Estimation of glottal source waveforms and vocal tract shape for singing voices with wide frequency range

Kyoko Takahashi* and Masato Akagi*
* Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: { kyoko.takahashi, akagi }@jaist.ac.jp

***Abstract*—Estimation of glottal vibration and vocal tract for singing voices is necessary for clarifying the mechanism of singing voice production. However, accurate estimation of glottal vibration and vocal tract shape in singing voices with a high fundamental frequency (f0) is difficult using simulated models such as the auto-regressive with exogenous input (ARX) model and Liljencrants–Fant (LF) model. This is caused by two problems: the inaccurate estimation method of the glottal closure instant (GCI) and the inappropriate estimation method of ARX model parameter values in singing voices with high f0. Therefore, this proposed method aims to accurately estimate glottal source waveforms and vocal tract shape for singing voices with wide frequency range. To achieve this objective, we propose two solutions: estimation of GCI using an electroglottogram (EGG) signal and estimation of ARX model parameter values using multi-stage optimization and an evaluation function including the leaking effect from forwarded periods. In experiments using simulated singing voices and real singing voices, it was indicated that the accurate estimation of GCI, the reliable estimation of the parameter values of the ARX model for singing voices with high f0, and the estimation of glottal vibration and vocal tract shape in singing voices with wide frequency range were achieved by the proposed method.**

I. Introduction

Singing voices are an essential communication tool and an essential factor in our lives. Moreover, non-linguistic information is an indispensable factor for smooth communication and is presented in singing voices more than speech. Additionally, singing voices are characterized by many kinds of voice quality, a higher fundamental frequency (f0) than that of speech, and sharp time-fluctuation of vocal quality and f0. However, the mechanism of singing voice production has not yet been completely clarified, which necessitates observing and analyzing the state of the speech production system during the singing of a song. Clarification of the mechanism of singing voice production can contribute to the clarification of the mechanism of production of non-linguistic information.

Many approaches have been taken to investigate the state of the glottis and vocal tract for speech. Direct observation of the glottis and vocal tract is carried out using MRI and a high-speed camera. This method can easily obtain the state of glottal vibration and vocal tract shape; however, it imposes an enormous drain on participants. An electroglottogram (EGG) signal can easily obtain the opening and closure instants of the glottis without causing stress; however, it cannot get detailed glottal conditions and vocal tract shape. Estimation using simulated models of glottal source waveforms and vocal tract shape depends on the quality of the simulated model, and using simulated models is not stressful to participants because analysis is performed using only speech signals.

Speech and singing voices are defined as an output of a vocal tract filter with a glottal source on the basis of the source-filter theory [1]. For simulation of the derivative of the glottal source signal, Fant *et al.* proposed the Liljencrants–Fant (LF) model [2] as shown in Fig. 1, and Klatt proposed the Rosenberg–Klatt (RK) model [3]. For estimation of vocal tract shape, Markel and Gray proposed the linear prediction coding (LPC) method [4], and Ding and Kasuya proposed a simulation model of speech production using the auto-regressive with exogenous input (ARX) model as shown in Fig. 2 [5].

Ding and Kasuya proposed a speech analysis-synthesis method based on the ARX-RK model [5]. Their method accurately estimated the vocal tract filter using the Kalman filer algorithm. Ohtsuka and Kasuya improved estimation to be able to analyze speaking voices with high f0 using the least square method [6]. They confirmed that the method can analyze voices spoken by females and children as well as males. Vincent *et al.* proposed another method for speech analysis and synthesis based on the ARX-LF model [7]. They analyzed simulated speech data and female speaking voices. Their method accurately estimated speech data with low f0. The simple analysis and simulation of glottal vibration and vocal tract shape were achieved by the simulated model.

Several methods based on the ARX-LF model have been reported for singing voice analysis and synthesis. Lu and Smith Ⅲ proposed the method for extraction and synthesis of the glottal aspiration noise in singing voices [8]. Motoda and Akagi investigated features of glottal source waveforms in each vocal register by analyzing singing voices using the ARX-LF model [9]. As a result, differences of glottal source waveforms were found among vocal registers.

