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

Title	Estimation of glottal source waveform and vocal tract shape for singing-voice analysis
Author(s)	Takahashi, Kyoko; Akagi, Masato
Citation	2018 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP2018): 691-694
Issue Date	2018-03-07
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/15088
Rights	Copyright (C) 2018 Research Institute of Signal Processing, Japan. Kyoko Takahashi and Masato Akagi, 2018 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP2018), 2018, 691-694.
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

2018 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing (NCSP2018)
Honolulu, Hawaii, USA, March 4-7, 2018

Estimation of glottal source waveform and vocal tract shape for singing-voice analysis

Kyoko Takahashi and Masato Akagi

Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology
1–1 Asahidai, Nomi, Ishikawa 923–1292 Japan
E-mail: { kyoko.takahashi, akagi } @jaist.ac.jp

Abstract

In this paper, an effective method to estimate the glottal source waveform and the vocal tract shape in singing voice was proposed based on ARX-LF model. Previous methods suffered from estimation of the glottal source waveform and the vocal tract shape in singing voices with high fundamental frequencies because of effects from forwarded periods. In the proposed method, parameters of the ARX-LF model were estimated accurately with exhaustive search in determined range and a simulated annealing method. Additionally, singing voice was re-synthesized using the estimated results of the vocal tract filter and periodic glottal source waveform with a length of settling time for considering the effects from forwarded periods. As a result of analysis using simulated singing voice data and actual sung voice data, the accuracy of estimation of the parameter values of the ARX-LF model from singing voices with wide range of fundamental frequency can be achieved by the proposed method.

1. Introduction

Speech and singing voice are defined as an output of a vocal tract filter with a glottal source based on source-filter theory [1]. Temporal fluctuations of glottal vibration and vocal tract shapes can be obtained by estimating glottal source waveform and vocal tract filter. Several analysis methods have been proposed based on the theory. Ding and Kasuya proposed a speech analysis-synthesis method based on autoregressive with exogenous input (ARX) model, which was a model of vocal tract filter, and Rosenberg-Klatt (RK) model, which was a model of glottal source waveform [2]. Their method accurately estimated the vocal tract filter using the Kalman filer algorithm. Ohtsuka and Kasuya improved estimation to be able to analyze high-pitch speaking voice using the Least Square method [3]. They reported that the method can analyze voices spoken by females and children as well as male voices. Vincent et al. proposed another method for speech analysis and synthesis based on the ARX model and Liljencrants-Fant (LF) model [4]. They analyzed simulated speech data and female speaking voice. Their method accurately estimated speech data with low fundamental frequency.

Several methods based on the ARX-LF model have been reported for singing voice analysis and synthesis. Lu and Smith Ⅲ focused on glottal aspiration noise in singing voice. They proposed the method for extraction and synthesis of the glottal aspiration noise in singing voices [5]. Motoda and Akagi investigated features of glottal source waveform in each vocal register by analyzing singing voices using the ARX-LF model [6]. As a result, differences of glottal source waveforms were found among vocal registers.

However, the previous methods [5, 6] suffered from accurate estimation of the glottal source waveform of singing voice. Inaccurate detection of the glottal closure instant (GCI) caused inaccurate estimation of the glottal source waveform. Li et al. detected GCI using Electroglottogram (EGG) signal and estimated parameter values of the ARX-LF model of emotional speech [7]. Nevertheless, their method cloud not accurately estimate the speech with high fundamental frequency (fo).

Singers can sing songs fluently changing fo in their voice freely not only within one vocal register but also among several registers. Therefore, estimation of the glottal source waveform and the vocal tract shape accurately in wide fo range is required to obtain characteristics of singing voices.

The objective of this paper is to propose a method of accurate estimation of the glottal source waveform and the vocal tract shape for singing voices whose fos are in wide range. In this paper, effects of forwarded periods are suggested as a cause of inaccurate estimation of singing voice with high fo. In singing voices with high fo, the settling time of the vocal tract filter exceeds a length of each periods and the responses of the vocal tract filter of the forwarded periods leak into the target period. Thus, singing voice is re-synthesized using the estimated ARX-LF parameter values to estimate the leaked components.

