A TIME VARYING ARMAX SPEECH MODELING WITH PHASE COMPENSATION USING GLOTTAL SOURCE MODEL

Keiichi Funaki, Yoshikazu Miyanaga and Koji Tochinai

Graduate School of Engineering, Hokkaido University
North-13 West-8, Kita-ku, Sapporo, 060, Japan
funaki@hudk.hokudai.ac.jp

ABSTRACT

This paper presents new speech analysis method based on a Glottal-ARMAX (Auto Regressive and Moving Average eXogenous) model with phase compensation. A Glottal-ARMAX model consists of two kinds of inputs: glottal source model excitation and a white gauss input, and a vocal tract ARMAX model. The proposed method can simultaneously estimate the glottal source model and vocal tract ARMAX model parameters pitch synchronously. In this method, ARMAX identification using a modified MIS(Model Identification System) method is adopted to estimate ARMAX parameters, and the hybrid approach of Genetic algorithm(GA) and Simulated annealing(SA) is employed to efficiently solve the non-linear simultaneous optimization of both parameters. Furthermore, phase compensation using an all-pass filter is introduced within a generation loop in the GA method in order to compensate phase distortion. Experiments using synthetic speech and natural speech demonstrate the efficacy of the proposed method.

1. INTRODUCTION

It has been thought that modeling the vocal tract and glottal source separately can improve the naturalness and intelligibility of speech quality in speech synthesis and speech coding. Several pitch synchronous speech analysis methods, which can estimate vocal tract and glottal source model parameters simultaneously, have already been proposed[1][2][3][4]. In these methods, speech production is modeled with a vocal tract AR or ARMA model combined with glottal source excitation, and the models are called a GARMA(Glottal-ARMA) model[1] or an ARX(Auto Regressive eXogenous) model[3]. However, a voiced speech source includes a random noise signal besides glottal source excitation. In order to represent speech production more efficiently and more generally, we introduce a time varying ARMAX(ARMA eXogenous) model combined with glottal source model excitation as well as a white gauss process input(henceforth, Glottal-ARMAX model), and we propose a speech analysis method based on the Glottal-ARMAX model to estimate the glottal source model and vocal tract ARMAX model parameters simultaneously[5]. In the proposed method, ARMAX identification is carried out by the modified version of the MIS method[6] with the assumed glottal source excitation.

The simultaneous estimation of the glottal source and vocal tract model parameters is considered to be a non-linear joint optimization problem that provides multiple local minimums. Optimization approachs such as Hill-climbing(HC) or Simulated annealing have been adopted[1][3] to solve the problem. However, in the HC method, there is a tendency for the estimated value to get stack at a local minimum, and huge computation is required in the SA method to obtain a global minimum. In this paper, a hybrid approach of Genetic algorithm(GA) and Simulated annealing(SA) is employed to solve the non-linear joint optimization problem efficiently in less iterations.

In addition, biased-glottal excitation estimation is inevitable owing to phase distortion at recording, and so on since these speech analysis methods estimate the parameters so as to minimize equation errors in the time-domain. Several phase estimation methods have already been proposed [7][8][9]. In [9], the phase characteristics of speech signal are represented by an all-pass filter, and then the coefficients are determined so as to minimize the errors between synthetic speech driven by the all-pass filtered glottal source excitation and the input speech signal to compensate phase distortion. In this paper, the phase compensation using an all-pass filter is embedded within the Glottal-ARMAX analysis method to compensate phase distortion of speech signal and to improve the accuracy of the estimation of both glottal source excitation and vocal tract characteristics.

This paper is organized as follows. The Glottal-ARMAX speech production model is explained in section 2. The proposed speech analysis method based on the Glottal-ARMAX model with phase compensation using an all-pass filter is presented in section 3. Section 4 describes the experiments using the Glottal ARMAX synthetic speech and natural speech signal.