Analysis and simulation of glottal vibration and vocal tract shape in singing voices have been difficult using the simulated model because of a large amount of aspiration noise and high f0 in singing voices. In the previous methods [8], [9], the inaccurate estimation of glottal source waveforms and vocal tract shape in singing voices with high f0 was caused by two problems: inaccurate estimation of the LF model parameter

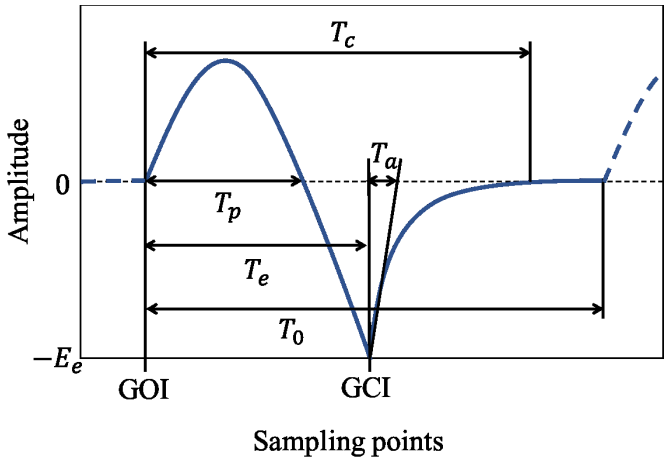


Fig. 1. Typical waveform from Liljencrants–Fant (LF) model

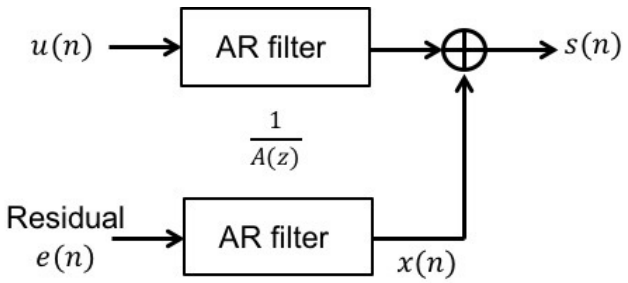


Fig. 2. Speech production process in auto-regressive with exogenous input (ARX) model

values because of the inaccurate estimation method of the glottal closure instant (GCI) and the inappropriate estimation method of ARX model parameter values in singing voices with high f0. Estimation of glottal source waveforms and vocal tract shape is difficult in typical singing voices with sharp time-fluctuation of f0 using the previous method.

Therefore, this proposed method aims to estimate glottal source waveforms and vocal tract shape for singing voices with wide frequency range. In order to achieve this objective, accurate estimation of GCI and reasonable estimation of ARX model parameter values in singing voices with high f0 are required. For the first problem, we suggest two solutions: 1) estimation of GCI using EGG signals and singing sound waves, and 2) synthesis of glottal source waveforms with a high value of the sampling frequency to express GCI in detail. For the second problem, we suggest two solutions: 1) a set of an evaluation function for the parameter values with the leaking effect from forwarded periods, and 2) estimation of ARX model parameter values using multi-stage optimization because of the time fluctuation of glottal source waveforms. The novelty of this study is the accurate estimation of glottal source waveforms and vocal tract shape for singing voices with high f0, which has been difficult until now. Additionally, the successful observation of time fluctuation of glottal vibration and vocal tract shape during singing with wide frequency range using the proposed method is another novelty point of this study. By achieving accurate estimation of glottal source waveforms and vocal tract shape for singing voices with wide range f0, the state of speech production systems can be observed and analyzed for any singing voices, and the proposed method is expected to clarify the mechanism of singing voice production.

II. ARX-LF MODEL

The LF model represents the derivative of the glottal source signal with 6 parameters [2], [10]: 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 the 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 the 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 to T_p, T_e, T_a, T_c, T_0, and E_e. In order to simplify the observation for the time fluctuations of glottal source waveforms, three parameters O_q, α_m, and Q_a were defined as the following equations:

$$O_q = \frac{T_e}{T_0}, \quad (2a)$$

$$\alpha_m = \frac{T_p}{T_e}, \quad (2b)$$

$$Q_a = \frac{T_a}{(1-O_q)T_0}, \quad (2c)$$

where T_0, T_e, T_p and T_a were the parameters of the LF model. O_q corresponds to the open quotient, α_m to the asymmetry coefficient, and Q_a to the return phase quotient.

