Analysis and Synthesis of Human Voice Considering the Nonstationary Based on the Glottis Open and Close Characteristics

Hiroyuki KAMATA, Hiroyuki OKA and Yoshihisa ISHIDA

School of Science and Technology,
Meiji University.
1-1-1, Higashi-Mita, Tama-Ku,
Kawasaki-City, 214, Japan.

ABSTRACT - In this paper, we present a new structure of transfer function which is suitable to reconstruct the wave form of voiced speech. In case of voiced speech, there are two different states every pitch period based on the glottis open and close. Thus it is difficult to identify the voice generation system by one transfer function.

In this paper, two transfer functions that are connected as the parallel structure are used. When the compensation of the estimation error is requested, one transfer function is added to the other two transfer functions. By these three transfer functions, the voiced speech of human voice is perfectly reconstructed. These transfer functions are estimated by least mean square (LMS) method.

Furthermore, the method for reduction of voice information is also discussed in this paper.

INTRODUCTION

There are many methods that estimate the vocal tract transfer function of human voice[1]. Specifically, linear prediction (LP) is one of the methods that estimate the vocal tract transfer function of an auto-regressive (AR) model[2] and it is widely used in the area of speech processing.

On the other hand, an auto-regressive and moving-average (ARMA) model estimation is suitable for analyzing the voice that includes nasal and frictional consonants[3,4]. Furthermore, GARMA model estimation that uses the modeled glottis wave as the input signal is effective to reconstruct the voiced speech and it is valid for reduction in the voice information[5,6]. However, these methods often need the iterative calculation. Thus it is difficult to realize the real time processing.

Furthermore, most methods for estimating the transfer function do not consider about the difference of characteristics based on the glottis open and close.

In case of voiced speech, there are two different states every pitch period. In one state, the glottis is open, and the other state, it is closed. When the glottis is closed, the speech signal is generated by the resonance of the vocal tract or nasal tract, mainly. However, when the glottis is open, the influence of the lung is added to the characteristics of the speech signal.

In case of the conventional methods for analyzing the speech signal, the parameters of transfer function are calculated in the domain of the auto-correlation or the spectrum[1]. Thus the glottis open and close characteristics are combined, it becomes difficult to separate the vocal tract characteristic and the excitation from human voice. The GARMA method is the one that is calculated in the time domain, but it does not consider the glottis open and close characteristics.

In this paper, we present a new form of transfer function for estimating the voiced speech[7-9]. Generally, the cascade structure of transfer function is used in the field of speech analysis. In this paper, we use three transfer functions that are connected as the parallel structure. These transfer functions are requested for

(1): representing a part of glottis wave that have non-minimum phase characteristic.

(2): reconstruction the vocal and nasal tract characteristics.

(3): compensation of the estimation error.

These three transfer functions are calculated in the time domain simultaneously. As the method for estimating of transfer functions, the least mean square (LMS) method is applied. When the LMS method is used, the iterative calculation is not needed on this estimation.

Furthermore, we discuss about the voice information reduction. In the proposed method, moving-average (MA) model transfer functions are used for modeling of glottis wave and the estimation error respectively because these waves have non-minimum phase and time-varying characteristics, it is difficult to reconstruct these waves using ARMA transfer function. In this paper, we discuss about the difficulty of the information reduction.

Finally, we demonstrate experimental examples based on the analysis and synthesis of human voice, and show that our method is effective.

DETECTION OF THE NON-STATIONARY CHARACTERISTICS ON THE VOICE GENERATION SYSTEM

In case of voiced speech, there are two different states based on the glottis open and close. The purpose of our study is that these two states are identified as two transfer functions respectively.
Therefore, it is needed to detect the changed point of the state. In this chapter, we try to detect the changed point of two states using the Prony method[10-12].