2. SPEECH PRODUCTION MODEL

The Glottal-ARMAX model shown in Fig.1 and Eq.(1) was introduced in order to represent speech production more accurately and more generally. In Fig.1, the MA model corresponding to glottal excitation, $C(z)$, is called the X(eXogenous) model(X-model) because glottal excitation can be regarded as an exogenous input(X-input).

$$\begin{aligned} y(k) &= h(k)^T \cdot p(k) + u(k) + g(k) + w(n) \\ p(k)^T &= [a(k,1), ..., a(k,l), b(k,1), ..., b(k,m), c(k,1), ..., c(k,n)] \\ h(k)^T &= [-y(k-1), ..., -y(k-l), u(k-1), \\ &\quad ..., u(k-m), g(k-1), ..., g(k-n)] \end{aligned} \tag{1}$$

where $y(k)$, $u(k)$, $g(k)$ and $w(k)$ denote an observed speech signal, an unknown white gauss process input, unknown glottal excitation, and an equation error at time k, respectively; and $a(k,i)$, $b(k,i)$ and $c(k,i)$ are i-th order time-varying AR, MA and X coefficients at time k. l, m, n are filter orders and T denotes transpose.

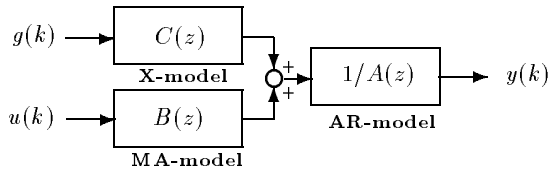


Fig.1 Glottal-ARMAX model

The Glottal-ARMAX model can represent speech production more accurately than conventional models do owing to the introduction of two kinds of inputs, i.e., glottal source excitation and a white gauss process input. Furthermore, an ARMAX model can involve all of the linear systems such as an AR, ARMA, and ARX model. Therefore, the Glottal-ARMAX model can be regarded as a generalized speech production model.

3. SPEECH ANALYSIS METHOD

3.1. Time-varying ARMAX identification

In order to identify the ARMAX model, the MIS method[6] was applied to ARMAX identification with an assumed X-input and an unknown white gauss process input. The block diagram of this identification is shown in Fig.2. In the identification, the ARMAX parameters are estimated so as to minimize the equation error $W(z)$ defined as Eq.(2), which is based on equation error formulation.

$$W(z) = \frac{1}{2\pi j} \oint_{|z|=1} |Y(z)\hat{A}(z) - \hat{B}(z)\hat{U}(z) - \hat{C}(z)\hat{G}(z)|^2 \qquad (2)$$

where $W(z)$, $\hat{U}(z)$, and $\hat{G}(z)$ denote the z-transform of an equation error $\hat{w}(k)$, an estimated white gauss process input $\hat{u}(k)$, and an assumed X-input $\hat{g}(k)$, respectively.

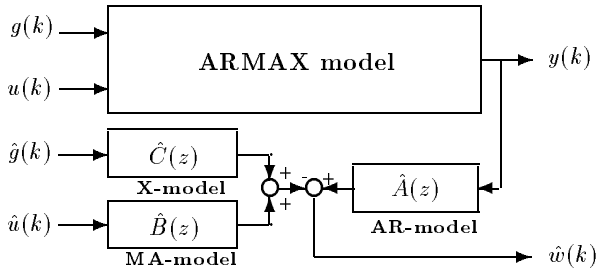


Fig.2 The block diagram of Glottal-ARMAX identification

The algorithm is described in Eq.(3). In the algorithm, the glottal source estimation $\hat{g}(k)$ is the assumed glottal model excitation and the white gauss process input $\hat{u}(k)$ is estimated in the same manner as the MIS method.