The ARX model simulates a vocal tract filter. The speech signal $s(n)$ is simulated as the following equation by means of an ARX model [5] as shown in Fig. 2:

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

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:

$$x(n) = -\sum_{k=1}^{p} a_k(n)s(n-k) + u(n). \quad (4)$$

III. Solution for the problems of the previous methods

In the previous methods [8], [9], the inaccurate estimation of glottal source waveforms and vocal tract shape for singing voices with high f0 was caused by two problems: the inaccuracy of estimating GCI and the difficulty of estimating ARX model parameter values in singing voices with high f0. Thus, we propose two solutions for these problems: estimation of GCI using EGG signals and estimation of ARX model parameter values using multi-stage optimization and reduction of the leaking effect from forwarded periods. Figure 3 shows an overview of the proposed procedure for estimation of the glottal source waveform and the vocal tract filter. The first solution is applied to Procedure I in Fig. 3, and the second solution is applied to Procedure II in Fig. 3.

A. The proposed solution for the first problem

In the previous methods [8], [9], the estimation method of GCI was not sufficient for accurate estimation of the LF model parameter values because GCI was estimated from singing sound waves. Li *et al.* proposed the estimation method of glottal source waveforms and vocal tract for emotional speech using the estimation of GCI with EGG signals and archived the accurate estimation of GCI for the speech of three males and a female [11]. However, GCI of singing voices could not be accurately obtained using their method. In the previous method [11], GCI was estimated using an EGG signal whose sampling frequency decreased from 44.1 kHz to 12 kHz. This means that the sampling frequency is not sufficient for expressing GCI because the value of f0 of singing voices is higher than that of speech, and this leads to the short length of each period in singing voices. Therefore, it appears that the value of the sampling frequency for expression of GCI and glottal source waveforms is important for accurately estimating glottal source waveforms.

The proposed estimation method of GCI consists of two steps: calculation of the value of GCI using an EGG signal and singing voice $s(n)$ and adjustment of the sampling frequency of a synthesized glottal source waveform.

The sampling frequency of the EGG signal and $s(n)$ is 44.1 kHz. GCI_{EGG} is calculated using the EGG signal, and GCI_s is calculated using $s(n)$ on the basis of the method proposed by Drugman and Dutoit [12]. With reference to the value of GCI_s, the relationship between GCI_s and GCI_{EGG} is represented by the following equation,

$$GCI_s > GCI_{EGG}, \tag{5}$$

because the EGG signal is measured on the position of the larynx. Consequently, the initial value of GCI for iteration, GCI_0, is expressed as the following equation:

$$GCI_0 = GCI_{EGG} + d, \tag{6a}$$

$$d = \frac{1}{P}\sum(GCI_s - GCI_{EGG}), \tag{6b}$$

where P is the number of periods in the data.

