JAIST Repository
https://dspace.jaist.ac.jp/

Title	Study on Simultaneous Estimation of Glottal Source and Vocal Tract Parameters by ARMAX-LF Model for Speech Analysis/Synthesis
Author(s)	Li, Kai; Unoki, Masashi; Li, Yongwei; Dang, Jianwu; Akagi, Masato
Citation	Proceedings, APSIPA Annual Summit and Conference 2021: 36-43
Issue Date	2021-12
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/18193
Rights	Copyright (C) 2021 APSIPA. This material is posted here with permission of APSIPA (Asia-Pacific Signal and Information Processing Association). Kai Li, Masashi Unoki, Yongwei Li, Jianwu Dang, Masato Akagi, Proceedings of APSIPA Annual Summit and Conference 2021,pp.36-43
Description	13th Asia Pacific Signal and Information Processing Association Annual Summit and Conference 2021 (APSIPA ASC), 14-17 December 2021, Tokyo, Japan

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Study on Simultaneous Estimation of Glottal Source and Vocal Tract Parameters by ARMAX-LF Model for Speech Analysis/Synthesis

Kai Li*, Masashi Unoki*, Yongwei Li†, Jianwu Dang* and Masato Akagi*
* Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: {kai_li, unoki, jdang, akagi}@jaist.ac.jp
† Institute of Automation, Chinese Academy of Sciences, Beijing, China
E-mail: yongwei.li@nlpr.ia.ac.cn

***Abstract*—Correct estimation of glottal source as well as vocal tract parameters is crucial for speech analysis and synthesis. Nearly all methods for estimating these parameters are based on the source-filter assumption. However, the separation and estimation of the source and filter parts are still challenging due to the unreasonable modeling related to physiological processes of speech production or inappropriate estimation procedures. We propose a model that combines the autoregressive moving average exogenous (ARMAX) and Liljencrants-Fant (LF) models, called the ARMAX-LF model, to accurately represent the physiological processes of speech production. The ARMAX model represents the vocal tract as a pole-zero filter with an additional exogenous residual signal, and the LF model represents glottal source waveform as a parametrized time-domain model. Furthermore, we propose a two-stage iterative estimation procedure to separately and simultaneously estimate the parameters of the ARMAX-LF model. The estimated parameters were evaluated objectively and subjectively with synthesized vowels, synthesized consonants, and natural speech. The results indicate that the ARMAX-LF model with the estimated parameters can separately represent the glottal source and vocal tract characteristics and can be widely used in speech analysis and synthesis.**

I. Introduction

Estimating parameters from the glottal source and vocal tract is crucial in many research fields, such as speaker verification [1], [2], speech coding [3], and speech synthesis [4]. The source-filter assumption, which models speech on the basis of exciting a vocal tract filter with a glottal source signal, is one of the most common assumptions for speech-production processes and is extensively applied to speech analysis/synthesis. However, most methods based on the source-filter model cannot separate and estimate parameters of the glottal source and vocal tract effectively due to the unreasonable modeling related to the physiological processes of speech production or inappropriate estimation procedures. For example, the vocal tract is assumed as a single tube that ignores the side branches within the vocal tract when using a linear autoregressive (AR) model. Therefore, parameter-estimation methods of the glottal source model and vocal tract filter using the AR model were extended to obtain more precise parameters using an autoregressive moving average (ARMA) model.

Theoretically, the linear predictive ARMA model estimates the vocal-tract transfer function better than that based on an AR model. Poles and zeros are calculated from the roots of the denominator and numerator polynomial given by the ARMA transfer function. There are many zeros in nasal, fricative, and stop consonants [5], [6]. Zeros can suppress the peaks and flatten the spectrum in the frequency domain of a vocal tract filter. Achieving accurate estimation in this type of speech by a finite number of poles in the all-pole AR model is not easy. With the ARMA model, pole-zero characteristics in the vocal-tract transfer function are assumed. It can provide information on zeros by using a low-order estimation [7]. Accuracy estimation of the ARMA model parameters is still an issue.

The most significant issue with ARMA-model estimation is to decrease the effect of the input excitation on the estimation of vocal tract. Glottal source waveforms possess complex spectral properties. Zeros and poles concealed in the glottal source waveform could be mistakenly interpreted as vocal-tract zeros and poles. Specifically, the estimated spectral envelope will describe the characteristic of the vocal-tract transfer function and contain information of the glottal source waveform. Reasonable modeling on the basis of the source-filter assumption and an effective estimation procedure are necessary to address this issue.

Traditional speech-analysis methods, such as linear prediction (LP) analysis, model the vocal tract and lip radiation in the same AR model while white noise excitation as input is assumed. More reasonable models using ARMA are also used instead of an AR model in LP analysis to obtain a broader range of applications. These methods ignore the complicated spectral characteristics of the glottal source waveform, leading to rough estimation. The estimated poles and zeros are related to the spectrum of input speech not a physiological vocal tract.