The Prony method

The Prony method will estimate time series signal $\{\hat{x}(n)\}$ by a linear combination of exponential functions.

$$\hat{x}(n)=\sum_{k=0}^{q-1} h(k)\cdot p(k)^{n} \qquad (1)$$

$$\text{for } 0 \le n \le N-1 \in Z$$

where the constants $h(k)$ and $p(k)$ are defined as

$$h(k)=A(k)\cdot \exp(j\theta(k)) \qquad (2)$$

$$p(k)=\exp\{(\alpha(k)+j2\pi f(k))T\} \qquad (3)$$

T : Sampling interval [sec]
$A(k)$:Amplitude of the complex exponential
$\alpha(k)$: Damping factor [sec^{-1}]
$f(k)$: Sinusoidal frequency [Hz]
$\theta(k)$: Sinusoidal initial phase [rad]

The z-transform of Eq. (1) can be written as

$$\hat{X}(z)=\sum_{k=0}^{q-1}\frac{h(k)}{1-p(k)\cdot z^{-1}} = \frac{\sum_{k=0}^{q-1}\left\{h(k)\cdot \prod_{i=0,i\neq k}^{q-1}\left[1-p(i)\cdot z^{-1}\right]\right\}}{\prod_{k=0}^{q-1}\left[1-p(k)\cdot z^{-1}\right]} \qquad (4)$$

Thus the Prony method can estimate the transfer function of ARMA model. The parameters $\{h(k)\}$ and $\{p(k)\}$ are obtained through the algorithm shown in Fig. 1.
In Fig. 1, the complex amplitude $\{h(k)\}$ is calculated by using time series signal $\{x(n): n=0,1,\cdots,q-1\}$. The estimated transfer function shows the characteristic of the short time signal therefore.
When the parameters obtained by the process are used, Eq. (4) can perfectly reconstruct the human voice for the short time signal. If the system has stationary characteristic, synthetic signals calculated by the estimated transfer function (Eq. (4)) must be similar to progressive real signals (Fig. 2).
We calculate the reconstruction error due to the non-stationary characteristic of signals.

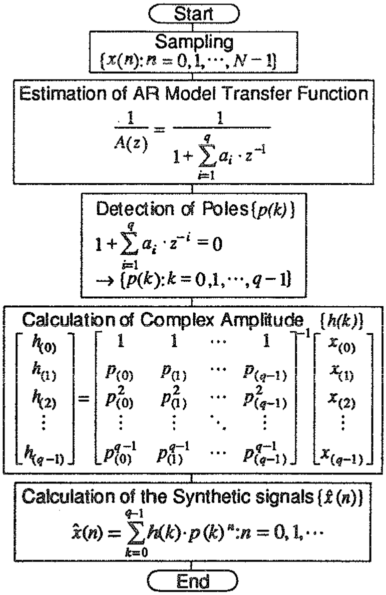

Start
Sampling
$\{x(n): n=0,1,\cdots,N-1\}$
Estimation of AR Model Transfer Function
$\frac{1}{A(z)}=\frac{1}{1+\sum_{i=1}^{q} a_i\cdot z^{-1}}$
Detection of Poles$\{p(k)\}$
$1+\sum_{i=1}^{q} a_i\cdot z^{-i}=0$
$\rightarrow \{p(k): k=0,1,\cdots,q-1\}$
Calculation of Complex Amplitude $\{h(k)\}$
$\begin{bmatrix} h_{(0)} \\ h_{(1)} \\ h_{(2)} \\ \vdots \\ h_{(q-1)} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ p_{(0)} & p_{(1)} & \cdots & p_{(q-1)} \\ p_{(0)}^{2} & p_{(1)}^{2} & \cdots & p_{(q-1)}^{2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{(0)}^{q-1} & p_{(1)}^{q-1} & \cdots & p_{(q-1)}^{q-1} \end{bmatrix}^{-1} \begin{bmatrix} x_{(0)} \\ x_{(1)} \\ x_{(2)} \\ \vdots \\ x_{(q-1)} \end{bmatrix}$
Calculation of the Synthetic signals$\{\hat{x}(n)\}$
$\hat{x}(n)=\sum_{k=0}^{q-1} h(k)\cdot p(k)^{n}: n=0,1,\cdots$
End

Fig. 1 The algorithm of the Prony method.

