F_0-noise-robust glottal source and vocal tract analysis based on ARX-LF model

Yongwei Li, Jianhua Tao, *Senior, IEEE,* Donna Erickson *Member, IEEE,* Bin Liu *Member, IEEE,* and Masato Akagi, *Member, IEEE*

***Abstract*—This paper proposes a robust automatic speech analysis method based on a source-filter model constructed of an Auto-Regressive eXogenous (ARX) model and the Liljencrants-Fant (LF) model. The proposed method estimates glottal source waveform and vocal tract shape parameters using an analysis-by-synthesis approach. Structurally, the first step is to initialize the glottal source parameters using the inverse filter method, and the second step is to simultaneously estimate the glottal source waveform and the vocal tract shape parameters using an analysis-by-synthesis approach with an iterative algorithm. The proposed method was verified on synthetic voices with different glottal noise (signal to noise ratio) from 0 dB to 50 dB and different fundamental frequency (F_0) from 80 Hz to 320 Hz levels. The results show that the proposed method achieved a much higher estimation accuracy than that of the state-of-the-art inverse filtering methods on both different glottal noise and different F_0 levels.**

***Index Terms*—Glottal source, vocal tract, source-filter model, ARX-LF model.**

I. Introduction

THE separation of glottal source and vocal tract filter from speech signals plays an important role in understanding speech production mechanisms. Glottal source and vocal tract cues are frequently used for many speech technologies with applications to speech recognition [1], speech synthesis [2], speech conversion [3], detection of language impairment [4], pathological voice detection [5], dysphonic voice analysis [6], speech emotion recognition [7], and speaker identification [8].

Manuscript received December xx, 2020; revised xxx, 2021; accepted xxx, 2021; Date of publication xxx, 2021. This work was supported in part by the National Key Research & Development Plan of China (No.2017YFC0820602), in part by the National Natural Science Foundation of China (NSFC) (No.61831022, No.61771472, No.61773379, No.61901473), and in part by the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300).The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Xie.
(Corresponding author: Jianhua Tao.)

Yongwei Li is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100101, China, E-mail: (yongwei.li@nlpr.ia.ac.cn).

Jianhua Tao is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100101, China, with the School of Artificial Intelligence, University of Chinese Academy of Sciences Beijing, 100101, China, and also with the Center for Excellence in Brain Science and Intelligence, Chinese Academy of Sciences, Beijing, 100101, China. E-mail: (jhtao@nlpr.ia.ac.cn).

Donna Erickson is with the Haskins Laboratories, U. S. A, E-mail: (ericksondonna2000@gmail.com).

Bin Liu is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100101, China, E-mail: (liubin@nlpr.ia.ac.cn).

Masato Akagi is with the Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Japan, E-mail: (akagi@jaist.ac.jp).

With the help of high-speed videoendoscopy [9] and magnetic resonance imaging (MRI) [10], the glottal source and vocal tract can be measured directly and somewhat accurately when uttering a speech sound. However, it is difficult to measure glottal source and vocal tract simultaneously on an utterance, and it is not always convenient to use such measuring equipment.

Based on the source-filter theory of speech production, the speech signal is represented by the output signal of a linear vocal tract filter with a glottal source excitation signal. Studies of separating glottal source and vocal tract from the speech signal based on the source-filter model have been going on for decades [11].

The earliest study for estimating vocal tract filters is linear prediction (LP) analysis [12]. It assumes the vocal tract filter as an auto-regressive model and its coefficients can be estimated from the speech signal. The LP residual signal is considered as the glottal source, and the frequency of periodic impulse in the LP residual signal is considered the fundamental frequency (F_0). However, the main problem of this method is the difficulty of removing the glottal source from the speech signal when estimating the vocal tract filter. To avoid the glottal source effects, vocal tract filters were estimated within the glottal closed phase, e.g., Wong $et\ al.$ [13] estimated vocal tract filters during glottal closed phase with LP analysis (CPLP). Yegnanarayana $et\ al.$ [14] estimated vocal tract characteristics using a closed phase inverse filter (CPIF), since no glottal source waveform occurs during the glottal closed phase. Although this solution can accurately estimate vocal tract filters in a prolonged glottal closed phase, it fails in the case of speech with short glottal source closed phases, which frequently happens in real conditions, such as female speech and aroused speech, where the F_0 is high, and the glottal period is short.

A straightforward method for estimating glottal source waveform is to process the speech signal using inverse filtering, such as CPLP, CPIF, and iterative and adaptive inverse filtering (IAIF) [15], where glottal sources can be considered as the residual signal or periodic pulse for voiced speech. However, these methods faced a fundamental problem that the oversimplified glottal source assumption could not describe the complex glottal source waveform. A more effective method is to process the residual signal of inverse filtering by parametric glottal source models [16], [17], such as the Liljencrants-Fant (LF) model [11], the Fujisaki-Ljungqvist (FL) model [18], and the Rosenberg-Klatt (RK) model [19]. The commonality of these glottal source models is the time-domain description of the glottal source waveform, whereas the difference is the number of parameters. These models provided a more appropriate assumption for describing the glottal source waveform. However, the source-filter interaction still remained since the vocal tract filter was estimated firstly to fit the residual signal by using glottal source models.

