Noname manuscript No.
(will be inserted by the editor)

Simultaneous estimation of glottal source waveforms and vocal tract shapes from speech signals based on ARX-LF model

Yongwei Li [1,2] · **Ken-Ichi Sakakibara** [3] · **Masato Akagi** [1]

Received: date / Accepted: date

Abstract Estimating glottal source waveforms and vocal tract shapes is typically done by processing the speech signal using an inverse filter and then fitting the residual signal using the glottal source model. However, due to source-tract interactions, the estimation accuracy is reduced. In this paper, we propose a method to estimate glottal source waveforms and vocal tract shapes simultaneously based on an analysis-by-synthesis approach with a source-filter model constructed of an Auto-Regressive eXogenous (ARX) model amd the Liljencrants-Fant (LF) model. Since the optimization of multiple parameters makes simultaneous estimation difficult, we first initialize the glottal source parameters using the inverse filter method, and then simultaneously estimate the accurate parameters of the glottal sources and the vocal tract shapes using an analysis-by-synthesis approach. Experimental results with synthetic and real speech signals showed that the proposed method had higher estimation accuracy than using the inverse filter.

Keywords glottal source waveform · vocal tract shape · ARX-LF model

Yongwei Li
yongwei@jaist.ac.jp

Ken-Ichi Sakakibara
kis@hoku-iryo-u.ac.jp

Masato Akagi
akagi@jaist.ac.jp

[1] Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan

[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Department of Communication Disorders, Health Science University of Hokkaido, 1757 Kanazawa, Tobetsu-cho, Ishikari-gun, Hokkaido, 061-0293, Japan

1 Introduction

The estimation of glottal source waveforms and vocal tract shapes is important for speech signal processing. It is utilized in many fields, such as speech recognition [1], speech synthesis [2], voice pathology detection [3], voice conversion [4], speaker recognition [5], and emotional speech recognition [6, 7], and in further understanding speech production mechanisms. Based on the source-filter theory of speech production [8], speech signals are modeled as output signals of a vocal tract filter with a glottal source excitation.

There are now many methods for estimating glottal source waveforms and vocal tract shapes based on a source-filter model. A widely used method to estimate vocal tract filters is linear prediction (LP) analysis, but the problem with this method is that it is difficult to estimate vocal tract filters without glottal source effects from the speech signals (source-tract interaction) [9]. To overcome this problem, Wong *et al*. estimated glottal source waveforms and vocal tract filters by LP analysis during the glottal closed phase, where there is no interaction [10]. However, this approach provides reliable estimations only in the long duration of glottal closure. It is difficult to find the glottal closed phase in real conditions, especially in the case of a very short glottal closed phase.

A simple and straightforward way to process speech signals to estimate glottal source waveforms is inverse

filtering, where glottal sources can be considered residual signals or periodic pulses for voiced sound/white noise for unvoiced sound [11, 12]. This oversimplified source assumption often faces the problem of inadequate description of glottal source characteristics. An improved method was proposed to deal with the residual signals by fitting parameters of glottal source models [13], e.g., the Liljencrants-Fant (LF) model [8], the Rosenberg-Klatt (RK) model [14], and the Fujisaki-Ljungqvist (FL) model [15]. Although a more accurate glottal source model is used, there is still the problem of source-filter interaction (due to first estimating the filter and then estimating the source), as mentioned in the above paragraph.

To reduce this source-filter interaction, glottal source parameters and vocal tract parameters can be estimated by a joint optimization process based on an analysis-by-synthesis scheme [16–20]. The main idea here is that a glottal source model is utilized as an input glottal excitation to a vocal tract filter that is assumed to be non-time-invariant. Thus, the source and filter are independent of each other. Among such source-filter models, the Auto-Regressive eXogenous with the LF (ARX-LF) model is widely used, in which the glottal source signal is represented by the LF model glottal waveform derivative and the vocal tract transfer function is represented by the ARX filter [19]. However, it is difficult to optimize/handle multiple parameters in the analysis and synthesis steps [20, 21].