A simple and direct method of eliminating the effect of the glottal source waveform on the estimation of vocal tract parameters and obtaining information from the source signal is to estimate the glottal source waveform using inverse filtering. This method has proven to be efficient in the separation of the glottal source waveform and vocal tract. Usually, LP

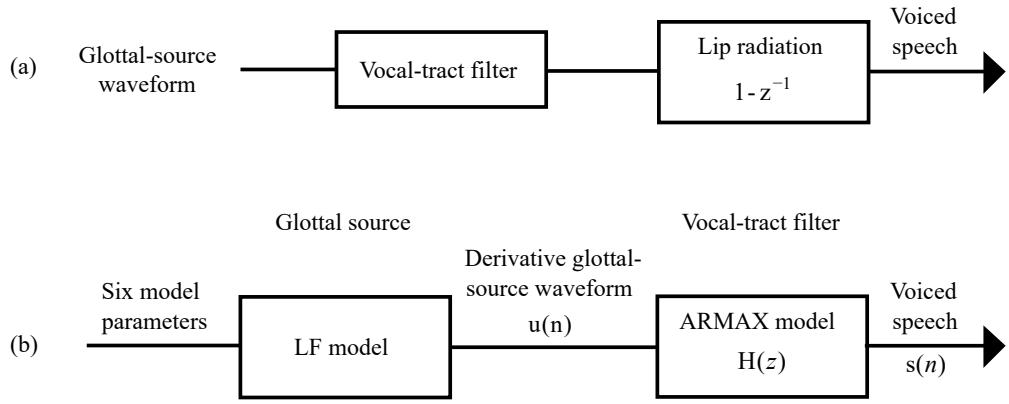


Fig. 1: Source-filter model (a) and its simplification using ARMAX-LF model (b) for voiced speech

analysis [8] incorporating a glottal source model [9], [10], which is simply assumed as a series of pulses for voiced speech or white noise signals for unvoiced speech, is used to obtain the approximate value of a vocal tract filter. These methods, however, are based on a simplistic assumption in the glottal source model, which cannot separate the characteristics of the glottal source waveform and vocal tract filter. More accurate models were proposed to approximate the glottal source waveform or its derivative, such as the Liljencrants-Fant (LF) model [11], the Fujisaki-Ljungqvist (FL) model [12], and the Rosenberg-Klatt (RK) model [13]. However, a problem remains in separating the estimation of the vocal tract filter and parameters of the glottal source models. Furthermore, accurate glottal source models combined with the autoregressive exogenous (ARX) model [14]–[18] on the basis of a joint-optimization process have gained much attention. However, optimizing multiple parameters in the analysis and synthesis stages is still a challenge due to the local optimization problem [19], [20].

Li et al. [21] estimated the parameters of the glottal source model and the vocal tract filter simultaneously on the basis of the ARX model with the LF model (ARX-LF model) using an iterative algorithm under analysis-by-synthesis methodology. With this methodology, the initial values of the LF model are first obtained using an inverse filter method. Accurate glottal source waveforms and vocal tract shapes are then estimated simultaneously on the basis of the ARX-LF model after 2,000 iterations. However, the all-pole autoregressive model in vocal tract modeling cannot always provide accurate pole and zero estimation because of the appearance of zeros. The parameters estimated from the glottal source waveform include zero information from the vocal tract.

We propose an autoregressive moving average exogenous (ARMAX)-LF model, or ARMAX-LF model, to accurately represent the physiological processes of speech production. The ARMAX-LF model represents the vocal tract as a pole-zero filter with an additional exogenous residual signal as an ARMAX model and derivative of the glottal source waveform as an LF model. To correctly estimate glottal source and vocal tract parameters of the ARMAX-LF model, we also propose a two-stage iterative estimation procedure for simultaneously and separately estimating parameters. The results with synthesized vowels, synthesized consonants, and natural speech indicate that the ARMAX-LF model combine with a two-stage estimation procedure can accurately estimate glottal source and vocal tract parameters for both vowels and consonants.

II. ARX-LF Model

As shown in Fig. 1(a), the glottal source waveform, vocal tract, and lip radiation are represented linearly and non-interactively in the linear source-filter model of speech production. The vocal tract and lip radiation filters are commutative since they are linear and time-invariant over short time frames. The effects of these filters can be represented by the derivation of the glottal source waveform. Therefore, we can obtain a simplified form of the source-filter model, as shown in the Fig. 1(b).

Estimating different parts involves coping with a complicated joint-optimization problem. To obtain an accurate estimate from the parameters of the glottal source model and vocal tract filter, the assumption on the physiological process of speech production is crucial. The physiological ARX-LF model represents the derivation of the glottal source waveform by using the LF model and vocal tract filter by using an ARX model. It is a state-of-the-art model for modeling speech production on the basis of the source-filter assumption [16], [21], as it not only has overall adaptability to common speech waveforms but is also flexible enough to represent extreme phonations [11].

A. Glottal source modeled by LF

The LF model proposed by Fant et al. [11] is a parametrized time-domain model for modeling the derivation of glottal source waveforms (glottal airflow). Its parameters can be approximated using inverse filtering from recorded speech. The properties of the LF model have been extensively studied.