The most complete assumption is to separate the glottal source and vocal tract parameters in a simultaneous manner, which would reduce the source-filter interaction [20]. However, it is difficult to simultaneously optimize multiple parameters of glottal source and vocal tract. To solve this problem, Funaki $et\ al.$ [21] presented a hybrid approach using a genetic algorithm and simulated annealing to optimize multiple parameters of the glottal source waveform with the RK model and parameters of the vocal tract filter with an auto regressive and moving average exogenous (ARMAX) model. Fu $et\ al.$ [22] presented a two step strategy for optimization, in which a simplified glottal source model (RK model) was used to estimate the initial values for a more complex glottal source model (LF model), and then the auto-regressive exogenous (ARX) model, as the vocal tract model, was combined for joint optimization. Vincent $et\ al.$ [23] and Ghosh $et\ al.$ [24] optimized the ARX-LF model parameter values by searching the entire possible space. Schleusing $et\ al.$ [25] presented a differential evolution approach to optimize the ARX-LF model parameters. Li $et\ al.$ [26] and Takahashi $et\ al.$ [27] presented an iterative algorithm to optimize the ARX-LF model parameters, in which an electro-glottograph (EGG) signal was used to estimate initial values of the LF model for the iteration. Due to the inconvenience of EGG in real conditions, in our previous study [28], we proposed a simple framework for estimating glottal source and vocal tract parameters, in which an inverse filter was used to estimate the LF model parameter values, and these values were used as the initial values for the iterative algorithm based on the ARX-LF model. We tested this method on synthetic vowels on clear conditions (without glottal noise) with an almost fixed F_0, and the results are comparable with the state-of-the-art method (IAIF with DyProg-LF) [17]. However, in the real-world scenario, glottal noises and F_0 have a wide range of variation, which appears frequently in many voices, e.g., varying glottal noise levels on different voice types and varying F_0 on female, male, and emotional voices. Therefore, it is important to know the robustness of the glottal source and vocal tract estimation method for different glottal noise and F_0 levels.

This paper extends our previous work [28] to more complex conditions of different glottal noise and different F_0 levels in order to further assess the interaction between glottal noise and F_0. In this present study, we propose a two-step strategy. The first step initializes the glottal source parameters using the inverse filter method; the second step simultaneously estimates accurate glottal source and vocal tract shape parameters using an analysis-by-synthesis approach with an iterative algorithm. The proposed method effectively estimates the glottal source and vocal tract parameters based on the ARX-LF model to show robustness for different amounts of glottal noise and F_0 levels.

The remainder of this paper is structured as follows. Section 2 describes the ARX-LF model of speech production. Section 3 presents the implementation of the estimation algorithm. Section 4 describes the detailed synthetic vowels conditions and the performance evaluations. The conclusions are given in section 5.

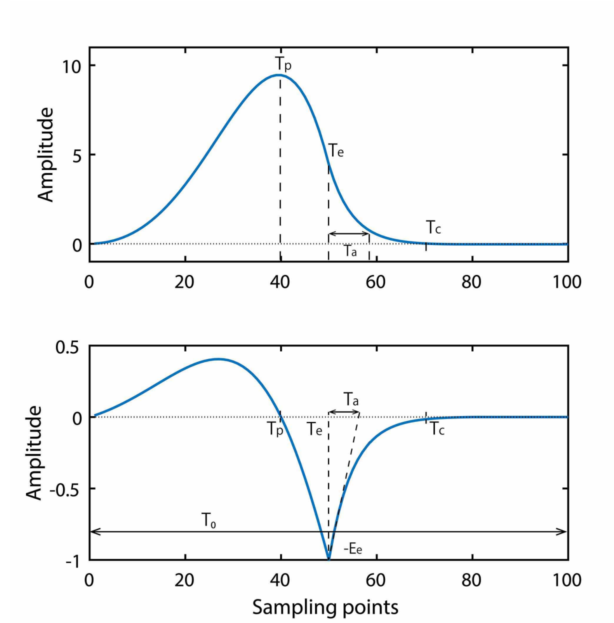

Fig. 1. One period of glottal source waveform (top) and its derivative waveform modeled by the LF model (bottom).

II. Source-filter model of speech production

Among source-filter models, the ARX-LF model is frequently used, in which the LF model is for glottal source and the ARX model is for vocal tract shape. The reason for choosing the ARX model is because the auto-regressive (AR) process models human speech production [29]. The reason for choosing the LF model is listed in the following, it is a suitable model for describing the glottal source waveform derivative [30], and is flexible enough for speech synthesis. Furthermore, the LF model has the smallest prediction error compared with other glottal source models [31]. Therefore, the ARX-LF model was chosen for this study and is introduced in this section.

A. ARX-LF model

The LF model mainly consists of six parameters to represent the glottal source waveform derivative, including five time-domain parameters T_p, T_e, T_a, T_c, T_0 and one amplitude parameter E_e. One period of glottal source waveform and its derivative waveform of the LF model is plotted in Fig. 1. T_0 is one period of glottal source waveform, T_p is the instant of the maximum glottal source waveform, T_e is the instant of the maximum negative differentiated glottal source waveform, 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. Since T_c is often set to T_0 in a simple

LF model version, five parameters are used in this paper. The LF model in the time domain can be formulated as:

$$u(n) = \begin{cases} E_1 e^{an} \sin(wn) & 0 \le n \le T_e \\ -E_2[e^{-b(n-T_e)} - e^{-b(T_0-T_e)}] & T_e \le n \le T_c \\ 0 & T_c \le n \le T_0 \end{cases} \quad (1)$$

where E_1, E_2, a, b and w are the parameters related to T_p, T_e, T_a ,E_e and T_0 [11].

Given the LF glottal source waveform derivative, the speech signal $s(n)$ can be synthesized by means of an ARX model:

$$s(n) = -\sum_{i=1}^{p} a_i(n)s(n-i) + b_0 u(n) + e(n). \quad (2)$$

where a_i are the coefficients of the p-order ARX model characterizing the vocal tract filter, b_0 is related to the amplitude of the LF glottal source waveform derivative and $e(n)$ is the glottal noise signal (residual signal).

III. Implementation of the Simultaneous Estimation Algorithm

In this section, the detailed implementation of the estimation algorithm based on the ARX-LF model is described. The structure of implementation is shown in Fig. 2. There are two components in the proposed structure, *initialization* and *iterative algorithm*.

A. Initialization

The purpose of this sub-section is to provide initial parameter values for the ARX-LF model, including glottal closure instant (GCI), T_p^0, T_e^0, T_a^0, and E_e^0.