To overcome this difficulty, three shape parameters of the LF model were transformed into a single shape parameter (R_d) by Fant [21]. Fu *et al.* provided initial values for the LF model, in which the RK model was first utilized to estimate initial values for the LF model, followed by the estimation of more accurate glottal source and vocal tract parameters [20]. Li *et al.* and Takahashi *et al.* proposed an iterative algorithm [22,23] to estimate accurate glottal source waveforms and vocal tract shapes, in which an electro-glottograph (EGG) signal was used to estimate initial values for the iteration. However, it is not always convenient to use EGG.

In this paper, we revisit the difficulty of optimizing multiple parameters. Following the idea of providing the initial values for the LF model, instead of the EGG signal, we first obtain the initial values of the LF model parameters using an inverse filter [13]. Then, the accurate glottal source waveforms and vocal tract shapes are estimated simultaneously based on the ARX-LF model using an iterative algorithm [22] .

2 ARX-LF model

On the basis of the source-filter theory of speech production, a source-filter model to represent speech production procedures was proposed by G. Fant [8]. Among the source-filter models, the ARX-LF model is widely used for representing glottal source waveforms and the vocal tract shapes of speech, as it not only has overall adaptability to common speech waveforms but is also flexible enough to represent extreme phonations [20]. Thus, the ARX-LF model is used in this paper.

The glottal source signal in the ARX-LF model is represented by the LF glottal flow derivative and the vocal tract transfer function is represented by the ARX filter. More specifically, the glottal flow derivative is formulated in the LF model by six parameters, five of which are related to time (T_p, T_e, T_a, T_c and T_0) and one to amplitude (E_e), as shown in Fig. 1. T_0 is one period of glottal flow, T_p is the instant of the maximum glottal flow model 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. T_c is often set to T_0 in a simple LF model version. Thus, five parameters are used in this paper.

A typical LF glottal flow derivative is plotted in Fig. 1. The explicit expression of the LF glottal flow derivative for one fundamental period is given by:

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

The ARX model simulates a vocal tract filter. Given the LF glottal flow derivative, the speech signal $s(n)$ can be synthesized by means of an ARX model, as

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

where $a_i(n)$ are the coefficients of the p-order ARX model characterizing the vocal tract, b_0 is used to adjust the amplitude of the differentiated glottal flow, and $e(n)$ is the residual signal.

E_1, E_2, a, b, and w are the parameters related to T_p, T_e, T_a ,E_e, and T_0 [8].

3 Estimation of glottal source waveform and vocal tract shape

Our proposed procedure for estimating glottal source waveforms and vocal tract shapes is shown in Fig. 2. It consists of two steps: initialization and implementation

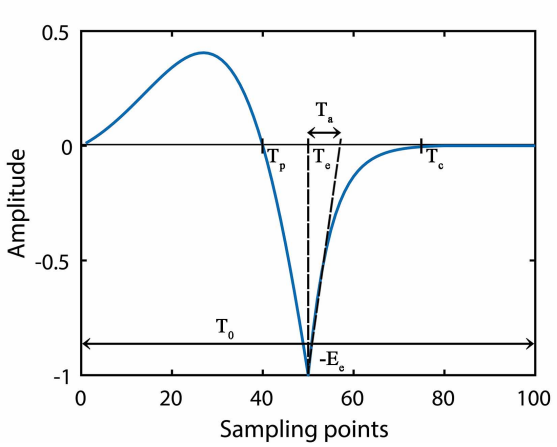

Fig. 1 One typical period of LF glottal flow derivative.

of simultaneous estimation. In the first step, instead of accuracy, initial values are prepared for the next step. The main step is the second step, in which precise glottal source waveforms and vocal tract shapes are estimated simultaneously by the proposed scheme based on the ARX-LF model.