In the continuous time domain, a typical period of the derivative of a glottal source waveform modeled using the LF

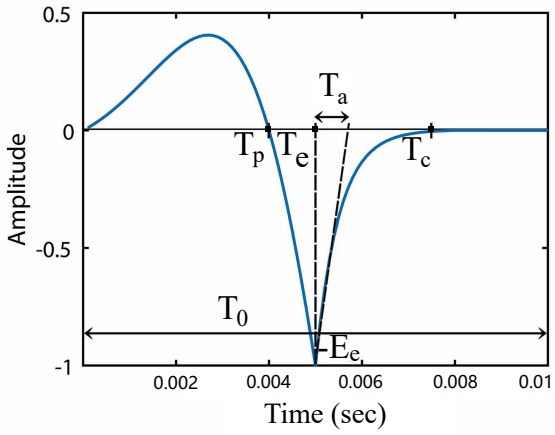


Fig. 2: Typical period of derivative of glottal source waveform represented using LF model.

model is shown in Fig. 2. In the LF model, six parameters (T_0, T_p, T_e, T_a, T_c and E_e) are used to describe the shape of a derivative of a glottal source waveform [11], where T_0 is one period of glottal flow, T_p is the instant of the maximum glottal flow waveform, T_e is the instant of the maximum negative differentiated glottal flow, T_a is the duration of the return phase, T_c is the instant at the complete glottal closure, and E_e is the amplitude at the glottal closure instant. The T_c is often set to T_0 in a simple LF model. Five of these parameters are time dependent (T_0, T_p, T_e, T_a and T_c) and one is amplitude related (E_e).

Assuming the sampling frequency is F_s and one period of glottal flow is T_0, the sampling period is then $T_s = 1/F_s$. The LF model in the discrete-time domain for one fundamental period can be expressed as

$$u(n) = \begin{cases} E_1 e^{\lambda n T_s} \sin(\omega n T_s), & 0 \le nT_s < T_e \\ -E_2[e^{-\mu(nT_s - T_e)} - e^{-\mu(T_c - T_e)}], & T_e \le nT_s < T_c \\ 0, & T_c \le nT_s \le T_0 \end{cases} \tag{1}$$

These direct synthesis parameters $\{E_1, \lambda, \mu, \omega\}$ can be derived with the following constraints [11]:

$$\begin{cases} \sum_{n=1}^{n=N_0} u(n) = 0 \\ \omega = \frac{\pi}{N_p} \\ \mu N_a = 1 - e^{-\mu(N_c - N_e)} \\ E_1 = -\frac{E_e}{e^{\lambda N_e} \sin(\omega N_e)} \\ E_2 = \frac{E_e}{\mu N_a} \end{cases}, \tag{2}$$

where $\{N_0, N_p, N_e, N_a, N_c\}$ are parameters in the discrete-time domain corresponding to $\{T_0, T_p, T_e, T_a, T_c\}$, respectively, and can be derived as

$$\begin{cases} N_0 = \lfloor T_0/T_s \rceil \\ N_p = \lfloor T_p/T_s \rceil \\ N_e = \lfloor T_e/T_s \rceil \\ N_a = \lfloor T_a/T_s \rceil \\ N_c = \lfloor T_c/T_s \rceil \end{cases}, \tag{3}$$

where $\lfloor \cdot \rceil$ denotes the rounding function.

B. ARX model

Given the above assumptions, the vocal tract can be simulated using an ARX model, which combines an all-pole AR model with an additional exogenous LF excitation. In the ARX model, the speech production in the time domain can be represented as

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + b_0 u(n) + e(n), \tag{4}$$

where $s(n)$ is the synthesized speech at time n, $e(n)$ is the error, $a_i, i = 1, ..., p$ are the coefficients of the ARX model, $u(n)$ is the exogenous input to the filter at n generated from the LF model, and b_0 is used to adjust the amplitude of the input.

III. ARMAX-LF Model

The ARMAX-LF model replaces the all-pole model in the ARX-LF model with a pole-zero model. In the ARMAX model, the speech production in the time domain can be represented as

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + \sum_{j=0}^{q} b_j u(n-j) + e(n), \tag{5}$$

where $a_i, i = 1, ..., p$ and $b_j, j = 1, ..., q$ are the coefficients of the ARMAX model. It can also be represented in the z-domain as

$$H(z) = \frac{S(z)}{U(z)} = \frac{\sum_{j=0}^{q} b_j z^{-j}}{\sum_{i=0}^{p} a_i z^{-i}} = \frac{\prod_{j=1}^{q}(1 - \beta_j z^{-1})}{\prod_{i=1}^{p}(1 - \alpha_i z^{-1})}, \tag{6}$$

where $a_0 = 1.0$, α_i and β_j refer to pole and zero in the vocal-tract transfer function and can be derived from a_i and b_j. To estimate a_i and b_j, we transform Eq. (5) into

$$e(n) = s(n) + \sum_{i=1}^{p} a_i s(n-i) - \sum_{j=0}^{q} b_j u(n-j). \tag{7}$$

For convenience, a_0 is set to 1, which transforms Eq. (7) to Eq. (8) and into a matrix form

$$e(n) = \sum_{i=0}^{p} a_i s(n-i) - \sum_{j=0}^{q} b_j u(n-j) \tag{8}$$

$$\mathbf{e} = \mathbf{Sa} - \mathbf{Ub} = \begin{bmatrix} \mathbf{S} \mid -\mathbf{U} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \mathbf{Fh}, \tag{9}$$