1) GCI determination: The purpose of this step is to find the vocal fold vibration period for the LF model, especially to find the start and end positions in each period, which correspond to the start and end point of one period of the LF model. It is well known that GCI is the easiest to detect during a vocal fold vibration period. Thus, GCIs, which correspond to the minimum amplitude position of the LF model waveform, are estimated first for the ARX-LF model.

There are various methods to estimate GCI from voice speech signals, such as hilbert envelope-based detection (HE) [32], dynamic programming phase slope algorithm (DYPSA) [33], yet another GCI algorithm (YAGA) [34], zero frequency resonator-based method (ZFR) [35], and speech event detection using the residual excitation and a mean based signal (SEDREAMS) [36], etc. Among these methods, the SEDREAMS technique shows the best performance [37]. Thus, the SEDREAMS method was chosen for GCIs estimation. T_0 is the distance between two continuous GCIs (T_0= GCI_{i+1}-GCI_i, i is number of periods).

2) Initial values of T_p^0, T_e^0, T_a^0, and E_e^0: The purpose of this step is to estimate initial parameter values (T_p^0, T_e^0, T_a^0, and E_e^0) for the next iterative algorithm of the ARX-LF model. In this step, the state-of-the-art of glottal inverse filtering (IAIF) is firstly used to process the voiced speech signals, then a Dynamic programming (DyProg-LF) is used to estimate the (T_p^0, T_e^0, T_a^0, and E_e^0) values. The detailed implementation of the IAIF and DyProg-LF algorithm was described in [16].

B. Implementation of the iterative algorithm

The optimal parameter values of the LF model and the ARX model are iteratively found in the sense of minimizing the mean square error (MMSE) for each three periods of the glottal source waveforms. There are two procedures in this step. The first procedure is under a fixed GCI condition. The glottal source waveform derivative $u(n)$ is synthesized by initial values of T_p^0, T_e^0, T_a^0, and E_e^0, $u(n)$ and then input to the ARX model. The ARX model parameters (vocal tract filter coefficients: a_i) can be estimated by using Eq. (2) with the least square method. Eq. (2) can be transformed to:

$$e(n) = s(n) - \sum_{i=1}^{p} a_i(n)s(n-i) - b_0 u(n). \quad (3)$$

the p-order ARX model coefficients a_i and b_0 can be calculated by $\mathbf{h}$ in Eqs. (4), (5), and (6). $s(n)$ is the speech waveform at time n, and $u(n)$ is the glottal source waveform derivative at time n. N is the number of sampling points in one glottal vibration period (T_0).

Eq. (3) can be transformed into a matrix form, as

$$\begin{aligned} \mathbf{e} &= \mathbf{x_0} + \mathbf{Xa} - \mathbf{u}_0 b_0 \\ &= \mathbf{x_0} + \begin{bmatrix} \mathbf{X} & | & -\mathbf{u}_0 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ - \\ b_0 \end{bmatrix} \\ &= \mathbf{x_0} + \mathbf{Fh}. \end{aligned} \quad (4)$$

where

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

$$\mathbf{F} = \begin{bmatrix} \mathbf{X} & | & -\mathbf{u}_0 \end{bmatrix},$$

$$\mathbf{h} = \begin{bmatrix} \mathbf{a} \\ - \\ b_0 \end{bmatrix},$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{bmatrix}, \quad (5)$$

$$\mathbf{u}_0 = \begin{bmatrix} u(n) \\ u(n-1) \\ \vdots \\ u(n-N+1) \end{bmatrix},$$

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}.$$

$$\mathbf{h} = -\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{x_0} \quad (6)$$

As shown in Fig. 2, $x(n)$ can be synthesized using $u(n)$ and the estimated coefficients, glottal noise $\widehat{e}(n)$ is calculated by the output of a inverse ARX model with $s(n)$- $x(n)$. In each iteration of this procedure, the LF model parameters are