3.1 Initialization

The objective of this step is to determine the period for the LF model. In one period of the LF model waveform, the glottal closure instant (GCI) is a discontinuity location, as shown in Fig. 1, and it is easier to detect than other locations in one period of the glottal source waveform. Thus, GCI is detected first, and the distance between two continuous GCIs is regarded as one period of T_0. Speech Event Detection using the Residual Excitation and a Mean-based Signal (SEDREAMS) is used for detecting the GCI [24], as it provides more accurate results. The detected GCI by the SEDREAMS method is denoted as GCI_0.

The iterative and adaptive inverse filters (IAIF) and the LF model are used to obtain the initial values of the LF model for the simultaneous estimation step. Dynamic programming (DyProg-LF) is utilized to fit the LF model parameters, and estimated glottal source parameters (LF model) are denoted as T_p^0, T_e^0, T_a^0 and E_e^0. The detailed implementation of the DyProg-LF algorithm has been described in [13] and [25].

3.2 Implementation of simultaneous estimation

In this step, a simultaneous estimation algorithm is implemented to accurately estimate the glottal source waveforms and vocal tract shapes on the basis of the ARX-LF model. There are two processes in this step.

First, the LF model parameters and vocal tract filter coefficients are obtained with a fixed GCI. The initial values of the LF model parameters (T_p^0, T_e^0, T_a^0, E_e^0) are used for synthesizing glottal source waveform derivative $u(n)$, and $u(u)$ is then exploited to synthesize $x(n)$ using the ARX model. The ARX model parameters can be estimated with mean square error (MSE) sense for $e(n)$ by the least square (LS) method. In each iteration of this optimization process, the LF model parameters are regenerated around the initial values of T_p^0, T_e^0, T_a^0 and E_e^0, and a glottal source waveform derivative is regenerated using these parameters. To estimate the p-oder ARX model coefficients $\mathbf{a}$, we transform Eq.(2) into a matrix form, as

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

$$\mathbf{e} = \mathbf{x_0} + \mathbf{Xa} - \mathbf{u}_0 b_0 = \mathbf{x_0} + \left[\mathbf{X} \mid -\mathbf{u}_0\right] \begin{bmatrix} \mathbf{a} \\ - \\ b_0 \end{bmatrix} = \mathbf{x_0} + \mathbf{Fh}. \tag{3}$$

where

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

$$\mathbf{X} = \left[\mathbf{x}_1\ \mathbf{x}_2 \cdots \mathbf{x}_p\right], \mathbf{u}_0 = \begin{bmatrix} u(n) \\ u(n-1) \\ \vdots \\ u(n-N+1) \end{bmatrix}, \tag{4}$$

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}, \ \mathbf{F} = \left[\mathbf{X} \mid -\mathbf{u}_0\right], \mathbf{h} = \begin{bmatrix} \mathbf{a} \\ - \\ b_0 \end{bmatrix}.$$

$s(n)$ is the speech waveform at time n, and $u(n)$ is the glottal source waveform at time n. For one period of glottal vibration, N is number of sampling points in T_0.

Then, the p-order ARX model coefficients $\mathbf{a}$ can be calculated as

$$\mathbf{h} = -\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{x_0}. \tag{5}$$

In the second process, we can estimate more accurate LF model parameters and vocal tract coefficients. Since the performance of the ARX-LF model is affected by the accuracy of GCI, as reported by Lu [26], we suggest further shifting the parameters of GCI around the value of GCI_0. GCI candidates are assumed to be in the four sampling points from GCI_0 left and right ($j <= 8$).