2. Estimation method for glottal source waveform and vocal tract

2.1 ARX-LF model

The LF model represents derivative of glottal source sig-

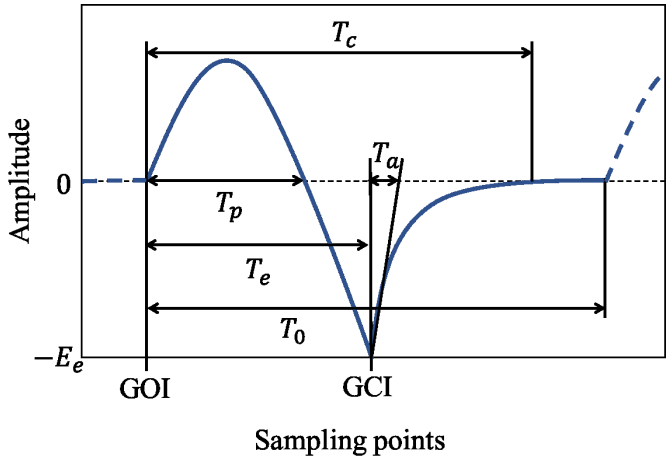

Figure 1: The Liljencrant-Fant model

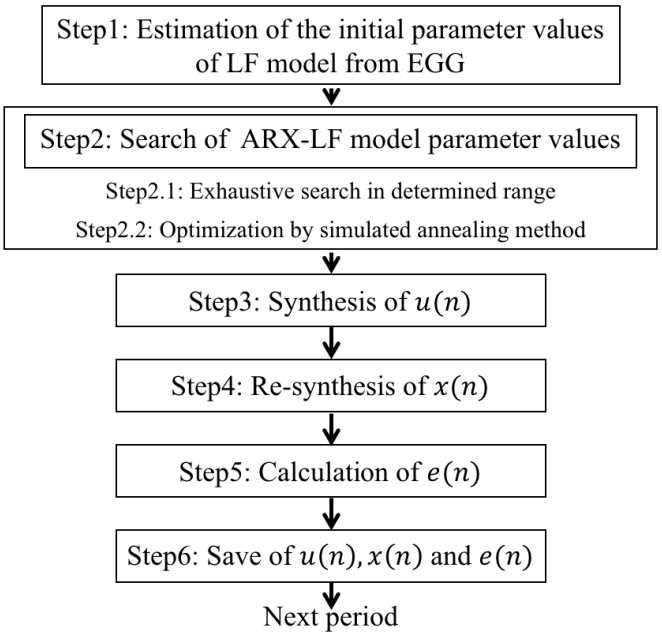

Figure 2: Analysis process of the proposed method

nal with 6 parameters [8, 9]: five parameters concerning time T_p, T_e, T_a, T_c and T_0, and one parameter concerning amplitude E_e as shown in Fig. 1. T_p is the phase where maximum value of glottal flow occurs, T_e is the open phase of the glottis, T_a is the return phase, T_c is the end of return phase and T_0 is the length of period. In Fig. 1, GOI is the glottal opening instant. The LF model in time domain is defined as the following equation:

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

where the parameters E_1, E_2, a, b and ω are related T_p, T_e, T_a, T_c, T_0 and E_e. The ARX model simulates a vocal tract filter. The speech signal $s(n)$ is simulated as a following equation by means of an ARX model [2]:

$$s(n) + \sum_{k=1}^{p} a_k(n) s(n-k) = u(n) + e(n) \quad (2)$$

where $a_k(n)$ is time-varying coefficients of the pth-order AR filter characterizing the vocal tract, $u(n)$ is the glottal flow derivative (periodic waveform) and $e(n)$ is the residual signal of the ARX model and the aspiration noise (aperiodic waveform). The output of the LF model is the input signal $u(n)$ to the vocal tract filter. The re-synthesized signal $x(n)$ is represented as the following equation:

$$x(n) = \sum_{k=1}^{p} a_k(n) s(n-k) + u(n) \quad (3)$$

2.2 Procedures for estimation

Figure 2 shows overview of the proposed procedure for estimation of the glottal source waveform and the vocal tract filter. In steps 1 and 2, parameter values of the ARX-LF model are estimated. Singing voice is re-synthesized to consider influence from forwarded periods in steps 2 and 4.

Step 1 is the phase for estimating of the initial parameter values of the LF model using EGG signal to determine the range of exhaustive search. According to the previous method [7], the initial parameter values of the LF model are obtained from a fixed GCI from a differential EGG. The ranges of exhaustive search are determined using these initial parameters of the LF model.

Step 2 is the phase for searching the optimum values of each parameter of the ARX-LF model. In step 2.1, value of each parameter of the ARX-LF model is explored using exhaustive search in each determined range. In step 2.2, the parameter values of the ARX-LF model are optimized using a simulated annealing method in the small range including the obtained parameters in step 2.1. The conditions of the exhaustive search and the simulated annealing are expressed as the following equations:

$$\begin{aligned} &minimize && f = \sum \{s(n) - x(n)\}^2 \\ &limitation && 0 < T_p < T_e < T_0 \\ & && 0.8 < T_c/T_0 < 1 \\ & && 0.01 < T_a/T_0 < 1 \end{aligned} \quad (4)$$

$u(n)$ and $x(n)$ are obtained by the same process as steps 3 and 4.