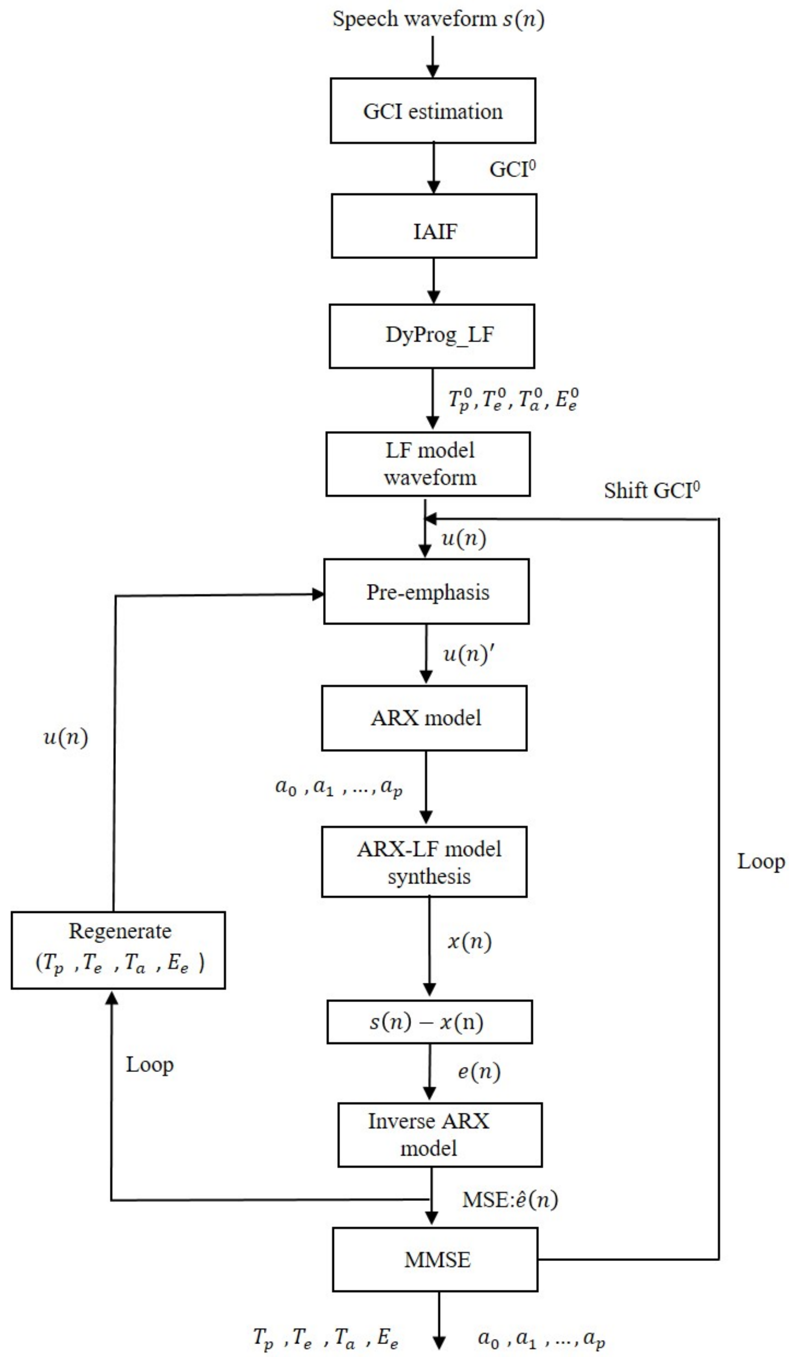

Fig. 2. Structure of the simultaneous estimation of glottal source and vocal tract parameters based on the ARX-LF model

ranged around the initial values of T_p^0, T_e^0, T_a^0, and E_e^0, and the glottal source waveform derivative is regenerated using these parameter values. Note that, in order to enhance the spectral flatness in the high frequency, the voiced speech signal $s(n)$ and glottal source waveform derivative $u(n)$ are pre-emphasized. This seems to improve the estimation accuracy in the high frequency region.

In the second procedure, although the high GCI estimation method (SEDREAMS) was used in the initial step, the estimation accuracy of the ARX-LF model is sensitive to the accuracy of GCI [38]. Therefore, the GCIs are further shifted around the initial GCI^0, to obtain more accurate GCI location. GCI^0 is further searched in the four sampling points from GCI^0 left and right. Then, the first procedure is run again for each shifted GCI. For a shifted GCI, the iteration processing in the MMSE optimization is set to 2000. After all the iterations, glottal source parameters (T_p, T_e, T_a, and E_e) and vocal tract filter coefficient values with the least MMSE are regarded as the optimal parameter values.

In this paper, the sampling frequency is set to 12000Hz, the vocal tract filter order p is set to 14, the frame length is set to 3 periods of the glottal source waveforms, and the frame shift

TABLE I
VARYING GLOTTAL SOURCE PARAMETERS FOR SYNTHESIZING VOICED SPEECH

Glottal source					
F_0	Noise (SNR)	T_p	T_e	T_a	E_e
80 : 20 : 320	0:10:50	0.75*T_e	0.35 : 0.1 : 0.85	0.08	1

TABLE II
VARYING VOCAL TRACT FILTERS PARAMETERS FOR SYNTHESIZING VOICED SPEECH

Vocal tract filters										
	/a/		/e/		/i/		/o/		/o/	
Formants	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2
Frequency (Hz)	960	1184	516	1959	301	1916	516	796	366	1206

is set to 1 period of the glottal source waveform.

IV. EXPERIMENTS AND RESULTS

Most studies test their methods on synthetic voiced speech[17], [20], [22], [24], [28], since the glottal source and vocal tract parameter values of synthesized speech are known as reference values, and the accuracy of the estimated parameter values can be calculated by comparing with the referenced parameter values. In our previous study [28], the performance of the proposed method was tested on the synthesized vowels that assume no glottal noises in the glottal source waveform and almost fixed F_0 conditions. In this paper, much closer to real conditions with varying glottal noise and F_0 levels on synthesized vowels are used for the performance evaluation of the proposed method.

A. Synthesized vowels

The source-filter model was used to synthesize the vowels that were the output signals of vocal tract filters with an input glottal exciton. The LF model was used to synthesize the glottal exciton that was input to the vocal tract filters/shape of five vowels (/a/, /e/, /i/, /o/, and /u/). Two steps were used for synthesizing vowels: the first step was to synthesize the glottal source waveform using the different parameter values of Table I; the second step was to synthesize the vocal tract filter shape for the vowels by using the formant frequencies listed in Table I. The detailed procedures of vowel synthesis have been described in [39].

To discuss the performance of the proposed method for varying glottal noises and F_0 levels, F_0 was varied from low levels (80 Hz) to high levels (320 Hz), and glottal noises were modeled by adding the white noise to the glottal source waveform derivative, which varied from strong noise conditions with signal-to-noise ratio (SNR = 0 dB) to almost clean conditions (SNR = 50 dB). Glottal source and vocal tract parameter values for synthesizing vowels are summarized in Table I. 2340 different conditions (6 T_e × 13 F_0 × 6 SNR × 5 filters =2340) were investigated for synthesized vowels, and each condition has 10 glottal source periods, thus, a total of 23400 periods of synthesized vowels for testing the proposed method.

B. Results and evaluation

To evaluate the performance of the proposed method and IAIF-DyProg-LF method, based on the structure in section III; the accuracy of the proposed method is compared with the IAIF-DyProg-LF method. The estimated LF model parameter values, and first formant frequency (F_1) and second formant frequency (F_2) were compared with the reference values. Let the reference values be vector $\beta \in \{T_p, T_e, T_a, E_e, F_1, F_2\}$ and the estimated values be vector $\hat{\beta}$. The estimation error of one parameter ($\gamma_m, m = 1, 2..., 6$) between reference and estimated values can be calculated by Eq. (7):

$$\gamma_m = \frac{|\hat{\beta_m} - \beta_m|}{\beta_m} \times 100\%. \tag{7}$$

1) Robustness to glottal noise: To examine the robustness on varying SNR levels of the proposed method and IAIF-DyProg-LF methods, as mentioned in section IV-A, white noise with various SNR levels has been added to the glottal source waveform derivative for synthesizing voiced speech. In each different SNR level, a total of 3900 glottal periods voiced speech (6 T_e × 13 F_0 × 5 filters × 10 periods) are analyzed by the proposed method and IAIF-DyProg-LF methods.