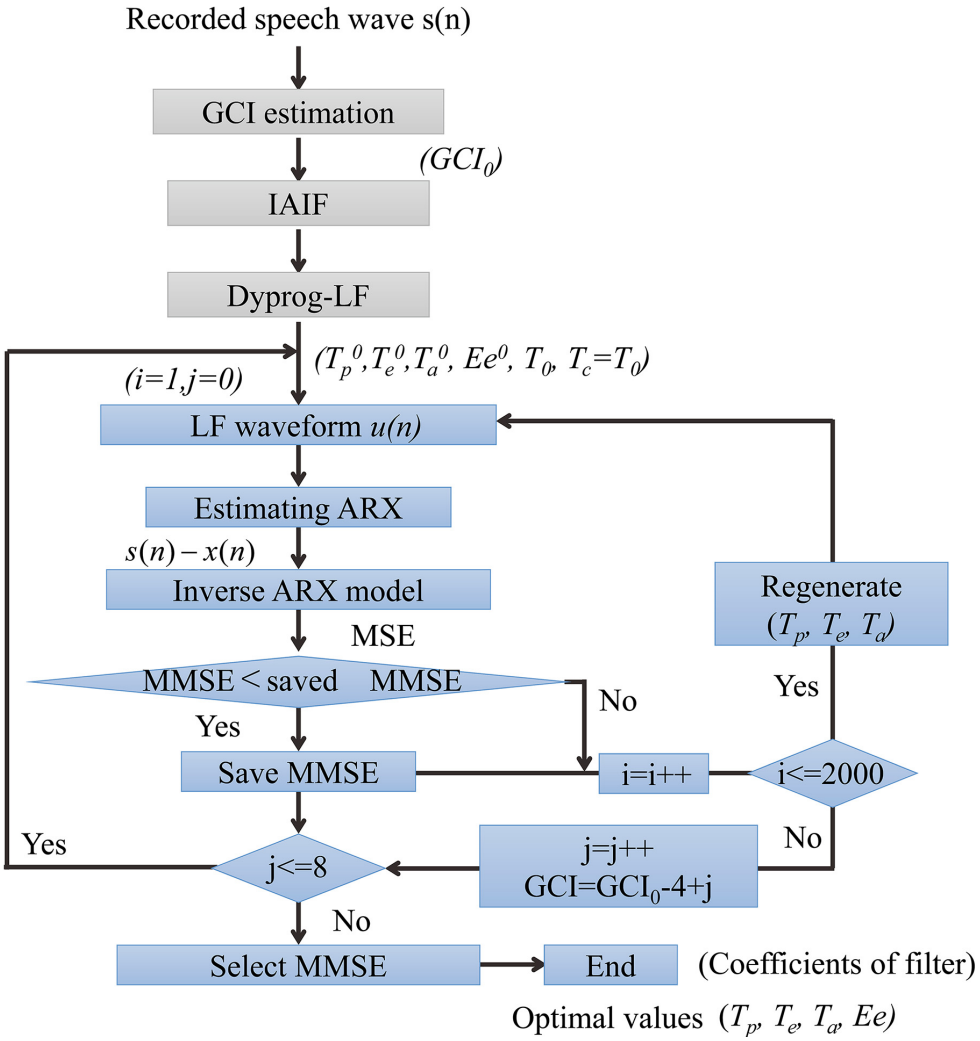

Fig. 2 Estimation scheme of glottal source waveform and vocal tract shape.

Table 1 Average estimation errors (ε) for synthesized vowels using two methods.

	Glottal source				Vocal tract	
	$T_p(\%)$	$T_e(\%)$	$T_a(\%)$	$E_e(\%)$	$F_1(\%)$	$F_2(\%)$
IAIF-Dyprog-LF	24.0	19.8	50.4	82.2	9.4	6.1
Proposed	11.4	9.5	60.0	23.2	2.3	1.1

Then, the first process is run again for each shifted GCI. For the given GCI each time, the iteration processing in the minimization of mean square error (MMSE) optimization is set to 2000 ($i <= 2000$). Finally, accurate glottal source parameters and vocal tract filter coefficients are estimated by MMSE.

In this work, the sampling frequency is 12000 Hz and p is set to 14. The estimation length of the frame is three periods of glottal source waveforms, and the shift frame is one period of a glottal source waveform.

4 Experimental results

First, we used synthetic vowels to test the proposed estimation method and the IAIF with Dyprog-LF method ([13]). The advantage of testing on synthetic vowels is that the accuracy of the proposed method can be investigated by comparing the estimated parameter values of glottal source waveforms and vocal tract shapes with referenced parameter values. After that, we estimated the glottal source waveforms and vocal tract shapes of real vowels to test the proposed estimation method and the IAIF with Dyprog-LF.