Steps 3, 4 and 5 are the phases for synthesizing $u(n)$, re-synthesizing $x(n)$ and calculating $e(n)$. In steps 3, 4 and 5, the vocal tract shape is assumed to be time-invariant during several periods. First of step 3, $u(n)$ is synthesized using the estimated the parameters of the LF model. Second, $u_l(n)$ with a length of over settling time is made using $u(n)$. In step 4, $x_l(n)$ is re-synthesized by inputting $u_l(n)$ into the estimated vocal tract filter. $x(n)$ is the backward periods of $x_l(n)$. In step 5, $e(n)$ is calculated using $s(n)$, $x(n)$ and the estimated vocal tract filter.

3. Evaluation

3.1 Experiment using simulated singing voice

Simulated singing voice were prepared using Kawahara's method [10]. The LF parameters were set as follows; T_e/T_0 is 0.3 to 0.5 with steps of 0.1 and $1/T_0$ (= fo) is fixed as 147, 221, 441 Hz. Each glottal source waveform is assumed as ideal condition without aspiration noise of glottis. This means that the power of the minimized error is 0 theoretically. The power of the minimized error $\varepsilon(n)$ is expressed as the following equation:

$$\varepsilon(n) = \frac{1}{M}\sum e(n)^2 \tag{5}$$

where M is the number of samples in $e(n)$. The typical vowel /a/ is considered for the filter. The sampling frequency of the simulated singing voices was 44.1 kHz.

Figure 3(a) shows the power of the minimized error by the previous method [6] in each simulated condition. Figure 3(b) shows the power of the minimized error by the proposed method. As Fig. 3(b), the power of the minimized error was almost 0 in each condition, especially the results of data with fo as 147 Hz and 221 Hz. Comparing Figs. 3(a) and (b), the power of the minimized error by the proposed method was smaller than that by the previous method: decreasing in 91.8% for 147 Hz, decreasing in 84.2% for 221 Hz and decreasing in 71.9% for 441 Hz. As a result, accurate estimation of singing voices in wide range fo is achieved using the proposed method.

3.2 Experiment using the real singing voice

Real singing voices were baritone singing voice sung with vowel /a/. One singing voice waveform is shown in Fig. 4(a). This data was offered by Prof. Tsuzaki, Kyoto City University of Arts, and included singing voice waves and EGG signals. The fo of this data was 233 Hz estimated by STRAIGHT (STRAIGHTV40_006b) [11]. The sampling frequency of the singing voice data was 44.1 kHz.

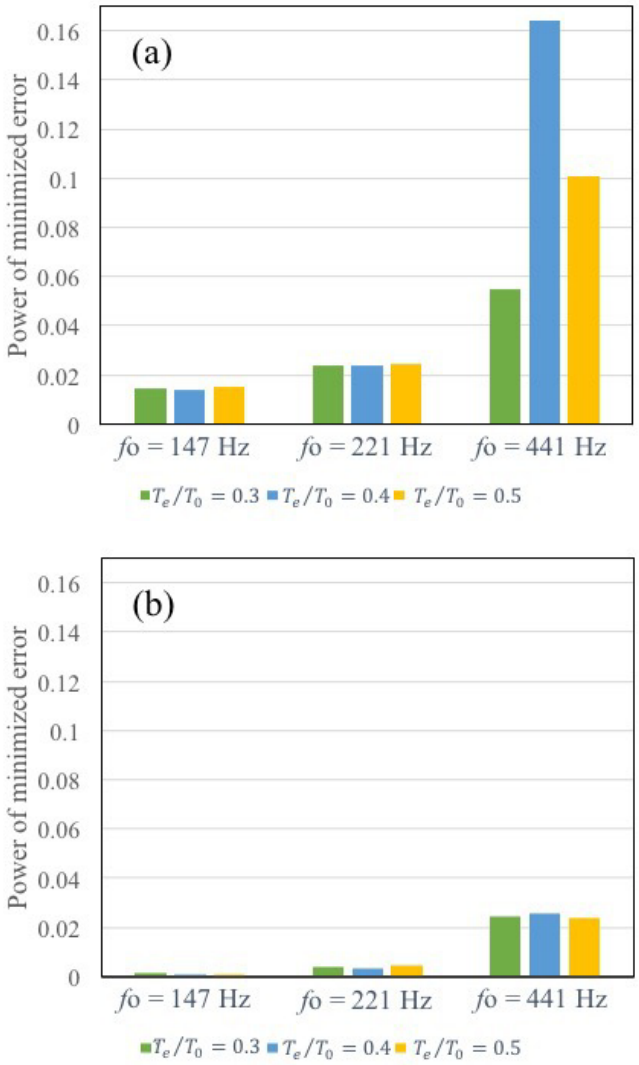

Figure 3: Power of minimized error of simulation data, (a) by the previous method and (b) by the proposed method

Figure 4(b) and (c) shows the estimated $e(n)$ by the previous method and the proposed method. $e(n)$ by the previous method was observed a periodic component shown in Fig. 4(b). $e(n)$ by the proposed method was observed only an aperiodic waveform in Fig. 4(c). Theoretically, $u(n)$ consists of periodic components and $e(n)$ consists of aperiodic components. Therefore, Fig. 4(c) is indicated that the proposed method could estimate glottal source waveform accurately.