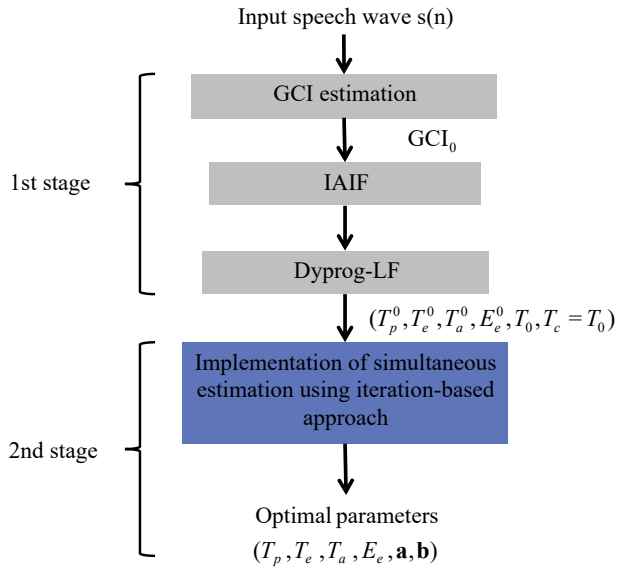


Fig. 3: Estimation scheme of glottal source waveform and vocal tract filter. Detailed implementation processes of stage two is described in Algorithm 1.

where

$$\mathbf{e} = \begin{bmatrix} e(n) \\ e(n-1) \\ \vdots \\ e(n-N+1) \end{bmatrix}, \quad \mathbf{s_i} = \begin{bmatrix} s(n-i) \\ s(n-i-1) \\ \vdots \\ s(n-i-N+1) \end{bmatrix},$$

$$\mathbf{S} = [\mathbf{s_0 s_1} \cdots \mathbf{s_p}], \quad \mathbf{u_j} = \begin{bmatrix} u(n-j) \\ u(n-j-1) \\ \vdots \\ u(n-j-N+1) \end{bmatrix}, \tag{10}$$

$$\mathbf{U} = [\mathbf{u_0 u_1} \cdots \mathbf{u_q}], \quad \mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_q \end{bmatrix},$$

$$\mathbf{F} = \begin{bmatrix} \mathbf{S} \mid -\mathbf{U} \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}.$$

For one period of glottal vibration, N is the number of sampling points in T_0. To obtain the optimal coefficients $\mathbf{h}$ from the ARMAX model, we seek to minimize the mean-square error (MSE) $E(e^T e)$, where $E(\cdot)$ denotes the mathematical expectation. Taking the gradient of $E(e^T e)$ with respect to a_i and b_j and equating to $\mathbf{0_{(p+q+2)\times 1}}$. Then, we can use Wiener-Hopf equation to obtain the optimal solution:

$$\mathbf{h} = -(\mathbf{F^T F})^{-1} \mathbf{F^T s_0}. \tag{11}$$

The estimated coefficients of the ARMAX model can be calculated at the minimum MSE in all iterations.

IV. Proposed Estimation Procedure of Voice Source and Vocal Tract Parameters

In the ARMAX-LF model, the glottal source is modeled by the derivative of the glottal source waveform using the parametrized LF model, and the vocal tract is represented using a pole-zero filter with an additional exogenous residual signal as an ARMAX model. When we estimate the ARMAX model parameters from the recorded speech, the glottal source excitation based on the LF assumption must be known. This estimation problem is a multi-parameter nonlinear joint-optimization problem. Furthermore, the problem of the source-tract interaction in the estimation processes will reduce estimation accuracy since the estimated spectral envelope will describe the characteristic of the vocal tract and contain information of the voice source. Accurate estimation of glottal source and vocal tract parameters is challenging.

As shown in Fig. 3, the estimation process for glottal source and vocal tract parameters is roughly divided into two stages. The first stage is the parameter of LF model initialization. In this stage, the initial value of the LF model (T_p^0, T_e^0, T_a^0, E_e^0, T_0, T_c) is estimated on the basis of advanced techniques in speech analysis. The second stage is the implementation of estimation using the iteration-based approach. In this stage, accurate glottal source and vocal tract parameters are estimated simultaneously and separately.

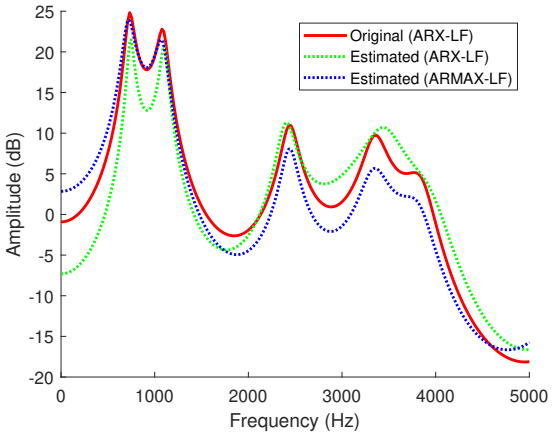
A. LF model initialization

The objective of the first stage is to initialize the parameters of each period of the LF model. Glottal closure instants (GCIs) refer to the instants of significant excitation of the vocal tract. The distance between two continuous GCIs is viewed as one period (T_0). The LF model is used to simulate the excitation signal within each period. GCIs are generally the first parameters for estimation as they can be identified relatively easily. We used the GCI-detection technique called speech event detection using the residual excitation and a mean-based signal (SEDREAMS) [22] to obtain accurate GCIs. The detected GCIs by using the SEDREAMS technique is denoted as GCI_0. As reported by Lu [26], the GCI significantly affects the final estimation. Therefore, to obtain the optimal estimation results, GCI_0 and other four sampling points from the GCI_0 left and right ($k \leq 4$) were assumed as the GCI candidates.