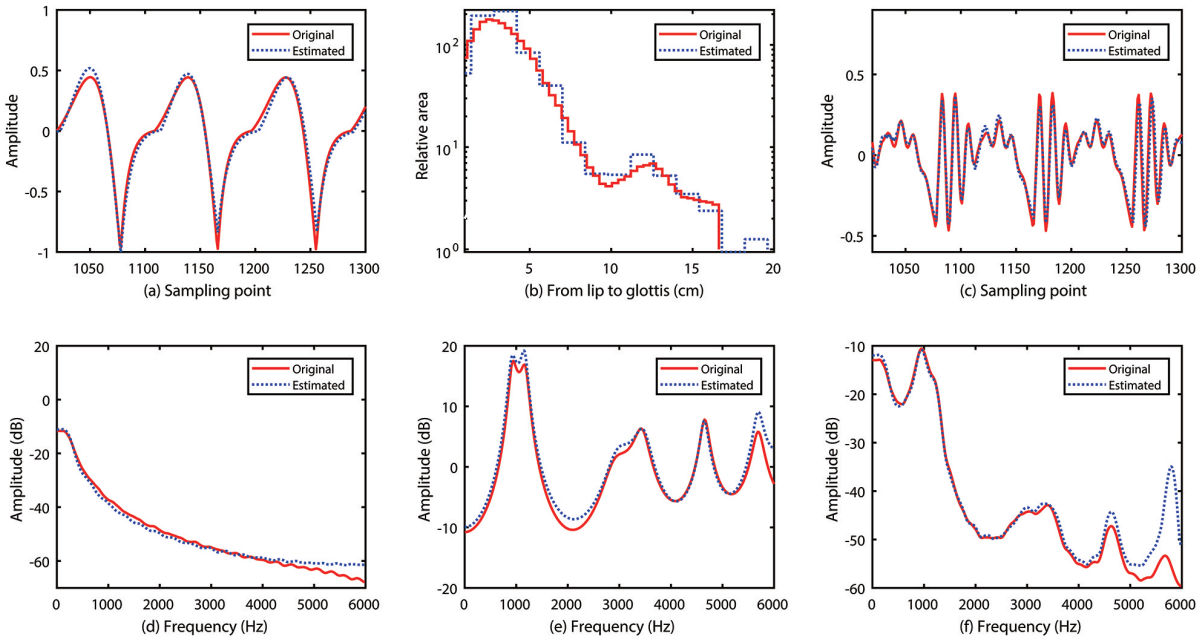

Fig. 3 Original and estimated glottal source waveforms in (a) time domain and (d) frequency domain. Original and estimated vocal tract shapes in (b) time domain and (e) frequency domain. Original and estimated voice waveforms in (c) time domain and (f) frequency domain.

4.1 Synthesized vowels

The synthesized vowels were produced according to the source-filter model, in which a glottal source waveform derivative is generated by the LF model. The vocal tract filters were taken from five Japanese vowels (/a/, /e/, /i/, /o/ and /u/) using Kawahara's method [27]. This is because the formant frequencies of the vocal tract in these vowels vary widely, especially the first and second formant frequencies (F_1 and F_2). We used a larger number of synthetic vowels with a wide range of the LF model parameter values in this paper. The LF model parameters were varied: T_e: 0.3 to 0.9 with steps of 0.05; T_p/T_e: 0.65 to 0.8 with steps of 0.05; and T_a: 0.03, 0.08, within the range suggested in [28]. In order to synthesize more realistic vowels, the fundamental frequency (F_0) was obtained from a real vowel, and 18 different F_0 ranging from 90 to 140 Hz were used for synthesis. Synthesized vowels with 9360 ($4[T_p] \times 13[T_e] \times 2[T_a] \times 18[F_0] \times 5[filter]$) different conditions were used for testing the proposed method.