The performance of the proposed method and IAIF-DyProg-LF method are compared according to Eq. (7). The averaged estimation errors of $\{T_p, T_e, T_a, E_e, F_1, F_2\}$ for the proposed method and IAIF-DyProg-LF methods in varying SNR levels (from 0 dB to 50 dB) are plotted in Fig. 3. As shown in Fig. 3, the estimation errors of the proposed method were smaller than those of IAIF-DyProg-LF under the differenent SNR levels. The estimation errors were different for each SNR level, the estimation errors of all parameters (except Ee of IAIF-DyProg-LF) were the largest in voiced speech with 0 dB (SNR), and the estimation errors of all parameters (except E_e of the proposed method) were the smallest in voiced speech with 50 dB (SNR). More specifically, the averaged estimation errors of the proposed method are in following ranges: T_p: between 20.4 % and 11.1 %; T_e: between 19.2 % and 9.7 %; T_a: between 15.3 % and 9.7 %; E_e: between 67.1% and 19.7 %; F_1: between 3.1 % and 2.7 %; F_2: between 33.3 % and 3.9 %. The averaged estimation errors of the IAIF-DyProg-LF methods are in the following ranges: T_p: between 23.1 % and

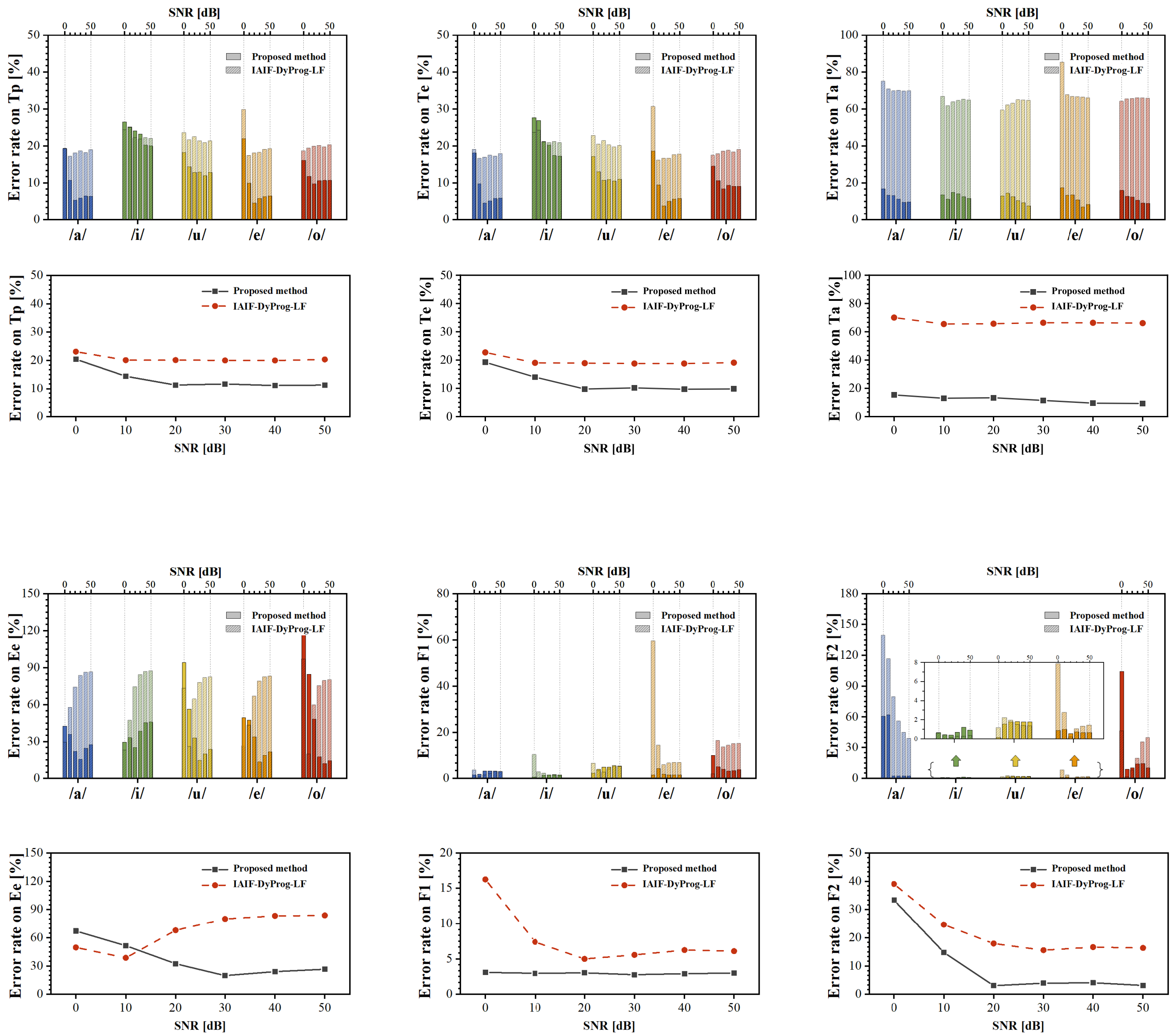

Fig. 3. Estimation errors of the six parameter values ($T_p, T_e, T_a, E_e, F_1, F_2$) of five vowels (different colors, first and third rows) for the proposed method and IAIF-DyProg-LF methods under varying SNR levels, and the averaged estimation errors (second and fourth rows): IAIF-DyProg-LF (red lines) and the proposed method

(black lines).