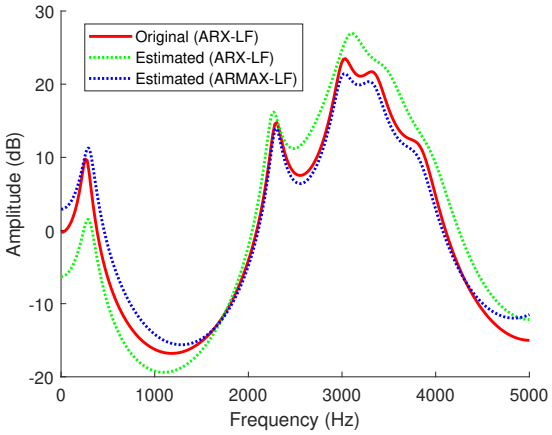
Inverse filtering is widely used in estimating glottal excitation. By using inverse filtering, the glottal source waveform can be obtained by canceling the effects of the vocal tract through the inverse of the transfer function of the vocal tract. It has been proven to be an efficient method in estimating the glottal source waveform. The iterative and adaptive inverse filtering (IAIF) method proposed by Alku [23] has become a representative method in inverse filtering [24]. This method is based on an iterative process between the vocal tract and glottal source to obtain the parameters of the glottal source model. The IAIF method and LF model are used to obtain the initial values of the LF model for the following simultaneous estimation stage. For the LF model fitting, dynamic programming (DyProgLF) proposed by Kane and Gobl [25]

Algorithm 1: Estimation process of the ARMAX model coefficients and glottal source parameters

Initialization: $p \leftarrow 14$, $q \leftarrow 6$, T_p^0, T_e^0, T_a^0, E_e^0, T_0, $T_c \leftarrow T_0$, GCI_0, $j \leftarrow 1$, $l \leftarrow 1$, $k \leftarrow 0$, $GCI \leftarrow GCI_0 - 2$, $MMSE \leftarrow 100$;
Input: Speech waveform *s(n)*;
while $j \leq length(GCI) - 1$ **do**
  while $k \leq 4$ **do**
    $s(n) = s(GCI(j) : GCI(j+1))$;
    while $l \leq 2000$ **do**
      generate LF waveform u(n) by Eq. (1);
      estimate coefficients ***a*** *and* ***b*** *by Eq. (11);*
      $x(n)$=*FILTER(****b****,* ***a****,* $u(n)$*);*
      $e(n)$=*FILTER(****a****,* ***b****,* $s(n) - x(n)$*);*
      MSE = $E\{e^T e\}$;
      if *MSE* <*MMSE* **then**
        MMSE = MSE;
        save T_p, T_e, T_a, E_e, ***a*** *and* ***b****;*
      end
      regenerate T_p^0, T_e^0 *and* T_a^0 *randomly around initial values;*
      l=l+1;
    end
    k=k+1;
    GCI = GCI+*k* ;
  end
  select optimal MMSE;
  j=j+1;
end
Output: *Estimated values of* T_p, T_e, T_a, E_e, **a**=$\begin{bmatrix} a_1 & a_2 & \dots & a_p \end{bmatrix}^T$ *and* $\mathbf{b} = \begin{bmatrix} b_0 & b_1 & \dots & b_q \end{bmatrix}^{\mathbf{T}}$;



(a) Vowel /a/ (poles: 730/90, 1090/110, 2440/170, 3350/250, and 3850/300)



(b) Vowel /i/ (poles: 270/90, 2290/110, 3010/170, 3350/250, and 3850/300)

Fig. 4: Comparison of ARX-LF and ARMAX-LF models in spectra estimated from synthetic vowels. These synthetic vowels were originally synthesized using ARX-LF model (p = 14 and q = 6 at F0 = 120 Hz).

is used due to its robustness. The estimated parameters of the LF model are denoted as T_p^0, T_e^0, T_a^0, and E_e^0.

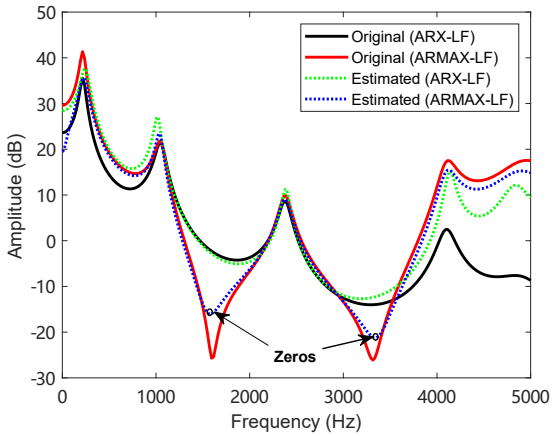
B. Implementation of simultaneous estimation using iteration-based approach

In the second stage, the LF-model parameters and vocal tract coefficients, including poles and zeros, are estimated simultaneously and separately with an iterative algorithm based on the proposed ARMAX-LF model. Detailed implementation processes can be found in Algorithm 1. In this algorithm, j is used to traverse all the period of derivative of glottal source waveform, k is the GCI candidate, and l is the iteration time. First, the LF model parameters obtained in the initialization stage are used to generate the derivative glottal-source waveform $u(n)$. Then, the coefficients of the ARMAX model (**a** and **b**) are estimated on the basis of the MSE. The $u(n)$ and estimated vocal tract parameters are then used to synthesize $x(n)$ using a one-dimensional digital filter shown in Eq. (5). The estimation error in the derivative glottal source is then calculated using the error inverse filtering method with $s(n) - x(n)$ as input. In each iteration, the LF model parameters are randomly generated around the initial value. Then, a new glottal source derivative is generated from these parameters to obtain a new estimation in the vocal tract. This method can avoid falling into local optimization with increased computation cost.