4.1.1 Results and discussion

Estimated values of the LF model parameters and F_1 and F_2 of the vocal tract were evaluated with respect to the reference values. Let the reference values be vector $\theta \in \{T_p, T_e, T_a, E_e, F_1, F_2\}$ and the estimated values be vector $\hat{\theta}$. The error (ε) between estimated and reference values can be calculated as

$$\varepsilon = \frac{|\hat{\theta} - \theta|}{\theta} \times 100\%. \quad (6)$$

The average errors (ε) are listed in Table 1. As shown, estimation errors of a glottal source were less than 12 except for those of T_a (since T_a was the smallest of all parameters as the denominator in Eq. 6), and the error was 60%. Estimation errors were less than 2.3% for the vocal tract. Figure 3 shows an example of the estimated results, in which the glottal source waveform and vocal tract shape were estimated from a synthe-

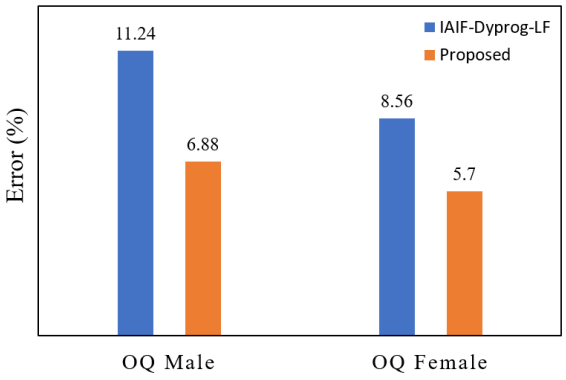

Fig. 4 Average estimation errors (ε_{OQ}) of proposed method and IAIF-Dyprog-LF.

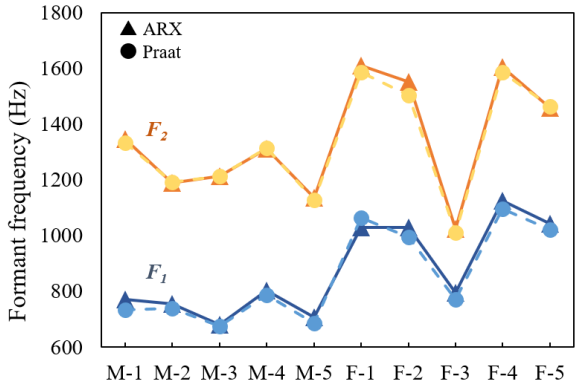

Fig. 5 Estimated first formant frequency and second formant frequency by proposed method and Praat for five males (M-1, M-2, M-3, M-4, and M-5) and five females (F-1, F-2, F-3, F-4, and F-5).

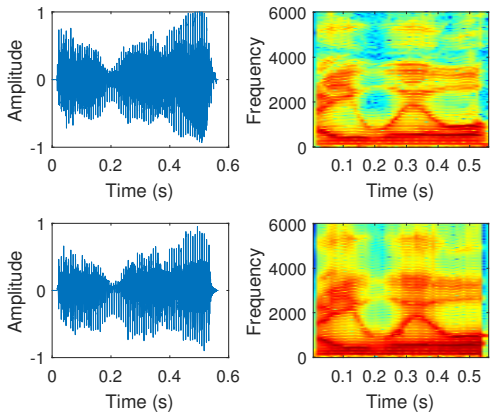

Fig. 6 Original speech waveform and its spectrogram (top), re-synthesized speech waveform and its spectrogram (bottom).

sized vowel /a/. As shown in the Fig. 3, the estimated glottal source waveforms and vocal tract shapes were very similar to the original ones in the time domain and frequency domain. The length of the vocal tract shapes was different between the estimated and original one because the sampling frequency and the order of the vocal filters were different between the synthesis step (Kawahara's method: 44100-Hz sampling frequency with 44th order) and the analysis step (ARX-LF: 12000-Hz sampling frequency with 14th order).