4. Conclusions

This paper proposed an effective method to estimate the glottal source waveform and the vocal tract filter of singing voices. Two evaluational experiments were carried out using

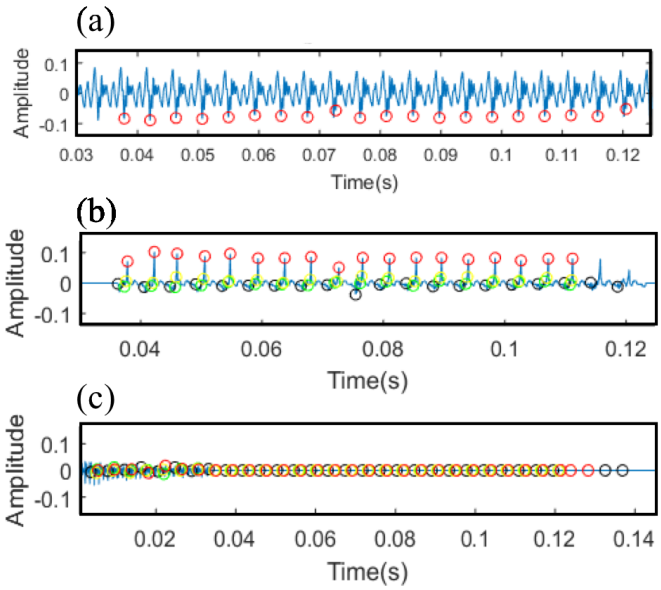

Figure 4: The results of a baritone singing voice /a/, (a) real singing voice, (b) estimated result of $e(n)$ by the previous method, (c) estimated result of $e(n)$ by the proposed method

the simulated singing voice whose *f*o are low, middle and high, and the real singing voice. Reduction of the power of the minimized error was found in the results of the simulated singing voice. As a result, it is indicated that the proposed method could accurately estimate the glottal source waveform and the vocal tract filter of singing voice with wide range of *f*o. The results of the real singing voice indicated that the proposed method could be applied to estimate the parameter values in real singing voices.

Acknowledgments

This study was supported by a rant-in-Aid for Scientific Research (A) (No. 25240026). The authors would like thank Prof. Minoru Tsuzaki and Jun Takahashi, in Kyoto City University of Arts, for offering the sung-voice data.

References

[1] G. Fant: The source filter concept in voice production, STL-QPSR, Vol. 22, No. 1, pp. 21–37, 1981.

[2] W. Ding and H. Kasuya: Simultaneous estimation of vocal tract and voice source parameters based on an ARX model, IEICE TRANS. INF. & SYST. Vol. E78–D, No. 6, 1995.

[3] T. Ohtsuka and H. Kasuya: An Improved speech analysis-synthesis algorithm based on the auto regressive with exogenous input speech production model, 6th International Conference on Spoken Language Processing (ICSLP 2000), Vol. 2, pp. 787-790, 2000.

[4] D. Vincent, O. Rosec and T. Chonavel: Estimation of LF glottal source parameters based on ARX model, Interspeech, pp. 333–336, 2005.

[5] H. Lu and J. O. Smith Ⅲ: Glottal source modeling for singing voice synthesis, in Proceedings of the 2000 International Computer Music Conference, Vol. 2000, 2000.

[6] H. Motoda and M. Akagi: A singing voices synthesis system to characterize vocal registers using ARX-LF model, 2013 International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'13), Hawaii, USA, pp. 93–96, 2013.

[7] Y. Li, K. Sakakibara, D. Morikawa and M. Akagi: Commonalities of glottal sources and vocal tract shapes among speakers in emotional speech, The 11th International Seminar on Speech Production (ISSP 2017), Tianjin, China, 2017.

[8] G. Fant, J. Liljencrants and Q. Lin: A four-parameter model of glottal flow, STL-QPSR, Vol. 26, No. 4, pp. 1–13, 1985.

[9] 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.

[10] 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, Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific, Hong Kong, pp. 520–529, 2015.

[11] H. Kawahara: STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds, Acoustic Science and Technology, Vol. 27, No. 6, pp. 349-353, 2006.