$$\begin{aligned}
H(k)^T &= [-y(k-1), ..., -y(k-l), \hat{u}(k-1), \\
&\quad ..., \hat{u}(k-m), \hat{g}(k-1), ..., \hat{g}(k-n)] \\
\hat{p}(k)^T &= [\hat{a}(k,1), ..., \hat{a}(k,l), \hat{b}(k,1), \\
&\quad ..., \hat{b}(k,m), \hat{c}(k,1), ..., \hat{c}(k,n)] \\
\hat{y}(k) &= H(k)^T \hat{p}(k) + \hat{u}(k) + \hat{g}(k) \\
\nu(k|k-1) &= y(k) - H(k)^T \hat{p}(k|k-1) - \hat{g}(k) \\
\hat{p}(k|k-1) &= \hat{p}(k-1) \\
F(k|k-1) &= F(k-1) \\
R(k) &= \frac{F(k|k-1)H(k)}{\lambda(k-1) + H(k)^T F(k|k-1)H(k)} \\
\hat{p}(k) &= \hat{p}(k|k-1) + R(k)\nu(k|k-1) \\
F(k) &= \frac{F(k|k-1) - R(k)H(k)^T F(k|k-1)}{\lambda(k-1)} \\
\hat{u}(k) &= \frac{\nu(k|k-1)}{1 + H(k)^T F(k|k-1)H(k)/\lambda(k-1)}
\end{aligned} \qquad (3)$$

3.2. Glottal source analysis

3.2.1. Glottal source model

Several glottal source models have already been proposed[1][2][11][12][13][14]. In this paper, the Rosenberg-Klatt model(RK-model)[11] was adopted to generate differentiated glottal source excitation because of its relatively easy implementation. The differentiation simulates a lip radiation. The RK-model consists of four parameters: pitch period($T0$), amplitude parameter(AV), open quotient parameter(OQ), and spectral tilt parameter(TL). The speech analysis method estimates the RK-model parameters of AV, OQ and TL, while $T0$ is determined in advance.

3.2.2. Estimation of glottal source parameters

As the simultaneous estimation of glottal source model and vocal tract model parameters gets stuck easily at a local minimum, an optimization approach based on the simulated annualing(SA) has been introduced[3] to get the value to converge a global minimum. In this paper, we adopt an optimization approach based on a hybrid approach of the genetic algorithm(GA)[10] and simulated annealing(SA) to obtain a global minimum with less computation. GA is a populated search algorithm that simulates each individual progress process in one species by means of simple modeling of selection, crossover and mutation. In this paper, the individual(chromosome) is regarded as the glottal model excitation and its genes are regarded as the RK-model parameters(AV,OQ,TL) with floating point precision. It is well known that although the GA method provides better performance of an exploration(global range) search, the performance of an exploitation(local range) search is poor. Therefore, the SA method is combined with the GA method to realize more accurate search. This is called a hybrid approach of GA and SA. In the hybrid approach, the GA method is employed to realize the exploration search, and the SA method with the only optimal individual that gives an optimal fitness is employed to realize the exploitation search. The procedure of the glottal source parameter estimation is described as follows.

(step0)Initialize the genes of each individual. In order to obtain the optimal solution in less iterations, one individual genes are set to those of the optimal individual at the previous pitch period
(step1)In each generation, ARMAX identification using Eq.(3) with the glottal source excitation corresponding to the genes of each individual is carried out, and then the fitness defined as Eq.(4) is calculated.
(step2)The exploitation search is carried out by the SA method with the genes of the optimal individual that gives optimal fitness.
(step3)If the procedure has been operated in a pre-determined number of generations(generation number), quit the procedure.
(step4)The GA operations(selection, one-point crossover and mutation) are carried out
(step5)Return to step1(to next generation).

The exploitation searched optimal individual is protected so as not to be diminished by the GA operations.

$$fitness = \frac{T0}{\sum_{k=0}^{T0-1} \left(y(k) - H(k)^T \hat{p}(k) - \hat{g}(k)\right)^2} \quad (4)$$

3.3. Phase compensation method

The proposed speech analysis method estimates the glottal source model and the ARMAX model parameters so as to minimize equation errors in the time-domain. Therefore, phase distortion due to recoding or phase differences between the speech signal and glottal source excitation make it difficult to accurately estimate the glottal source excitation. The phase characteristics can be modeled with the all-pass filter defined as Eq.(5), and the filter coefficients can be estimated by linear prediction[9]. In Eq.(5), $D(z)$ is a causal minimum polynominal and $D(z^{-1})$ is an uncausal maximum polynominal. In this case, pure delay is neglected. In order to estimate more accurate glottal source excitation, the all-pass filter is determined in every pitch period, and phase distortion is compensated with the inverse all-pass filter[9].