Table1 shows that the estimation accuracy of the proposed method is higher than that of IAIF with Dyprog-LF.

4.2 Natural vowels

The voiced vowel (/a/) was pronounced by five male and five female Japanese speakers. The speech signals were recorded together with EGG signals. Thus, ten real voiced vowels were used to test the proposed method and the IAIF with Dyprog-LF.

4.2.1 Results and discussion

There is no direct reference parameter available for the glottal sources and vocal tracts in real vowels. Therefore, to evaluate glottal sources, as a reference value, we calculated the open quotient (OQ) to evaluate the accuracy of the proposed method. The recorded vowels were analyzed by the proposed method to estimate T_e, which is often considered the OQ_{LF}, and referenced OQ_{EGG} was calculated from a differentiated EGG (dEGG) signal by examining the glottal opening instant (GOI) and GCI. Thus, the estimation errors can be calculated by Eq. 6. The average estimation errors (ε) are shown in Fig. 4. Compared with the value of OQ obtained from the dEGG signal, the accuracy of the proposed method was higher than that of IAIF with Dyprog-LF.

Vocal tract parameters F_1 and F_2 were estimated by the proposed method and a widely used formant extractor (Praat), respectively. Results are shown in Fig. 5, where the values of F_1 and F_2 estimated by the proposed method were very similar to those extracted by Praat. Furthermore, for ten speakers, most of the F_1 values estimated by the proposed method were a little higher than those estimated by Praat, and the F_2 values of the two methods were mostly the same. This results indicates that the proposed method can effectively estimate the vocal tract parameters.

We also tried using the proposed method for a more advanced operation by examining a continuous speech (/aiueo/) pronounced by a male speaker. It is impossible to discuss glottal source parameters since there was no EGG signal recorded together with a speech signal. The waveform and the spectrogram of the original speech signal and the resynthesized speech signals by the ARX-LF model were plotted in Fig. 6. As shown, the original speech signal was very similar to the resynthesized speech signal in the time and frequency domains, and the signal-to-noise ratio (SNR) between the original speech signal and residual (noise) was 23.1 dB. The spectrogram clearly shows that the formant frequencies were the same as the original one, which demonstrates the high accuracy of the proposed method in estimating the vocal tracts of continuous speech. The distance of mel-frequency cepstrum coefficients (order=14) between the original speech and the resynthesized speech signal was 0.93 dB. The synthesized speech could be perceived as well as the original one simply by an informal perception test. This demonstrates that the proposed method is also suitable for

continuous speech. The slight difference between the original speech and the resynthesized one may stem from using only the ARX-LF model, in which $e(n)$ was not added in the synthesis process.

The above results demonstrate that the proposed method has a higher estimation accuracy than IAIF with Dyprog-LF, and that it can be utilized for continuous speech.

5 Conclusion

In this paper, we proposed a simultaneous estimation of glottal source waveforms and vocal tract shapes from speech signals based on the ARX-LF model. The estimation procedures consists of two steps: first, obtaining the initial values of glottal source parameters, in which an inverse filter and the LF model are used, and second, using a simultaneous estimation procedure to obtain accurate glottal sources and vocal tract parameters with the ARX-LF model. We tested the proposed method and the IAIF with Dyprog-LF method for both the real and synthesized vowels and found that the proposed method has a higher estimation accuracy. Moreover, the proposed method can be utilized for continuous speech. In future work, we will apply the proposed method to voice conversion.

Acknowledgements This study was supported by a Grant-in-Aid for Scientific Research (A) (No. 25240026), JST-Mirai Program (JP-MJMI18D1) and China Scholarship Council (CSC).