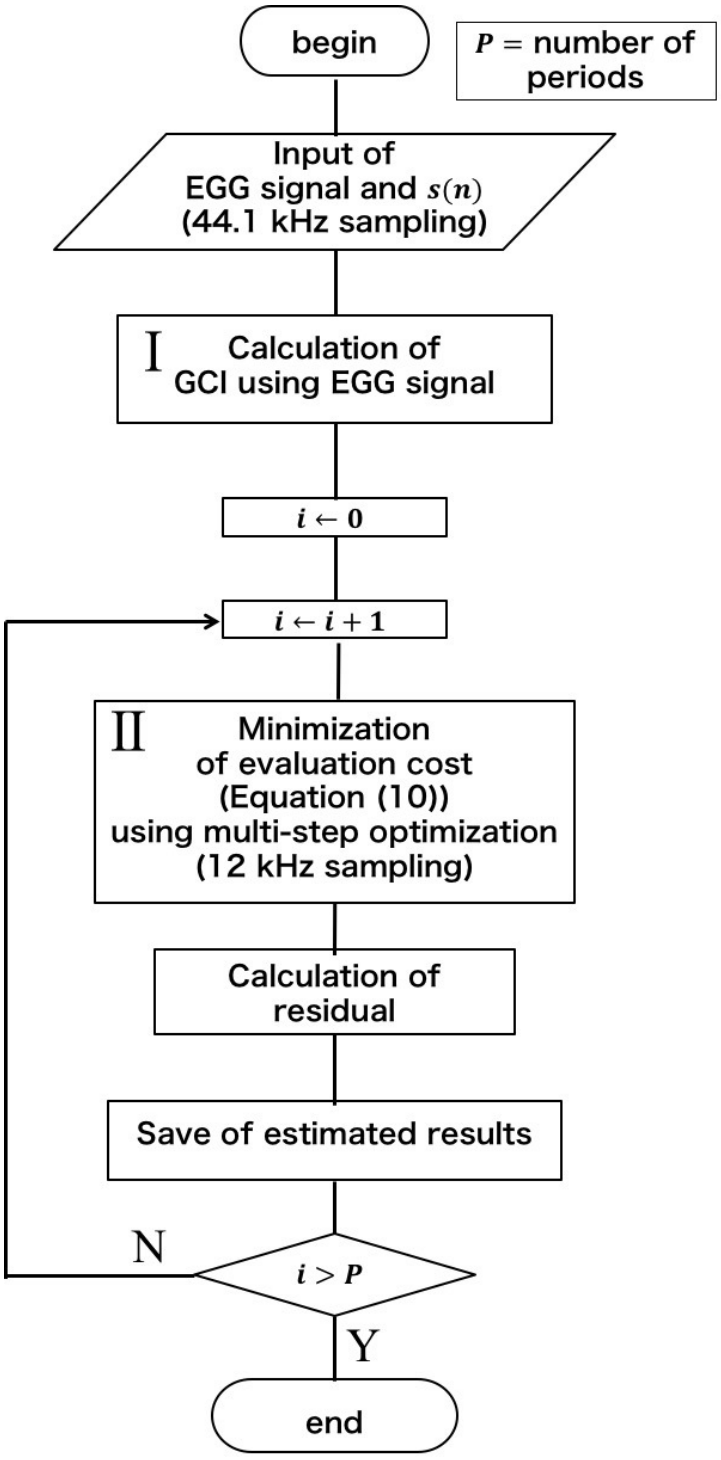


Fig. 3. Process of proposed procedure

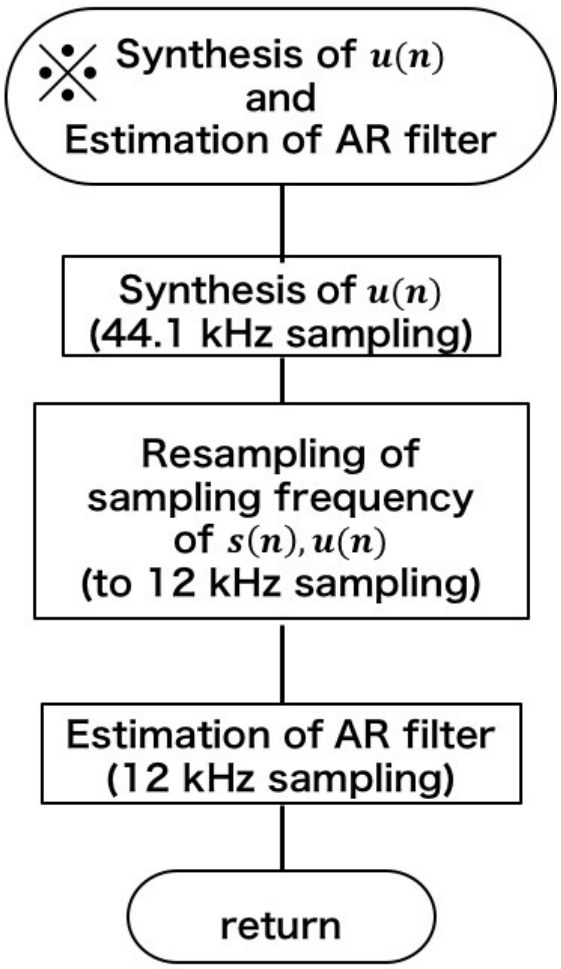


Fig. 4. Synthesis and adjustment process of derivative of glottal source waveform $u(n)$ for estimation of AR filter

Second, the sampling frequency of synthesized $u(n)$ is adjusted before the parameter values of the ARX model are estimated. Figure 4 shows the procedure of this step. In the first step, the sampling frequency is 44.1 kHz for accurate estimation of GCI and detailed expression of GCI in $u(n)$. However, the high sampling frequency leads to bad fitting of the vocal tract filter. Therefore, the sampling frequency of synthesized $u(n)$ is adjusted from 44.1 kHz to 12 kHz when estimating the ARX filter.