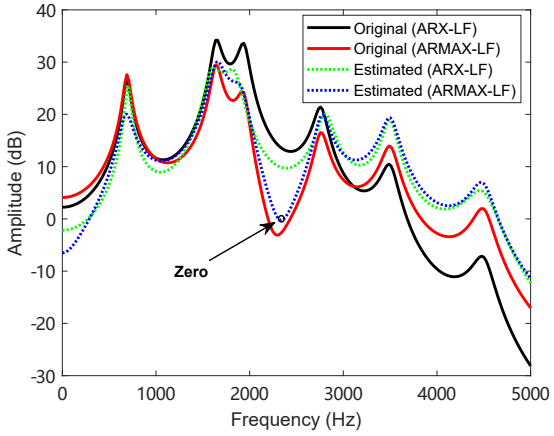
In this study, the sampling frequency was 12,000 Hz, and the orders of the poles and the zeros were set to 14 and 6, respectively. To control variables and indicate the superiority of the ARMAX-LF model, the length of the vocal tract was set to a fixed value of 17.35 cm, and the length of the synthesized speech was set to 1 s. In each iteration, estimated parameters can be calculated using Eq. (11). After 10,000 iterations, the accurate glottal source and vocal tract filter parameters are finally estimated in the period with the smallest minimum

TABLE I: Averaged error rates [%] for parameters of ARMAX-WN, ARX-LF, and ARMAX-LF models with synthesized vowels and consonants. Averaged error rate represented as '/' means corresponding model could not estimate this parameter. 'Average' means average error rate calculated from poles as well as zeros by using Eq. (12).

		Glottal source					Vocal tract		
		T_p	T_e	T_a	T_c	E_e	Poles	Zeros	Average
Vowels	ARMAX-WN	/	/	/	/	/	2.15	/	2.15
	ARX-LF	8.78	8.24	126.73	64.99	24.37	1.78	/	1.78
	ARMAX-LF	15.44	15.72	93.57	64.64	23.10	1.80	/	1.80
Consonants	ARMAX-WN	/	/	/	/	/	4.06	6.44	4.44
	ARX-LF	13.83	13.27	167.97	65.62	26.64	3.57	/	3.57
	ARMAX-LF	14.03	14.54	166.73	65.43	25.48	2.20	6.41	2.95



(a) Nasalized consonant /m/ (poles: 220/60, 1050/100, 2380/120, and 4100/180; zeros: 1600/70 and 3320/130)



(b) Nasalized consonant /$\widetilde{\varepsilon}$/ (poles: 690/70, 1640/100, 1940/110, 2760/130, 3500/160, and 4500/200; zeros: 2260/250)

Fig. 5: Comparison of ARX-LF and ARMAX-LF models in spectra estimated from synthetic nasalized consonants. These synthetic nasalized consonants were originally synthesized using ARX-LF or ARMAX-LF model (p = 14 and q = 6 at F0 = 120 Hz).

MSE. Poles and zeros are calculated from the roots of the denominator and numerator polynomial given by the ARMAX transfer function.

V. Results and Discussion

In this section, we confirm whether the ARMAX-LF model with the proposed estimation procedure can correctly represent synthesized vowels, synthesized consonants, and natural speech. Synthesized vowels (/a/, /e/, /i/, /o/, and /u/) were obtained using Kawahara's method [27]. To obtain synthesized consonants, we replaced the AR model in Kawahara's method with an ARMA filter in the vocal tract modeling to synthesize nasalized consonants such as /m/ and /$\widetilde{\varepsilon}$/. Moreover, we statistically evaluated the estimation accuracy in each period of natural speech (/m/). We also implemented the ARMAX and white noise (WN) models, the combination of which is called the ARMAX-WN model, with our proposed two-stage estimation procedure as a comparison experiment to show the superiority of the ARMAX-LF model. In the ARMAX-WN model, the glottal source signal is represented as white noise, and the vocal tract is represented as a pole-zero filter by using the ARMAX model. Note that we cannot obtain the real parameters in natural speech. Therefore, it is difficult to evaluate the estimation accuracy by creating a large database.

A. Results on synthesized speech

The spectra of the vocal tract transfer function estimated from synthesized vowels (/a/, /i/) and nasalized speech (/m/, /$\widetilde{\varepsilon}$/) are shown in Fig. 4 and Fig. 5, respectively. The frequency/bandwidth of each synthesized speech is shown in the corresponding subfigure. With synthesized vowels, the ARMAX-LF model estimated spectra of the vocal-tract transfer function similarly to or better than the ARX-LF model. For synthesized consonants, the ARMAX-LF model could clearly estimate the zeros and was very close to the spectra from synthetic speech.