References

1. Cohen, J., Kamm, T., & Andreou, A. G. (1995). Vocal tract normalization in speech recognition: Compensating for systematic speaker variability. *The Journal of the Acoustical Society of America*, 97(5), 3246–3247.
2. Raitio, T., Suni, A., Pulakka, H., Vainio, M., & Alku, P. (2011). Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4564–4567.
3. Drugman, T., Dubuisson, T., & Dutoit, T. (2009). On the mutual information between source and filter contributions for voice pathology detection. *Tenth Annual Conference of the International Speech Communication Association.*
4. Childers, D. G. (1995). Glottal source modeling for voice conversion. *Speech communication*, 16(2), 127–138.
5. Plumpe, M. D., Quatieri, T. F., & Reynolds, D. A. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, 7(5), 569–586.
6. Iliev, A. I., Scordilis, M.S., Papa, J. P., & Falcão, A. X. (2010). Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech & Language*, 445–460.
7. Li, X., & Akagi, M. (2018). A Three-Layer Emotion Perception Model for Valence and Arousal-Based Detection from Multilingual Speech. *In interspeech*, 3643–3647.
8. Fant, G., Liljencrants, J., & Lin, Qi.g. (1985). A four-parameter model of glottal flow. *STL-QPSR*, 4. 1–13.
9. Rabiner, L.R., & Schafer, R. W. (1987). Digital processing of speech signals. *Prentice-hall Englewood Cliffs, NJ*, 100.
10. Wong, D., Markel, J., & Gray, A. (1979). Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4), 350–355.
11. Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2-3), 109–118.
12. Drugman, T., Bozkurt, B., & Dutoit, T. Complex cepstrum-based decomposition of speech for glottal source estimation. *Interspeech*, 116-119.
13. Kane, J., & Gobl, C. (2013). Automating manual user strategies for precise voice source analysis. *Speech Communication*, 55(3), 397–414.
14. Klatt, D.H., & Klatt, L.C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, 87(2), 820-857.
15. Fujisaki, H., & Ljungqvist, M. (1986). Proposal and evaluation of models for the glottal source waveform. *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 11, 1605–1608.
16. Ding, W., Kasuya, H., & Adachi, S. (1995). Simultaneous estimation of vocal tract and voice source parameters based on an ARX model. *IEICE transactions on information and systems*, 78(6), 738–743.
17. Fujisaki, H., & Ljungqvist, M. (1996). Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform. *Recent research towards advanced man-machine interface through spoken language.* 52–60.
18. Fröhlich, M., Michaelis, D., & Strube, H. W. (2001). SIM-simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. *The Journal of the Acoustical Society of America*, 110(1), 479–488.
19. Vincent, D., and Rosec, O., & Chonavel, T. (2005). Estimation of LF glottal source parameters based on an ARX model. *Ninth European Conference on Speech Communication and Technology*, 333-336.
20. Fu, Q., & Murphy, P. (2006). Robust glottal source estimation based on joint source-filter model optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2), 492–501.
21. Fant, G. (1995). The LF-model revisited. Transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2(3), 119–156.
22. Li, Y., and Sakakibara, K.I., Morikawa, D., & Akagi, M. (2017). Commonalities of glottal sources and vocal tract shapes among speakers in emotional speech. *International Seminar on Speech Production*, 24-34.
23. Takahashi, K., & Akagi, M. (2018). Estimation of glottal source waveforms and vocal tract shape for singing voices with wide frequency range. *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1879–1887.
24. Drugman, T., Thomas, M., Gudnason, J., Naylor, P., & Dutoit, T. (2012). Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3), 994–1006.

25. Kane, J., Yanushevskaya, I., Ní Chasaide, A., & Gobl, C. (2012). Exploiting time and frequency domain measures for precise voice source parameterisation. *Speech Prosody 2012*, 143–146.
26. Lu, H.L. (2002). Toward a high-quality singing synthesizer with vocal texture control. *Stanford University*.
27. Kawahara, H., Sakakibara, K.I., Banno, H., Morise, M., Toda, T., & Irino, T. (2015). Aliasing-free implementation of discrete-time glottal source models and their applications to speech synthesis and F0 extractor evaluation. *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*, 520–529.
28. Drugman, T., Bozkurt, B., & Dutoit, T. (2012). A comparative study of glottal source estimation techniques. *Computer Speech & Language*. 26(1), 20–34.