B. The proposed solution for the second problem

In the previous methods [8], [9], estimating ARX model parameter values in singing voices with high f0 is difficult because of two reasons: unsuitable optimization and the ignored leaking effect from forwarded periods.

In analysis of glottal vibration and vocal tract shape using the simulated models, the glottal source waveform and vocal tract filter for speech and singing voices are estimated simultaneously. Furthermore, the LF model has the bounds of each parameter value. Therefore, global optimization of the multi-funnel function is necessary for estimation of the glottal source waveform and vocal tract filter. However, it is assumed that the optimization of previous methods [8], [9] leads not to global optima but local optima.

Previous studies [3], [13] reported incomplete closure of the glottis and an increase in aspiration noise in speech and singing voices with high f0. From these facts, it is assumed that the glottal source waveform is different in each period. Furthermore, the length of each period is short in singing voices with high f0. This leads to an unignorable leaking effect from the different responses of each forwarded period.

Therefore, the proposed estimation method of ARX model parameter values consists of two steps: multi-stage optimization of the parameter values of the ARX-LF model and a set of an evaluation function of optimization including the leaking effect from forwarded periods.

1) Multi-stage optimization: Figure 5 shows the estimation procedure of the parameter values of the ARX-LF model using multi-stage optimization. In this procedure, an exhaustive search method and a simulated annealing method are used in a sequence for optimization of the parameter values of the ARX-LF model in detail. In the exhaustive search method, the search range is determined with reference to the initial parameter values of the LF model obtained using the EGG signal. In the simulated annealing method, the initial values of optimization are the optimum values of the exhaustive search method. The conditions of the exhaustive search and the simulated annealing method are expressed as the following equations:

subject to

$$0 < T_p < T_e < T_0, \tag{7a}$$
$$0.8 < T_c/T_0 < 1, \tag{7b}$$
$$0.01 < T_a/T_0 < 1, \tag{7c}$$
$$0.5 < E_e/H < 1.5, \tag{7d}$$
$$GCI_0 - 15 < GCI < GCI_0 + 15, \tag{7e}$$

where H is the amplitude of $s(n)$ at GCI_0.

2) The evaluation function including the leaking effect from forwarded periods: In the evaluation of the parameter values for optimization, resynthesized signal $x(n+l)$ with the leaking effect from forwarded periods is used as shown in Fig. 6. $x(n+l)$ has the length of r-period because of the estimation of the AR filter for singing voices with high f0. In Fig. 6, the red box with the dotted line shows the target period of estimation in the time domain. $\hat{u_0}(x)$ means the glottal source waveform with the length of r-period, $\hat{h_0}(n)$ means the AR filter of the target period, and T_0 means the length of the target period. The re-synthesized signal i periods ago from the target period is expressed as the following equation:

$$x_i(n + \sum_{k=1}^{i} T_k) = u_i(n + \sum_{k=1}^{i} T_k) * h_i(n), \tag{8}$$

where T_k is the length of period k periods ago. The following equation expresses $x(n)$ including the leaking effect from