20.0 %; T_e: between 22.7 % and 18.7 %; T_a: between 70.0 % and 65.3 %; E_e: between 83.8 % and 38.6 %; F_1: between 16.3 % and 5.0 %; F_2: between 39.0 % and 16.4 %.

2) Robustness to F_0: To examine the robustness on varying F_0 levels of the proposed method, as mentioned in section IV-A, F_0 was varied form 80 Hz to 320 Hz with steps of 20 Hz for synthesizing voiced speech. In each different F_0 level, a total of 1800 glottal periods of voiced speech (6 T_e $\times$ 6 SNR $\times$ 5 filters $\times$ 10 periods) are analyzed by the proposed method and IAIF-DyProg-LF methods.

The performance of the proposed method and IAIF-DyProg-LF methods are compared according to Eq. (7), the averaged estimation errors of $\{T_p, T_e, T_a, E_e, F_1, F_2\}$ for the proposed method and IAIF-DyProg-LF methods in different F_0 levels (from 80 Hz to 320 Hz) are plotted in Fig. 4. As shown in Fig. 4, for the two methods, the estimation errors were different for each F_0 level: the estimation errors of T_p, T_e, T_a, F_1, and F_2 were the largest in voiced speech with 320 Hz (F_0), and were smallest in voiced speech with 80 Hz (F_0). The estimation errors parameter E_e of the proposed method kept nearly the same values for each different F_0 level, whereas estimation errors of the parameter IAIF-DyProg-LF were the largest in voice speech with 80 Hz (F_0), and were smallest in voice speech with 320 Hz (F_0). More specifically, the averaged

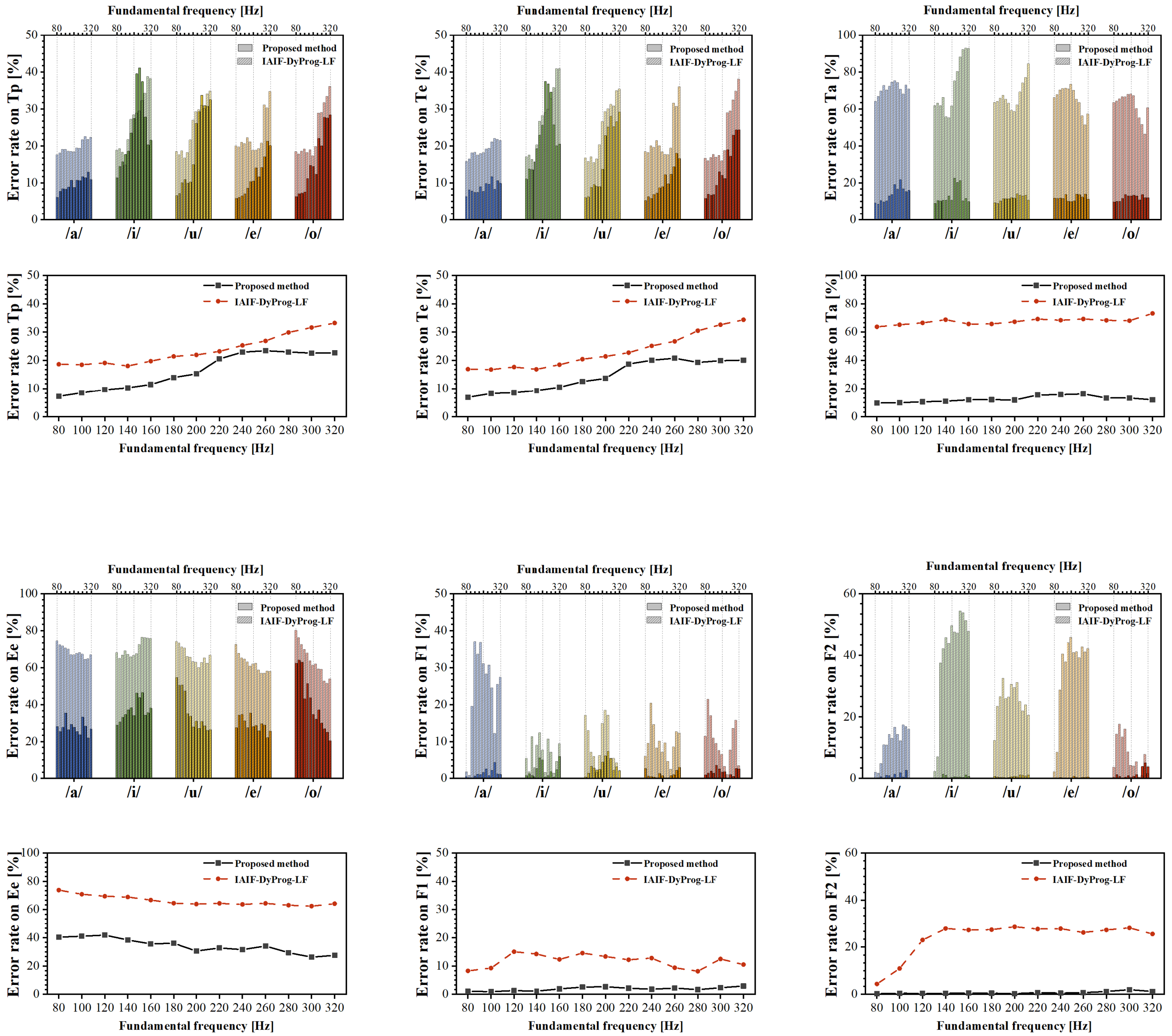

Fig. 4. Estimation errors of the six parameter values ($T_p, T_e, T_a, E_e, F_1, F_2$) of five vowels (different colors, first and third rows) for the proposed method and IAIF-DyProg-LF methods under varying F_0 levels, and the averaged estimation errors (second and fourth rows): IAIF-DyProg-LF (red lines) and the proposed method (black lines).

estimation errors of the proposed method have the following ranges: T_p: between 22.9 % and 7.2 %; T_e: between 20.0 % and 6.8 %; T_a: between 13.2 % and 9.7 %; E_e: between 41.7% and 26.7 %; F_1: between 2.8 % and 0.9 %; F_2: between 1.9 % and 0.2 %. For averaged estimation errors of the IAIF-DyProg-LF methods, the ranges are as follows: T_p: between 33.2 % and 18.0 %; T_e: between 34.3 % and 16.7 %; T_a: between 69.1.0 % and 63.7 %; E_e: between 73.7 % and 62.4 %; F_1: between 14.5 % and 8.1 %; F_2: between 28.2 % and 4.3 %.