The difference in estimating vocal tract parameters increases as the fundamental frequency F0 increases. To evaluate the accuracy in different F0, a large number of vowels and consonants were synthesized with 45 F_0s, varying from 80 to 300 Hz. The average error rate ($\bar{\epsilon}$) calculated using Eq. (12) was used to evaluate the distance between the reference and estimation values.

TABLE II: Performance of ARMAX-LF model with natural speech /m/. M and SD refer to mean and standard deviation of estimated values respectively. P_n and Z_n refer to nth pole and zero, respectively. Unit: Hz.

	Utterance 1		Utterance 2		Utterance 3		Utterance 4		Utterance 5		Mean	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
P_1	232.9	10.06	227.0	10.52	224.2	8.65	217.4	6.24	220.1	6.29	224.3	8.35
P_2	1403.4	51.10	1428.6	77.94	1403.7	70.26	1265.6	77.16	1210.8	79.74	1342.4	71.24
P_3	2238.0	23.78	2233.6	70.31	2230.9	43.60	2118.0	77.78	2228.5	29.99	2209.8	49.09
P_4	3181.2	27.94	3208.5	48.42	3185.5	50.35	3009.5	42.78	3135.0	76.72	3143.9	49.24
P_5	4193.1	46.05	4234.6	65.20	4259.4	39.42	4212.6	72.18	4223.0	98.02	4224.5	64.17
Z_1	679.7	202.26	747.6	207.91	738.2	233.23	753.5	130.03	473.9	321.96	678.6	219.08
Z_2	2597.6	326.82	2544.2	317.46	2393.2	336.86	2381.6	357.03	2948.2	329.25	2573.0	339.48
Z_3	5287.9	560.83	5289.5	456.43	5110.0	558.99	5141.4	425.53	5160.5	580.06	5197.9	516.37

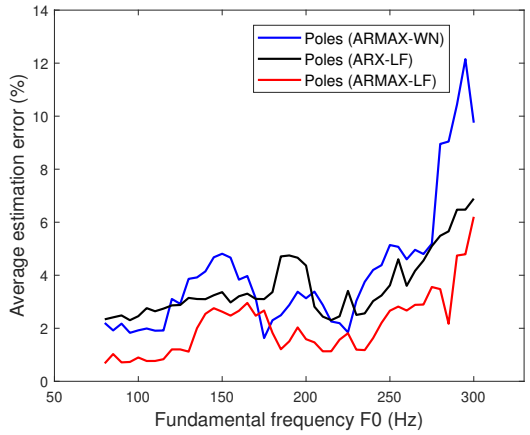


Fig. 6: Averaged estimation error rate of poles by using three methods on basis of synthesized consonants, in which F0 ranged from 80 to 300 Hz.

$$\bar{\epsilon} = \frac{\sum_{i=1}^{i=\chi} \frac{|\widehat{\xi}-\xi|}{\xi}}{\chi} \times 100, \tag{12}$$

where $\widehat{\xi}$ refers to the estimated value, ξ refers to the true value, and χ is the total number of estimated points. As shown in Table I, the ARMAX-LF model's performance was superior to the ARMAX-WN model. When comparing the estimation errors of the ARX-LF and ARMAX-LF models in the glottal source parameters of vowels, the ARMAX-LF model performed relatively poorly, with 15.44 and 15.72% in T_p and T_e, respectively. However, the T_a, T_c and E_e with the ARMAX-LF model were 93.57, 64.64 and 23.10%, respectively, which are better than those with the ARX-LF model. The estimation error comparison of the ARX-LF and ARMAX-LF models in glottal source parameters of consonants was similar to that in vowels. Regarding the estimation of the vocal tract parameters, the $\bar{\epsilon}$ of the ARMAX-LF model with vowels (1.80%) was similar to that of the ARX-LF model (1.78%). However, with consonants, the $\bar{\epsilon}$ of the ARMAX-LF model were 2.20 and 6.41% in the separate estimations of poles and zeros, respectively. The error rate of the ARMAX-LF model (2.95%) was better than that of the ARX-LF (4.29%) on average. In summary, compared with the ARX-LF model, our ARMAX-LF model achieved a comparable performance in glottal-source-parameters estimation and superior performance in vocal-tract-parameter estimation.

To demonstrate the superiority of the ARMAX-LF model, the comparison of the averaged estimation error rate of poles with the three different methods is also illustrated in Fig. 6. These results are based on synthesized consonants with different F_0 ranging from 80 to 300 Hz. Fig. 6 show that the degradation occurred with the increase of F_0. However, our ARMAX-LF model was better in pole estimation (red line) than the ARX-LF model (black line) and ARMAX-WN model (blue line). These results indicate that our ARMAX-LF model is superior to the ARX-LF and ARMAX-WN models in vocal-tract-parameter estimation.

B. Results on natural speech

To evaluate the accuracy of our ARMAX-LF model with natural speech, five utterances of natural speech data (/m/) were recorded in a soundproof room of the JAIST AIS laboratory using Audacity software. All the utterances were uttered by the same speaker at a 44,100-Hz sampling rate in a 16-bit, mono-quality format. The duration of each utterance was restricted to 2 s. To maintain stability of pronunciation, a 1-s utterance was selected from the middle part of each utterance. The mean and standard deviation (SD) of estimated poles and zeros for each utterance were calculated. Five poles and three zeros of each speech were demonstrated, as shown in Table II.