$$\begin{aligned} T(z) &= D(z^{-1})/D(z) \\ D(z) &= 1 + \sum_{i=1}^{q} d(i) z^{-i} \end{aligned} \quad (5)$$

Since the all-pass filter $T(z)$ compensates the phase of glottal source excitation, the all-pass filter can be estimated by solving the following linear equation [9]:

$$\begin{aligned} X \cdot d &= -x \\ (X)_{i,j} &= \sum_{k=0}^{T0-1} (y(k-i) - s(k+i))(y(k-j) - s(k+j)) \\ (x)_j &= \sum_{k=0}^{T0-1} (y(k) - s(k))(y(k-j) - s(k+j)) \\ d^T &= [d(1), d(2), ..., d(q)] \end{aligned} \quad (6)$$

where $y(k)$ and $s(k)$ respectively denote an input speech signal and synthetic speech using an ARX filter with the estimated glottal source excitation.

Phase compensation can be implemented by inverse filtering of $T(z)$ for speech signal. The phase compensated glottal source excitation can be estimated by the Glottal-ARMAX analysis with the phase compensated speech signal $Y(z)/T(z)$. ($Y(z)$ denotes the z-transform of speech signal $y(k)$)

3.4. Analysis flow

The vocal tract ARMAX analysis, glottal source analysis and phase compensation have already been discussed in sections **3.1**, **3.2**, and **3.3**. In this section, total analysis flow of the proposed method is described. Phase compensation is implemented within both the generation loop and the pitch loop of the Glottal-ARMAX analysis method.

(1)A pitch period($T0$) is determined automatically by using the SIFT(Simplified Inverse Filtering Transform) algorithm[15].
(2)The differentiated auto-correlation LPC residuals are estimated by glottal inverse filtering with the differentiated speech signal. Then, a negative peak point(PP) of the LPC residuals, which corresponds to the negative peak point of glottal source excitation, is searched to determine the glottal source excitation position.
(3)Simultaneous estimation of the glottal source model and the ARMAX model parameters is carried out with one pitch period of speech signals extracted by $T0$, PP and each OQ by using the Glottal-ARMAX identification described in **3.1** and **3.2**. The phase compensation described in **3.3** is carried out for each generation of the GA method by using the estimated ARX filter with the glottal source excitation.
(4)The phase compensation described in **3.3** is carried out for each pitch period by using the estimated ARX filter with the glottal source excitation.

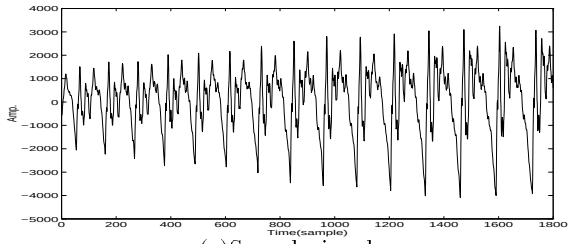
Because of the poor estimation accuracy of PP, the estimation procedure (3) has to be carried out at samples around PP.

