdots & \vdots & & \vdots \\ H_{p,1}^1 & H_{p,2}^1 & \cdots & H_{p,p}^1 \end{bmatrix} - \begin{bmatrix} H_{1,1}^2 & H_{1,2}^2 & \cdots & H_{1,p}^2 \\ H_{2,1}^2 & H_{2,2}^2 & \cdots & H_{2,p}^2 \\ \vdots & \vdots & & \vdots \\ H_{p,1}^2 & H_{p,2}^2 & \cdots & H_{p,p}^2 \end{bmatrix} \quad (A \cdot 4)$$

where

$$H_{q,z}^1 = \frac{2}{\hat{s}^2 \alpha} \sum_{i=1}^M s_{i-z} s_{i-q} \frac{1}{(1+d_i)^2} \quad (A \cdot 5)$$

$$H_{q,z}^2 = \frac{2}{\bar{s}^2 \alpha} \sum_{i=1}^M s_{i-z} s_{i-q} \frac{d_i}{(1+d_i)^2}, \quad d_i = \frac{(\varepsilon_i / \bar{s})^2}{\alpha} \quad (\text{A} \cdot 6)$$

Since H^1 and H^2 are both positive definite, the inequality

$$H^1 > H^2, \quad (\text{A} \cdot 7)$$

the elements of matrix H^1 are larger than the elements of matrix H^2 , has to be fulfilled to ensure that H is positive definite. The simulation results on many human speech data show that Eq.(A.7) is always satisfied; the ratio between the smallest value of matrix H^1 and the largest value of matrix H^2 is about 10.

Appendix C: The Inverse System Stability

By using the matrix G in Eq.(20) we built the Lyapunov equation

$$-R = Q^T G Q - G \quad (\text{A} \cdot 8)$$

where

$$Q = \begin{bmatrix} -a_1 & -a_2 & \cdots & \cdots & -a_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \quad (\text{A} \cdot 9)$$

is the companion matrix. At the local minima, $\nabla(a) = 0$ so that

$$Ga = -v, \quad v = [G_{1,0} G_{2,0} \cdots G_{p,0}]^T. \quad (\text{A} \cdot 10)$$

By substituting Eq.(A.9) and Eq.(A.10) into Eq.(A.8) we obtain

$$R = K + G^1 - G^2 \quad (\text{A} \cdot 11)$$

$$= \begin{bmatrix} k & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} + \begin{bmatrix} G_{1,1} & \cdots & G_{1,p} \\ \vdots & & \vdots \\ G_{p,1} & \cdots & G_{p,p} \end{bmatrix} - \begin{bmatrix} G_{0,0} & \cdots & G_{0,p-1} \\ \vdots & & \vdots \\ G_{p-1,0} & \cdots & G_{p-1,p-1} \end{bmatrix} \quad (\text{A} \cdot 12)$$

where

$$k = \sum_{i=1}^{M+p} s_i \varepsilon_i w_i \quad (\text{A} \cdot 13)$$

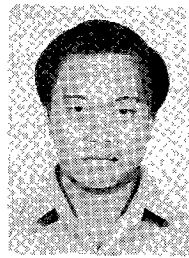
Since

$$k = [1 \quad a^T] \begin{bmatrix} G_{0,0} & v^T \\ v & G \end{bmatrix} \begin{bmatrix} 1 \\ a \end{bmatrix} \quad (\text{A} \cdot 14)$$

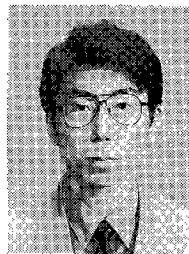
is always positive, K , G^1 and G^2 are positive definite. The inequality

$$G^1 \geq G^2, \quad (\text{A} \cdot 15)$$

all elements of matrix G^1 are larger than the elements of matrix G^2 , has to be fulfilled to ensure that matrix R is positive definite. The simulation results on many human speech data show that Eq.(A.15) is always satisfied; the ratio between the smallest value of G^1 and the largest value of G^2 is about 5. Thus the obtained system is minimum phase.



Junibakti Sanubari was born in Surabaya, Indonesia. He received B.Sc and M. E. E. degree in electronics engineering from Satya Wacana University, Salatiga, Indonesia in 1983 and Eindhoven International Institute, Eindhoven University of Technology, Eindhoven, The Netherlands, in 1986, respectively. He is currently a graduate school student at Tokyo Institute of Technology, where he is working toward the Dr. Eng. degree. His research interests include speech analysis, synthesis and coding.



Keiichi Tokuda was born in Nagoya, Japan, in 1960. He received the B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr. Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986 and 1989, respectively. He is currently a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. His research interests include digital signal processing, speech analysis and coding, and adaptive signal processing. He is a member of IEEE.



Mahoki Onoda was born in Tokyo, Japan on February 22, 1935. He received the BS degree in electronic engineering from Tokyo Institute of Technology in 1957. He has been in Tokyo Institute of Technology since 1957, and now he is a professor of the same university. His research fields include simulation and design of integrated circuits, switched capacitor circuits, etc. He also has great interests in the technology education in

the developing countries.