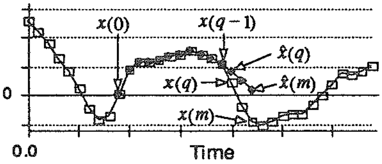


Fig. 2 Difference between the real signals $x(n)$ and synthetic signals $\hat{x}(n)$.

As the calculation method, Eq. (5) is used.

$$\varepsilon=\sqrt{\sum_{n=q}^{m}\{x(n)-\hat{x}(n)\}^{2}} \qquad (5)$$

where

m : The number of samples
q : The order of the Prony method.

Fig. 3 shows the reconstruction error ε of voiced speech. In this figure, there is a point that ε has a peak for each pitch. We have already known that the position of the peak is the same time when the glottis is closed approximately[5]. Thus we can detect the fundamental frequency when the result is used.

ANALYSIS OF VOICED SPEECH

Decision of the analysis area

In case of the proposed method, speech signals are analyzed on each pitch period. As a suggestion, we suppose that the place where the reconstruction error ε indicates the peak is adopted as the beginning point for analyzing and the place of next peak is adopted as the ending point. However, this assumption has some problems as follows:

(A) In the place of the peak of reconstruction error ε, the level of the speech signal has large value. Thus lag window for the time domain will be requested when the signal is analyzed.

(B) The amplitude of speech signal is increasing after the signal decreased. When such signal is analyzed, unstable poles are often estimated.

The voice generation system is a stable system. We have already known that the part of the increasing signal is generated by the glottis wave[8]. As the result, the place which indicates the peak of reconstruction error ε is not suitable for adopting as the beginning and ending points of analysis.

As the other proposal, we adopt the place s_1 and s_2 as the beginning and ending points for analyzing respectively (Fig. 3 (a)). The reasons are as follows:

(1) The place that the speech signal intersected line of zero.

(2) The place that the tendency of the speech signal changes toward increase.

By the proposal, the problem (A) is cleared.

A new transfer function's structure

For solving the problem (B), we propose a new structure of transfer function shown in Fig. 4.

In this Fig. 4, a delay unit $z^{-(p+1)}$ and two transfer functions $G_1(z)$and $G_2(z)$ are used.

$$G_1(z) = \sum_{i=0}^{p} c_i \cdot z^{-i} \qquad (6)$$

$$G_2(z) = \frac{\sum_{i=0}^{r} b_i \cdot z^{-i}}{1 + \sum_{i=1}^{q} a_i \cdot z^{-i}} = \frac{B(z)}{A(z)} \quad q>r \qquad (7)$$

As the transfer function $G_1(z)$, MA model transfer function is used. The reason is as follows:

(1) The impulse response ends in finite time length.

(2) We hope that $G_1(z)$ reconstructs the signal of the area $[s_1, p_1]$, and $G_2(z)$ reappears in the area $[p_1, s_2]$.

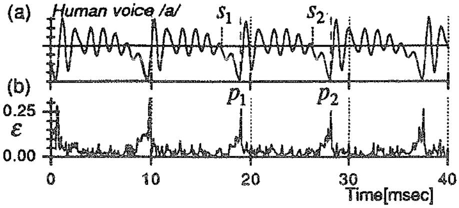


Fig. 3 Reconstruction error ε using the Prony method. (*T=* Frame shift=*0.1* [*msec*], *q=* Frame length =*12* , *m=13*)
(*a*) *Human voice /a/,* (*b*) *Reconstruction error.*

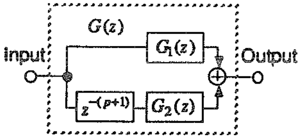


Fig. 4 The basic diagram of the proposed model.

The delay unit $z^{-(p+1)}$ is used to prevent that the impulse response of $G_1(z)$ and $G_2(z)$ are overlapped. Therefore the transfer functions $G_1(z)$ and $G_2(z)$ can be estimated independently in the time domain.