4. EXPERIMENTS

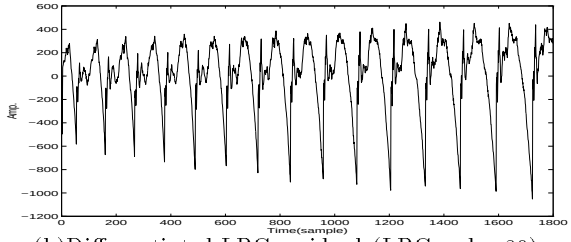
In order to evaluate the estimation accuracy of the proposed method, the ARMAX synthetic signal with the RK-model and the white gauss process input(Ref. Table 1) shown in Fig.3(a) was analyzed using the proposed method. The AR order, MA order, and X order in the synthetic speech were 10,4, and 4, respectively. Analysis conditions were as follows. Analysis AR order, MA order, and X order were 14,8 and 8, respectively, because AR, MA and X orders are set higher than those of a true one in order to estimate the optimal parameters in a small amount of data[6]. In ARMAX identification, $F(-1) = 100I$(Unit matrix), $\hat{p}(-1) = 0$(zero vector), $\lambda(k) = 0.995$. In GA operation, population number, generation number, crossover rate, and mutation rate were 6, 3, 0.7, and 0.01, respectively. TL is fixed at 10[dB] in the experiment. The all-pass filter order was 1. In addition, the pre-emphasis operation with a differencing filter was applied with both speech signal and glottal source model excitation to improve efficiency in the speech analysis, because the glottal source model excitation is approximated by an LPF and it results in poor estimation in the high frequency region. The differentiated auto-correlation LPC residuals(LPC order 20), which determine the glottal excitation position, are shown in Fig.3(b). The estimated glottal excitation(dashed line) and real glottal excitation(solid line) are shown in Fig.3(c). This figure indicates that the proposed method can accurately estimate glottal excitation. These results confirm the efficacy of the proposed method. We also carried out an experiment using natural speech uttered by an adult male speaker. The speech signal was 10kHz sampled Japanese vowel and voiced consonant of /ugeN/, which was converted from a 20KHz sampled ATR database and whose speaker was MYI. The speech signal, differentiated auto-correlation LPC residuals(LPC order 20), and the estimated glottal excitation are shown in Fig.4(a), Fig.4(b), and Fig.4(c), respectively. In the experiment, the analysis orders were 10, 6, and 6 and the other analysis conditions were same as those for synthetic speech. Fig.4(c) shows that the proposed method can accurately estimate glottal excitation. In addition, the phase compensation can also reduce error of the glottal excitation position(PP), and as a result, the phase compensation can reduce a large amount of computation.

5. CONCLUSIONS

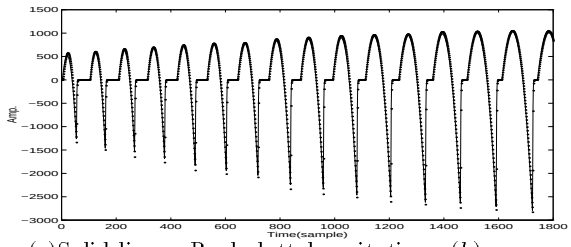
This paper has proposed a new speech analysis method that can effectively estimate glottal source parameters and vocal tract ARMAX parameters simultaneously. The method consists of time-varying adaptive ARMAX identification, glottal source analysis based on the hybrid approach of GA and SA methods, and the phase compensation method with an all-pass filter. Experiments using Glottal-ARMAX synthetic speech and a natural speech signal proved that the proposed method can accurately estimate both the glottal source and ARMAX model parameters.



(a)Speech signal

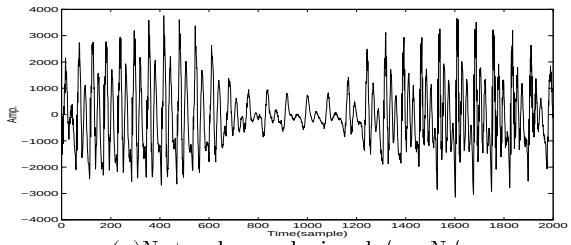


(b)Differentiated LPC residuals(LPC order 20)

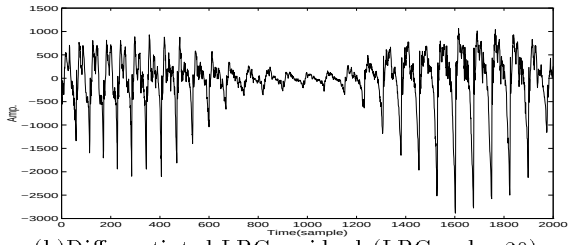


(c)Solid line Real glottal excitation $g(k)$
Dashed line Estimated glottal excitation $\hat{g}(k)$