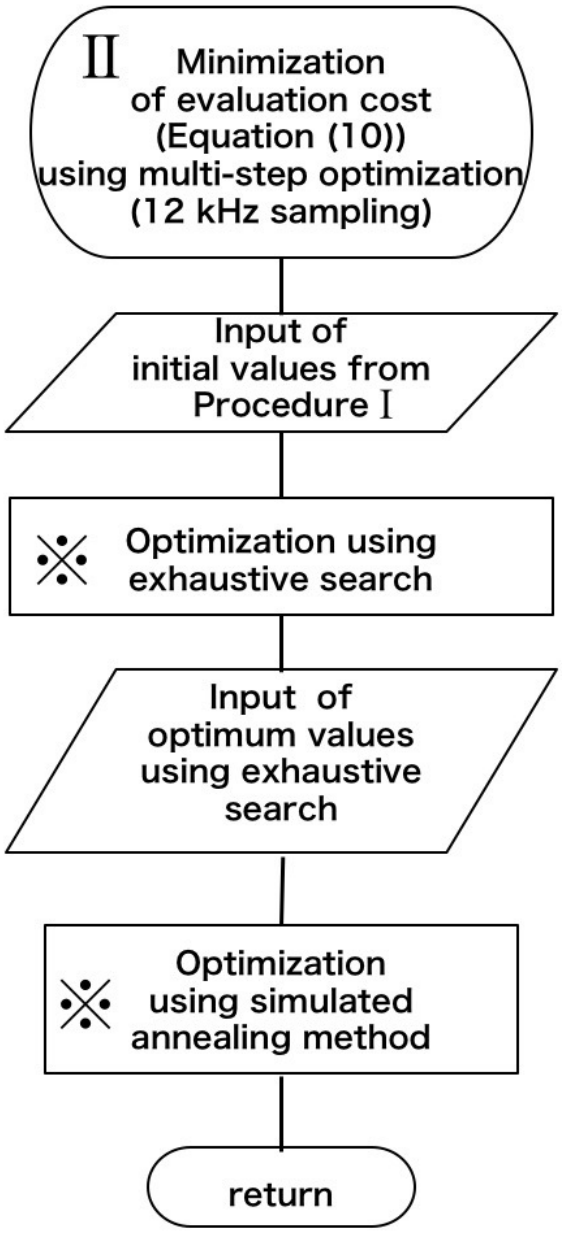


Fig. 5. Process of minimization of evaluation cost using multi-step optimization

forwarded results which were estimated from m periods ago to 1 period ago.

$$x(n) = \sum_{i=1}^{m} \{x_i(n + \sum_{k=1}^{i} T_k)\} + \hat{u_0}(n) * \hat{h_0}(n), \qquad (9)$$

where m is the number of periods before the target period. Therefore, the evaluation function is expressed as the following equation:

$$\text{Minimize} \qquad f = \sum_{l=-rT_0}^{0} \{s(n+l) - x(n+l)\}^2. \qquad (10)$$

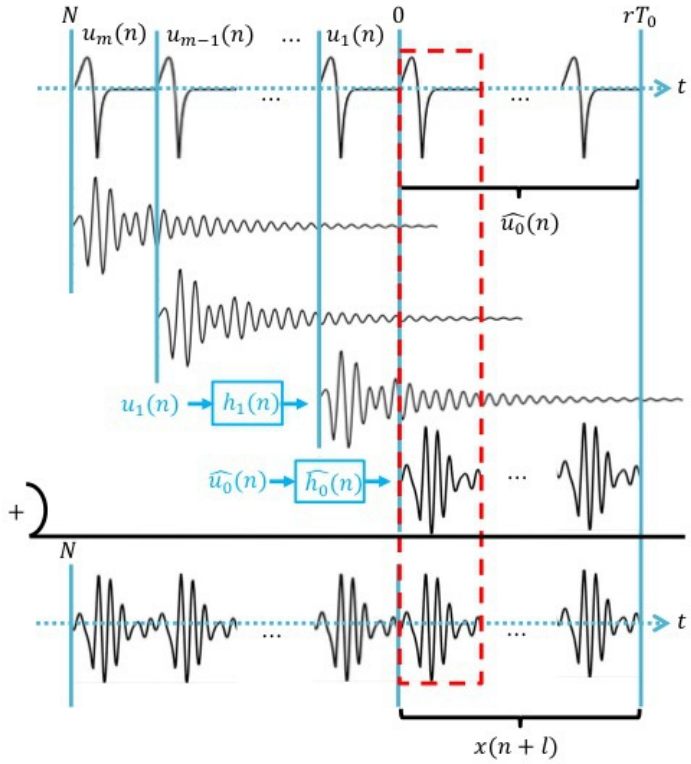


Fig. 6. Re-synthesis procedure including leaking effect from forwarded period for calculation of evaluation cost

IV. Procedures for Estimation

Figure 3 shows an overview of the proposed procedure for estimation of the glottal source waveform and the vocal tract filter. i means a loop counter between 0 and P. The proposed procedure consists of five steps: input of the EGG signal and $s(n)$, estimation of the initial values of the LF model parameters, minimization of the evaluation cost (Equation (10)), calculation of the residual, and a save of all results of this period.