(3) The signal of the area $[s_1, p_1]$ has the tendency toward increase (Problem (B)). If AR or ARMA model is adopted, estimated poles will include the unstable one.

Furthermore, the total transfer function $G(z)$ of Fig. 4 is given as follow:

$$G(z) = G_1(z) + z^{-(p+1)} \cdot G_2(z) \qquad (8)$$

Estimation of the proposed model transfer function

In this paper, the conventional least mean square method is adopted for identification of the proposed model.

When Eq. (8) is reduced to a common denominator, the expression is written as

$$G(z) = \frac{G_1(z) \cdot A(z) + z^{-(p+1)} \cdot B(z)}{A(z)} \qquad (9a)$$

$$\therefore \quad = \frac{\sum_{i=0}^{p+q} d_i \cdot z^{-i}}{1 + \sum_{i=1}^{q} a_i \cdot z^{-i}} = \frac{D(z)}{A(z)} \quad (\because q > r) \qquad (9b)$$

Then, the order of numerator *D(z)* becomes *p+q*. The proposed model needs to estimate a transfer function such that the order of numerator is grater than the denominator's one.

The algorithm of the least mean square method is shown in Fig. 5.

Next, it is necessary to resolve the estimated $D(z)$ into $G_1(z)$ and $B(z)$ by using the recursive process (Fig. 6). If it is needed to calculate the coefficients $\{c_k\}$ and $\{b_k\}$ uniquely, the order q and r must have the relation of $r=q-1$.

EXPERIMENT

In this chapter, we try to estimate the real human voice by using the proposed method. The algorithm of the method is shown in Fig. 7. The method which decides the order is as follows:

(1) The order of $A(z)$ corresponds to the number of formants.

(2) The order of $B(z)$ is q-1.

(3) The order of $G_1(z)$ is determined by the interval between s_1 and p_1.

The experimental result using the proposed method is shown in Fig. 8. In this experimental results,

(a) The estimated coefficients of $G_1(z)$ coincide with the real signals of the area $[s_1, p_1]$ approximately.

(b) In the amplitude characteristic of $G_2(z)$, the first and second formants are obtained precisely. The third and other formants are generated by the addition of $G_1(z)$ and $G_2(z)$.

(c) The synthesized signal based on the estimation result is similar to the real voice. In our hearing experiment, the synthetic voice is obtained as the high quality sound.

Compensation of the estimation error

We think that there are some reasons for producing the error as follows:

(i) The effect of the impulse response of $G_2(z)$ that is estimated on the past analysis area.

(ii) The estimated error by using the least mean square method.

(iii) The non-linear, time-varying and non-stationary characteristics of human voice.

In case of the (i), we can eliminate the effect by the process as follows:

(1) The past impulse response that is generated by past $G_2(z)$ subtracts from the real signal on the area of $[s_1, p_1]$ because $G_1(z)$ and $G_2(z)$ are connected by parallel structure. Of course, the area $[p_1, s_2]$ is not able to subtract.

(2) The least mean square method calculates the transfer function $G(z)$ based on the subtracted signal.

Usually, the inverse system is adopted to eliminate the effect of the past impulse response. However, the inverse system is not necessary when the parallel structure of transfer function is used.