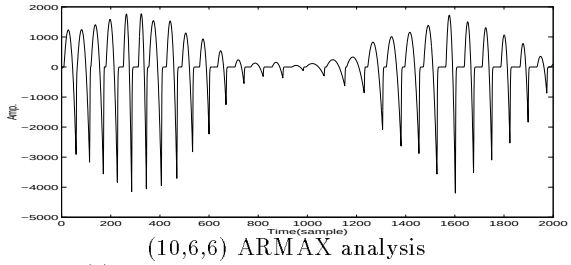
Fig.3 Analysis results for synthetic speech



(a)Natural speech signal /ugeN/



(b)Differentiated LPC residuals(LPC order 20)



(10,6,6) ARMAX analysis

(c)Estimated glottal source excitation

Fig.4 Analysis results for natural speech /ugeN/

ACKNOWLEDGEMENT

The authors would like to thank Dr. M.Hiroshige of Hokkaido University for his valuable discussion regarding this work.

Table 1 Synthetic speech

RK-model parameters

$T0$	100+2*pitch number[sample]
AV	100+20*pitch number
OQ	0.5+0.025*pitch number
TL	10[dB] fixed

White gauss input

mean	0	variance	10

Pole-Zero Frequency/Bandwidth[Hz] of ARMAX filter

	AR	MA	X
No.1	400/80	1500/50	700/50
No.2	1000/60	3400/50	2500/50
No.3	2000/50	-	-
No.4	3000/70	-	-
No.5	3800/100	-	-

REFERENCES

[1] H.Fujisaki and M.Ljungqvist "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform,"IEEE Proc.ICASSP87,pp.637-640,1987.

[2] K.Funaki et.al."A speech analysis method based on a glottal source model," Proc.ICSLP90,pp.45-48,Nov.1990.

[3] W.Ding, H.Kasuya and S.Adachi "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model,"IEICE Trans.Vol.E78-D,No.6,Jun.1995.

[4] P.Hedelin "A glottal LPC-vocoder," IEEE proc. ICASSP84, pp.1.6.1-1.6.4,Mar.1984.

[5] K.Funaki, Y.Miyanaga and K.Tochinai "A time-varying ARMAX speech analysis method based on glottal source model," Proc. 3rd joint-meeting between ASA. and ASJ.,Dec.1996.

[6] Y.Miyanaga, N.Miki and N.Nagai "Adaptive Identification of a time-varying ARMA speech model," IEEE Trans.ASSP-34,pp.423-433,Mar.1986.

[7] H.Kanai,M.Abe and K.Kido,"Accurate autoregressive spectrum estimation at low signal-to-noise ration using a phase matching technique,"IEEE Trans. ASSP,vol.35,pp.1264-1272,Sept.1987.

[8] N.Miyazaki and N.Hamada "An Identification Method for Non-minimum Phase ARMA System Using Phase Equivalent MA System," IEICE Trans. Vol.J78-A,No.7,pp.848-855,Jul.1995.

[9] P.Hedelin "Phase compensation in all-pole speech analysis,"IEEE Proc.ICASSP88,pp.339-342,Apr.1988.

[10] J.Holland,"Adaptation in natural and artificial systems,"The University of Michigan,1975.

[11] D.Klatt and L.Klatt"Analysis, synthesis, and perception of voice quality variations among female and male talkers ," J.Acoust. Soc. America, Vol.87,pp.820-857,Feb.1990.

[12] G.Fant, J.Liljencrants and Q.Lin"A four-parameter model of glottal flow,"STL-QPSR,Vol.4,pp.1-13,1986.

[13] T.V.Ananthapadmanabha "Acoustic analysis of voice source dynamics," STL-QPSR, 2-3/1984, pp.1-24,Oct.1984.

[14] A.E.Rosenberg "Effect of glottal pulse shape on the quality of natural vowels,", J.Acoust.Soc.Am., 49,2,pp.583-588,Feb.1971.

[15] J.D.Markel "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. AU-20, No.5, pp.367-377,Dec.1972.