In estimation of the initial values of the LF model parameters, the initial values, GCI_0, GOI_0, and T_0, are estimated using the EGG signal and $s(n)$ on the basis of the steps of section III-A. Then, the O_q is calculated for determination of the search range of an exhaustive search. The calculated O_q is the sharply scattered values in each period because GOI_0 does not appear as a specific value. However, it is likely that O_q cannot change sharply on the basis of the structure of the glottis. Thus, the time sequence of O_q is fitted to a 2nd-order polynomial curve.

The step of minimization of the evaluation cost (Equation (10)) is based on the first steps of Section III-B. In the exhaustive search method, the parameter values of the LF model are searched as follows:

- $t_e(p)$: $O_q(p) - 0.02$ to $O_q(p) + 0.02$ with steps of 0.01,

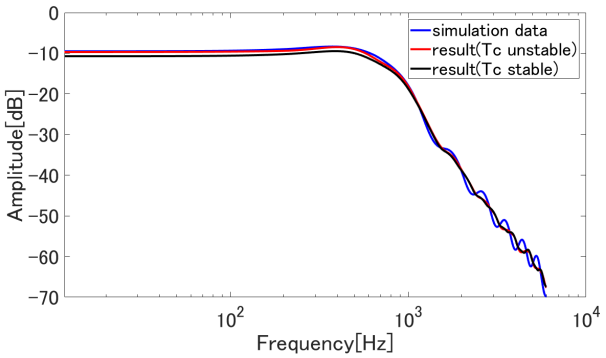


Fig. 7. Spectra of glottal source waveform of simulated singing voice

- $t_p(p)$: $0.65T_e(p)$ to $0.9T_e(p)$ with steps of 0.02,
- $t_a(p)$: 0.01 to 0.1 with steps of 0.01,
- $t_c(p)$: 0.8,
- $E_e(p)$: $0.5H$ to $1.5H$ with steps of $0.01H^2$,
- $GCI(p)$: $GCI_0(p) - 15$ to $GCI_0(p) + 15$ with steps of 1,

where $t_e(p)$, $t_p(p)$, $t_a(p)$, and $t_c(p)$ are defined as the following equations (11).

$$t_e(p) = \frac{T_e(p)}{T_0(p)}, \tag{11a}$$

$$t_p(p) = \frac{T_p(p)}{T_0(p)}, \tag{11b}$$

$$t_a(p) = \frac{T_a(p)}{T_0(p)}, \tag{11c}$$

$$t_c(p) = \frac{T_c(p)}{T_0(p)}. \tag{11d}$$

The ranges of $t_p(p)$ and $t_a(p)$ depend on Fu and Murphy's method [10]. T_c relates to the spectral tilt of $u(n)$. Figure 7 shows the spectrum of estimated $u(n)$ in an experiment using a simulated singing voice that was prepared using Kawahara's method [14].

The blue line shows the spectrum of $u(n)$ of simulation data, the red line shows the spectrum of the estimated $u(n)$ with t_c that was estimated between 0.8 to 1, and the black line shows the spectrum of the estimated $u(n)$ with fixed t_c as 0.8. As a result, the tilt of the spectrum of the estimated $u(n)$ did not change whether or not T_c was a constant value as in Fig. 7. Then, t_c is the constant value for accurately estimating the values of other parameters.

In the simulated annealing method, the lower bounds and upper bounds of each parameter of the LF model are set as follows:

- t_e: $t_e(p) - 0.02$ to $t_e(p) + 0.02$,
- t_p: $t_p(p) - 0.01$ to $t_p(p) + 0.01$,
  - If $t_p(p) - 0.01 \leq 0$, lower bound is set to 0,
  - If $t_p(p) + 0.01 \geq t_e(p) - 0.02$, upper bound is set to $t_e(p) - 0.02$,
- t_a: $t_a(p) - 0.01$ to $t_a(p) + 0.01$,
- t_c: -,
- E_e: $E_e(p) - 0.01H^2$ to $E_e(p) + 0.01H^2$,
- GCI: -.