C. Discussion

Fig. 3 clearly shows that the performance of the two methods was strongly affected by the different glottal noise levels. The results show that estimation errors of the proposed method were much smaller than those of IAIF-DyProg-LF on different glottal noise levels, which indicates the estimation accuracy of the proposed method was higher than that of IAIF-DyProg-LF on different glottal noise levels. It is further noted that the estimation errors of the two methods (except parameter E_e of the IAIF-DyProg-LF) decrease with the increase of SNR level. This result is similar to the findings in [40] which reported a high SNR level has high performance of IAIF-

DyProg-LF method. As expected, more aperiodic components in the glottal source waveform result in estimation performance decrease; therefore, it may be more difficult to analyze whisper and breathy speech (high level aperiodic components) than normal speech, either by the the proposed method or the IAIF-DyProg-LF method. More importantly, for conditions of glottal noise level greater than 20dB, estimation errors of the proposed method for glottal source parameters and formant frequencies (F_1 and F_2) were smaller than 15% and 5%, respectively. Moreover, the estimation errors of the proposed method for F_1 and F_2 are insensitive to the different glottal noise levels. These results indicate that the proposed method has a strong robustness with regard to the different glottal noise levels.

Fig. 4 clearly shows that the performance of the two methods was strongly affected by the different F_0 levels. The results show that estimation errors of the proposed method were much smaller than these of IAIF-DyProg-LF on different F_0 levels, which indicates the estimation accuracy of the proposed method was higher than that of IAIF-DyProg-LF on different F_0 levels. It is further noted that the estimation errors of the two methods increase with the increase of F_0 levels for T_p, T_e, and T_a, whereas the estimation errors of the two methods keep spectral flatness or decrease with the increase of F_0 for parameters E_e, F_1, and F_2, which is in line with the findings in [17]. More importantly, estimation errors of the proposed method were smaller than 20% for all parameters (except E_e). Noted also is that the estimation errors of the proposed method for parameters T_a, E_e, F_1, and F_2 are insensitive to the different F_0 levels. These results indicate that the proposed method has strong robustness with regard to the different F_0 levels.

Figs. 3 and 4 clearly show that the estimation error of the proposed method was much smaller than those of IAIF-DyProg-LF on different glottal noise and F_0 levels. It is noted that the estimation errors of the proposed method for F_1 and F_2 on different glottal noise and F_0 levels are different; the estimation error differences for F_1 and F_2 were also found in [25]. Moreover, the estimation errors of the proposed method for F_1 and F_2 were insensitive to both different glottal noise and F_0 levels.

All the above results indicate that the proposed method has strong robustness with regard to the different glottal noise and F_0 levels, and the estimation accuracy of the proposed method is higher than that of IAIF-Dyprog-LF for both conditions. It suggests that the proposed method can be used to analyze speech signals with high F_0 and low SNR, such as falsetto voice of females, more breathy voice quality, and high arousal emotional voice.

V. Conclusion

In this paper, an automatic speech analysis method to estimate the glottal source and vocal tract parameters was proposed based on the ARX-LF model; then the performance of the proposed method on different glottal noise and F_0 levels was discussed. The glottal source and vocal tract parameters of the synthesized vowels with different glottal noise and F_0 levels were estimated by the proposed method and IAIF-DyProg-LF methods. The results show that (1) the estimation accuracy of the proposed method is higher than that of IAIF-Dyprog-LF for both different glottal noise and different F_0 levels, and (2) the performance of the proposed method is insensitive for different glottal noise and different F_0 levels. It indicates that the proposed method is robust for estimating glottal source and vocal tract parameters.

Limitations of this study are that the performance of the proposed method was tested only for the synthesized vowels (/a/, /e/, /i/, /o/ and /u/); real voice speech should be taken into account. Also, the focus of this study was the estimation accuracy of the proposed method. Since an iterative algorithm was used, we analyzed vowel of glottal vibration in 10 periods, which takes an average of 20 seconds. Future work is necessary to reduce analysis time.

References

[1] T. Claes, I. Dologlou, L. T. Bosch, and D. Van Compernolle, "A novel feature transformation for vocal tract length normalization in automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 6, pp. 549–557, 1998.

[2] L. Juvela, B. Bollepalli, V. Tsiaras, and P. Alku, "Glotnet—a raw waveform model for the glottal excitation in statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1019–1030, 2019.

[3] O. Perrotin and I. Mcloughlin, "Glottal flow synthesis for whisper-to-speech conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 889–900, 2020.

[4] M. K. Reddy, P. Alku, and K. S. Rao, "Detection of specific language impairment in children using glottal source features," *IEEE Access*, vol. 8, pp. 15 273–15 279, 2020.

[5] N. P. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67 745–67 755, 2020.

[6] H. Vinod and R. Sharma, "Glottal wave analysis of dysphonic voice using inverse filtering," 2018.

[7] Z. Xiao, Y. Chen, and Z. Tao, "Contribution of glottal waveform in speech emotion: A comparative pairwise investigation," in *2018 IEEE International Conference on Progress in Informatics and Computing (PIC)*, 2018, pp. 185–190.

[8] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, 1999.

[9] G. Degottex, E. Bianco, and X. Rodet, "Usual to particular phonatory situations studied with high-speed videoendoscopy," in *The 6th International Conference on Voice Physiology and Biomechanics, ICVPB*, 2008.