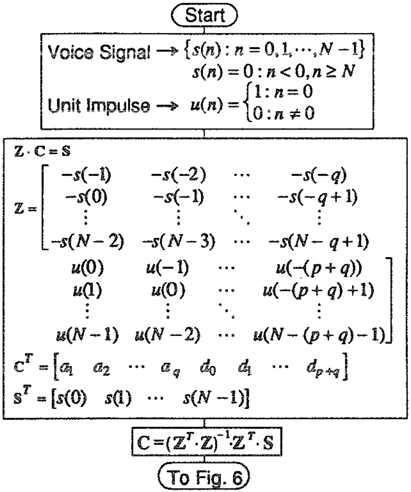

Start
Voice Signal → $\{s(n): n = 0,1,\cdots,N-1\}$
$s(n) = 0 : n < 0, n \geq N$
Unit Impulse → $u(n) = \begin{cases} 1 : n = 0 \\ 0 : n \neq 0 \end{cases}$
$\mathbf{Z} \cdot \mathbf{C} = \mathbf{S}$
$\mathbf{Z} = \begin{bmatrix} -s(-1) & -s(-2) & \cdots & -s(-q) & u(0) & u(-1) & \cdots & u(-(p+q)) \\ -s(0) & -s(-1) & \cdots & -s(-q+1) & u(1) & u(0) & \cdots & u(-(p+q)+1) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -s(N-2) & -s(N-3) & \cdots & -s(N-q+1) & u(N-1) & u(N-2) & \cdots & u(N-(p+q)-1) \end{bmatrix}$
$\mathbf{C}^T = [a_1 \; a_2 \; \cdots \; a_q \; d_0 \; d_1 \; \cdots \; d_{p+q}]$
$\mathbf{S}^T = [s(0) \; s(1) \; \cdots \; s(N-1)]$
$\mathbf{C} = (\mathbf{Z}^T \cdot \mathbf{Z})^{-1} \mathbf{Z}^T \cdot \mathbf{S}$
To Fig. 6

Fig. 5 The algorithm of LMS method.

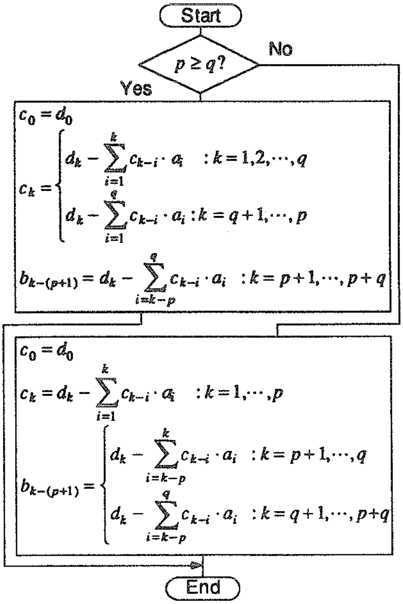

Start
$p \geq q$?
No
Yes
$c_0 = d_0$
$c_k = \begin{cases} d_k - \sum_{i=1}^{k} c_{k-i} \cdot a_i & : k = 1,2,\cdots,q \\ d_k - \sum_{i=1}^{q} c_{k-i} \cdot a_i & : k = q+1,\cdots,p \end{cases}$
$b_{k-(p+1)} = d_k - \sum_{i=k-p}^{q} c_{k-i} \cdot a_i \quad : k = p+1,\cdots,p+q$
$c_0 = d_0$
$c_k = d_k - \sum_{i=1}^{k} c_{k-i} \cdot a_i \quad : k = 1,\cdots,p$
$b_{k-(p+1)} = \begin{cases} d_k - \sum_{i=k-p}^{k} c_{k-i} \cdot a_i & : k = p+1,\cdots,q \\ d_k - \sum_{i=k-p}^{q} c_{k-i} \cdot a_i & : k = q+1,\cdots,p+q \end{cases}$
End

Fig. 6 Resolution the estimated $D(z)$ into $G_1(z)$ and $B(z)$.

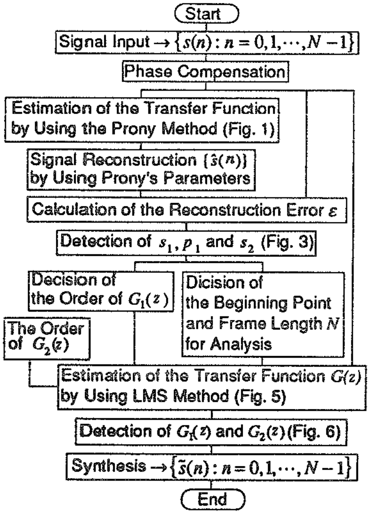
By the property, we try to calculate the error based on (ii) and (iii). The structure for considering the estimation error is shown in

Fig. 9 and experimental result is shown in Fig. 10.

$$E(z)=\sum_{n=0}^{N-(p+1)}\{s(n+p+1)-\tilde{s}(n+p+1)\}\cdot z^{-n} \quad (10)$$

$s(n)$: The real signal.
$\tilde{s}(n)$: Synthesized signal
(is not included the error).
N : Length of pitch period.