The results indicate that the SDs in the zero estimation are obviously larger than in the pole estimation, meaning zero estimation is much more difficult. Moreover, as the frequencies of the zeros and poles increase, the SDs also increase. This phenomenon suggests that estimation difficulty increases as the frequency increases because high frequencies are more dependent on small vocal tract details than low frequencies. Thus, our ARMAX-LF model can achieve acceptable accuracy in natural speech.

VI. Summary

We proposed the ARMAX-LF model to accurately represent the physiological processes of speech production. In the

ARMAX-LF model, the glottal source signal is represented as the derivative of the glottal source waveform by using the LF model and the vocal tract is represented as a pole-zero filter by using the ARMAX model. In contrast to the conventional ARMA model, the ARMAX model represents the vocal tract as an ARMA filter with an additional exogenous residual signal, which is more reasonable for modeling speech production. Furthermore, to decrease the effect of the input excitation on the estimation of vocal tract characteristics and accurately estimate parameters of the ARMAX model as well as parameters of the LF model, we also proposed a two-stage iterative estimation procedure to estimate the parameters of these two models simultaneously and separately.

Vocal tract characteristics estimated from synthesized vowels and consonants were used to show the estimation accuracy of our ARMAX-LF model. The average error rate was statistically calculated from a large amount of synthesized speech and natural speech. From the results of estimation accuracy, our ARMAX-LF model can accurately estimate poles and zeros in the vocal-tract transfer function. These results also indicate that the model can separately and accurately estimate the parameters of the glottal source model and vocal tract filter due to the reasonable modeling related to the physiological processes of speech production and appropriate estimation procedures. Furthermore, the estimated glottal source and vocal tract parameters can support important information for speech analysis/synthesis as well as speaker recognition.

VII. Acknowledgements

This work was supported by JSPS-NSFC Bilateral Joint Research Projects/Seminars (JSJSBP120197416), a Grant-in-Aid for Scientific Research (Grant number: 20H04207).

References

[1] P. Rose,Forensic speaker identification. cRc Press, 2002.

[2] E. Enzinger, P. Balazs, D. Marelli, and T. Becker, "A logarithmic based pole-zero vocal tract model estimation for speaker verification," in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 4820–4823, 2011.

[3] G. Fant, Acoustic theory of speech production. Walter de Gruyter, no. 2, 1970.

[4] E. Keller, Fundamentals of speech synthesis and speech recognition: basic concepts, state-of-the-art and future challenges. John Wiley and Sons Ltd., 1995.

[5] H. Morikawa and H. Fujisaki, "Adaptive analysis of speech based on a pole-zero representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 1, pp. 77–88, 1982.

[6] N. Ouaaline and L. Radouane, "Pole–zero estimation of speech signal based on zero-tracking algorithm," *International Journal of Adaptive Control and Signal Processing*, vol. 12, no. 1, pp. 1–12, 1998.

[7] J. Cadzow, "High performance spectral estimation–a new arma method," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 524–529, 1980.

[8] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[9] Y. Grenier, B.-G. Lee, I. Song, and S. Ann, "Robust estimation of AR parameters and its application for speech enhancement," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, pp. 309–312, 1992.

[10] L. R. Rabiner, "Digital processing of speech signal," *Digital Processing of Speech Signal*, 1978.

[11] G. Fant, J. Liljencrants, and Q.-g. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[12] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11, IEEE, pp. 1605–1608, 1986.

[13] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990.

[14] W. Ding, H. Kasuya, and S. Adachi, "Simultaneous estimation of vocal tract and voice source parameters based on an arx model," *IEICE transactions on information and systems*, vol. 78, no. 6, pp. 738–743, 1995.

[15] 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," in *Recent Research Towards Advanced Man-machine Interface Through Spoken Language*. Elsevier, pp. 52–60, 1996.

[16] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of LF glottal source parameters based on an ARX model," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[17] Y. Li, K.-I. Sakakibara, D. Morikawa, and M. Akagi, "Commonalities of glottal sources and vocal tract shapes among speakers in emotional speech," in International Seminar on Speech Production. Springer, pp. 24–34, 2017.

[18] K. Takahashi and M. Akagi, "Estimation of glottal source waveforms and vocal tract shape for singing voices with wide frequency range," in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, pp. 1879–1887, 2018.

[19] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 492–501, 2006.

[20] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, vol. 2, no. 3, p. 40, 1995.

[21] Y. Li, K.-I. Sakakibara, and M. Akagi, "Simultaneous estimation of glottal source waveforms and vocal tract shapes from speech signals based on ARX-LF model," *Journal of Signal Processing Systems*, pp. 1–8, 2019.

[22] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2011.

[23] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[24] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech Language*, vol. 26, no. 1, pp. 20–34, 2012.

[25] J. Kane and C. Gobl, "Automating manual user strategies for precise voice source analysis," *Speech Communication*, vol. 55, no. 3, pp. 397–414, 2013.

[26] H.-L. Lu,Toward a high-quality singing synthesizer with vocal texture control, Stanford University, 2002.

[27] H. Kawahara, "SparkNG: Interactive MATLAB tools for Introduction to Speech Production, Perception and Processing Fundamentals and Application of the Aliasing-Free LF Model Component." in *INTERSPEECH*, pp. 1180–1181, 2016.