[10] K. Honda, H. Takemoto, T. Kitamura, S. Fujita, and S. Takano, "Exploring human speech production mechanisms by mri," *IEICE Transactions on Information and Systems*, vol. 87, no. 5, pp. 1050–1058, 2004.

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

[13] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[14] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 313–327, 1998.

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

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

[17] J. Kane and C. Gobl, , "Evaluation of automatic glottal source analysis," LNAI 7911, pp. 1-8, 2013.

[18] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for glottal source waveform," in *IEEE International Conference on Acoustics, Speech, Signal Processing*, 1986.

[19] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.
[20] 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.
[21] K. Funaki, Y. Miyanaga, and K. Tochinai, "Recursive armax speech analysis based on a glottal source model with phase compensation," *Signal Processing*, vol. 74, no. 3, pp. 279–295, 1999.
[22] F. Qiang and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Transactions on Audio Speech Language Processing*, vol. 14, no. 2, pp. 492–501, 2006.
[23] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of lf glottal source parameters based on an arx model," in *European Conference on Interspeech -eurospeech*, pp. 333–336, 2005.
[24] P. K. Ghosh and S. S. Narayanan, "Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter," *Speech Communication*, vol. 53, no. 1, pp. 98–109, 2011.
[25] O. Schleusing, T. Kinnunen, B. H. Story, and J. Vesin, "Joint source-filter optimization for accurate vocal tract estimation using differential evolution," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1560–1572, 2013.
[26] Y. Li, K. Sakakibara, D. Morikawa, and M. Akagi, "Commonalities of glottal sources and vocal tract shapes among speakers in emotional speech," in *Studies on Speech Production, ISSP 2017. Lecture Note in Computer Science*, Springer, vol. 10733, pp. 24–34， 2018.
[27] K. Takahashi and M. Akagi, "Estimation of glottal source waveforms and vocal tract shape for singing voices with wide frequency range," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* pp. 1879–1887, 2018.
[28] Y. Li, K. 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*, vol. 92, pp. 831–838, 2020.
[29] K. Funaki, Y. Miyanaga, and K. Tochinai, "A time-varying armax speech analysis method based on the glottal source model," *The Journal of the Acoustical Society of America*, vol. 100, no. 4, p. 2602, 1996.
[30] D. G. Childers and C. Ahn, "Modeling the glottal volume‐ velocity waveform for three voice types," *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 505–519, 1995.
[31] H. Strik, "Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses," *Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2659–2669, 1998.
[32] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 309–319, 1979.
[33] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.
[34] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.
[35] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.
[36] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Tenth Annual Conference of the International Speech Communication Association*, pp. 2891–2894， 2009.
[37] T. Drugman, M. R. P. Thomas, J. Gudnason, P. A. 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, 2012.
[38] H. L. Lu, "Toward a hig hquality singing synthesizer with vocal texture control," *Stanford University*, 2002.
[39] H. Kawahara, K. Sakakibara, H. Banno, M. Morise, T. Toda, and T. Irino, "Aliasing-free implementation of discrete-time glottal source models and their applications to speech synthesis and f0 extractor evaluation," *Asia-pacific Signal and Information Processing Association Summit and Conference, (APSIPA ASC)*, pp. 520–529, 2015.
[40] 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.

Yongwei Li received his M.S. and Ph.D. in information science form the Japan Advanced Institute of Science and Technology (JAIST) in 2014 and 2018. He is currently an Assistant Professor in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CAS), Bejing, China. His current research interests include modeling of speech production, voice quality, and speech emotion recognition and synthesis.

Jianhua Tao (Senior Member, IEEE) received the M.S. degree from Nanjing University, Nanjing, China, in 1996, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001. He is currently a Professor with NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 200 papers on major journals and proceedings including the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. His current research interests include speech recognition, speech synthesis, human computer interaction, affective computing, and pattern recognition. He is the Board Member of ISCA, the Chair or Program Committee Member for several major conferences, including Interspeech, ICPR, ACII, ICMI, ISCSLP, etc. He was the Steering Committee Member for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, and is an Associate Editor for Journal on Multimodal User Interface and International Journal on Synthetic Emotions. He was the recipient of several awards from the important conferences, such as Interspeech, NCMMSC, etc.

Donna Erickson received her B.A. from The Ohio State University in 1966, her M.A. from the University of Michigan in 1968, and her Ph.D. from the University of Connecticut in 1976. Her Ph.D. thesis was on the laryngeal electromyographic activity underlying the tones of Thai. She taught linguistics and English as a Second Language at Earlham College and at The Ohio State University from 1982-1996. During this time, she was also a Research Scientist at the Center for Cognitive Linguistics at The Ohio State University, working with Professor Osamu Fujimura on voice quality and analyzing X-ray Microbeam data. In 1998, she was a Visiting Professor at Kanazawa University in Japan, from 2000-2006, a Professor at Gifu City Women's College, Japan, and from 2006-2012, a Professor at Show University of Music, Kawasaki, Japan. She retired from full-time teaching in 2012, and is currently an Affiliated Research Scientist at Haskins Laboratories, New Haven, Connecticut. She continues to be active in research, focusing on the acoustic and articulatory characteristics of voice production, as it pertains to emotional and social affective expressions, and also voice production changes during singing.

Bin Liu (Member, IEEE) received the B.S. degree and the M.S. degree from the Beijing institute of technology, Beijing, China, in 2007 and 2009 respectively. He received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include affective computing and audio signal processing.

Masato Akagi received his B.E. from Nagoya Institute of Technology in 1979, and his M.E. and Ph.D. Eng. from the Tokyo Institute of Technology in 1981 and 1984. He joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science of the Japan Advanced Institute of Science and Technology (JAIST) and is now a full professor. His research interests include speech perception, modeling of speech perception mechanisms in human beings, and signal processing of speech. Dr. Akagi received the IEICE Excellent Paper Award from the IEICE in 1987, the Best Paper Award from the Research Institute of Signal Processing in 2009, and the Sato Prize for Outstanding Papers from the ASJ in 1998, 2005, 2010 and 2011.