In case of this experiment, the synthesized signals are reconstructed perfectly because the estimation error series is used. When information reduction of voice signal is requested, the estimation error must be compress by using any methods.
However, in our study, we have found that the estimation error series has the characteristics like a chaos signal. Therefore it is difficult to compress the error signal by using the transfer function. About the method, we wish to report in the other chance.



Start
Signal Input → $\{s(n): n=0,1,\cdots,N-1\}$
Phase Compensation
Estimation of the Transfer Function by Using the Prony Method (Fig. 1)
Signal Reconstruction $\{\hat{s}(n)\}$ by Using Prony's Parameters
Calculation of the Reconstruction Error ε
Detection of s_1, p_1 and s_2 (Fig. 3)
Decision of the Order of $G_1(z)$
Dicision of the Beginning Point and Frame Length N for Analysis
The Order of $G_2(z)$
Estimation of the Transfer Function $G(z)$ by Using LMS Method (Fig. 5)
Detection of $G_1(z)$ and $G_2(z)$ (Fig. 6)
Synthesis → $\{\tilde{s}(n): n=0,1,\cdots,N-1\}$
End

Fig. 7 The algorithm of the proposed method.

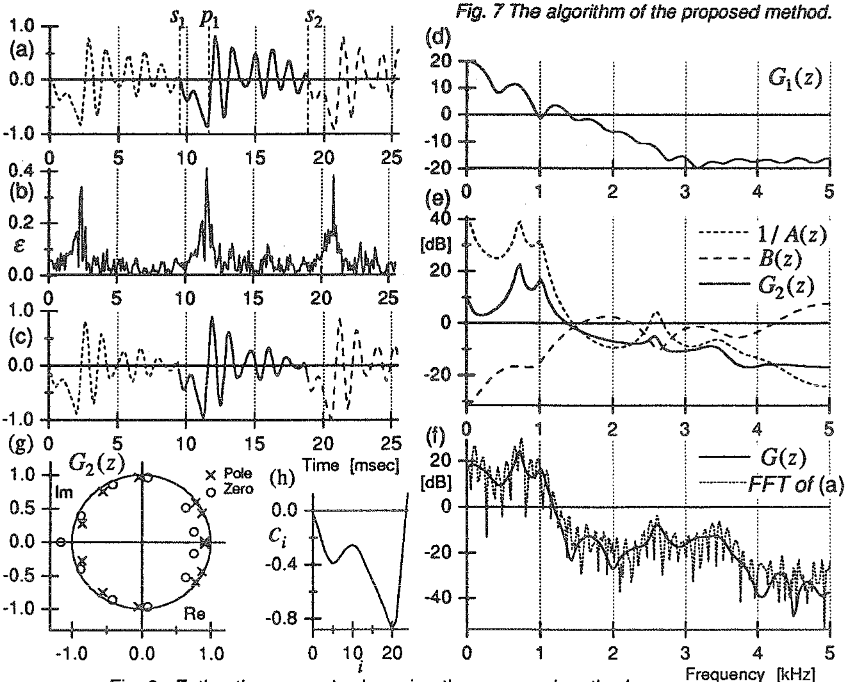


Fig. 8 Estimation examples by using the proposed method.
(T=0.1 [msec], N=90, p=24, r=11, q=12, m=13)
(a) Human voice /a/, (b) Reconstruction error. (c) Synthesized voice. Amplitude characteristics: (d) $G_1(z)$, (e) $G_2(z)$, (f) $G(z)$.
(g) Poles and zeros of $G_2(z)$. (h) Coefficients of $\{c_i\}$.

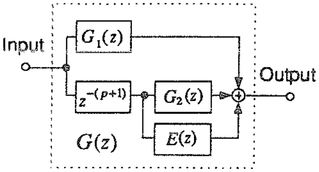


Fig. 9 The diagram of the proposed method considering the estimation error.

CONCLUSION

We have proposed a new model of transfer function for the voiced speech. In this paper, two different type transfer functions for corresponding to non-stationary characteristics of the voice are used and estimated simultaneously. Furthermore the synthesized signal is perfectly reconstructed by adding one transfer function for compensating the estimation error.

However, the proposed method has a problem that it is no good to reduce the voice information because MA model transfer functions are used for recording the estimation error series. And we have not reported about the application to the unvoiced signal. We wish to present about these problems our research again.

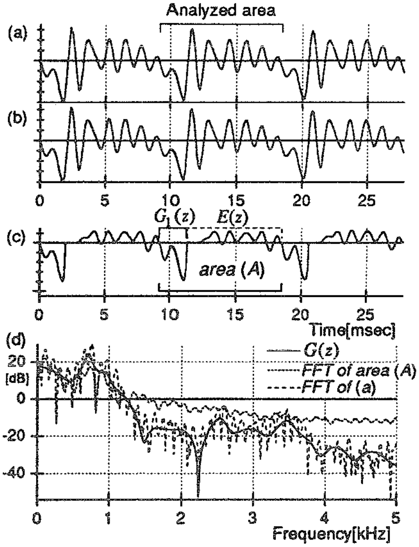


Fig. 10 Experimental examples by the proposed method considering the estimation error and the past impulse response.
(a) Human voice /a/, (b) Synthesized voice.
(c) Output Signal of $G_1(z)$ and $E(z)$.
(d) Amplitude Characteristics.

References

[1] Marple Jr.,S. L. (1987) *Digital Spectral Analysis with Applications*, (Prentice-Hall)

[2] Markel, J. D. & Gray, A. H. (1976) *Linear Prediction of Speech*, (Springer-Verlag)

[3] Morikawa, H. & Fujisaki, H. (1982) "Adaptive Analyze of Speech Based on a Pole-Zero Representation", IEEE Trans. ASSP-30, No. 1,

[4] Grenier, Y. (1983) "Time-Dependent ARMA Modeling of Nonstationary Signals", IEEE Trans. ASSP-31, No. 4

[5] Ljungqvist, M. & Fujisaki, H. (1986) "A Method for Estimating ARMA Parameters of Speech Using a Wave form Model of the Voiced Source", IEICE Technical Report, SP86-49, pp39-45

[6] Seza, K., Tasaki, H. & Takahashi, S. (1992) "Fully Vector Quantized ARMA Analysis Combined with Glottal Model for Low Bit Rate Coding", Proc. ICSLP'92, pp29-32

[7] Kamata, H., Oka, H. & Ishida, Y. (1993) "Estimation of vocal tract transfer function considering the glottis open and close characteristics", Proc. IEEE Pac. Rim."93, pp137-140

[8] Oka, H., Kamata, H. & Ishida, Y. (1994) "A proposition of the transfer function for the voiced speech", IEICE Technical Report, SP93-120, pp17-24

[9] Asano, T., Kamata, H. & Ishida, Y. (1994) "A method for estimation of pitch frequency based on detection of nonstationary on voiced speech", IEICE Technical Report, SP93-119, pp9-16

[10] Castanie, F. & Daymier, E. (1986) "Prony Spectral Analysis of Stationary Processes", IEEE Signal Processing, Vol. 3, pp283-286

[11] Parks, T.W. & Burrus, C. S. (1987) *Digital Filter Design*, pp226-228 (John Wiley and Sons)

[12] Kamata, H. & Ishida, Y. (1992) "A Method to Estimate the Transfer Function of ARMA Model of Speech Wave Using Prony Method and Homomorphic Analysis", Proc. ICSLP'92, pp1649-1652