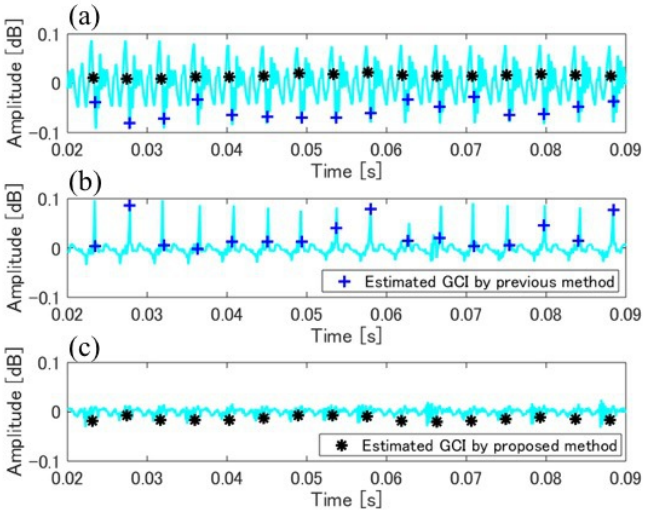


Fig. 8. Results of a baritone singing voice /a/, (a) real singing voice, (b) estimated result of $e(n)$ by previous method, (c) estimated result of $e(n)$ by proposed method

During each optimization as in Fig. 5, $u(n)$ is synthesized, its sampling frequency is adjusted from 44.1 kHz to 12 kHz, and the AR filter is estimated as shown in Fig. 4. These procedures are the second step of Section III-B. The AR filter is estimated as the 14th-order filter using the least square method.

V. Evaluation

A. Experiment using the real singing voice

Real singing voices were baritone singing voices sung with the vowel /a/. One singing voice waveform is shown in Fig. 8(a). These data were offered by Prof. Tsuzaki, Kyoto City University of Arts, and included singing sound waves and EGG signals. The f0 of these data was 233 Hz estimated by STRAIGHT [15]. The sampling frequency of the singing voice data was 44.1 kHz.

Figure 8(b) and (c) show the estimated $e(n)$ by the previous method [9] and the proposed method, respectively. The blue cross marks plotted in Fig. 8(a) and (b) indicate the estimated GCI using the previous method. The black asterisk marks plotted in Fig. 8(a) and (c) indicate the estimated GCI using the proposed method. $e(n)$ by the previous method [9] has

TABLE I
AVERAGE PERCENTAGE OF ERROR OF EACH OF THE ARX-LF MODEL PARAMETER VALUES IN EACH SIMULATION DATASET BY PREVIOUS METHOD [9] [%]

f0	O_q	α_m	Q_a	F_{R1}	F_{R2}
147 Hz	49.5	0.695	80.2	10.4	2.36
221 Hz	38.5	0.666	79.6	11.0	4.96
441 Hz	18.1	12.8	30.0	9.10	14.5

TABLE II
AVERAGE PERCENTAGE OF ERROR OF EACH OF THE ARX-LF MODEL PARAMETER VALUES IN EACH SIMULATION DATASET BY PROPOSED METHOD [%]

f0	O_q	α_m	Q_a	F_{R1}	F_{R2}
147 Hz	3.13	1.17	30.2	5.63	1.11
221 Hz	4.38	1.69	30.2	5.80	1.43
441 Hz	3.91	4.35	36.6	8.84	1.78

a periodic component at the position of each GCI shown in Fig. 8(b). This means that the error of GCI estimated by the previous method [9] is large. In Fig. 8(c), the error of GCI almost disappears by using the proposed method. Therefore, Fig. 8(c) indicates that the proposed method can estimate GCI accurately.

B. Experiment using the simulated singing voice

Simulated singing voices were prepared using Kawahara's method, "SparkNG: Matlab realtime speech tools and voice production tools" [14]. The LF parameters were set as follows: T_e/T_0 was 0.3 to 0.5 with steps of 0.1, and $1/T_0$ (= f0) was fixed at 147, 221, and 441 Hz. Each glottal source waveform was assumed as an ideal condition without aspiration noise of the glottis. This implies that the power of the minimized error was theoretically 0. The power of the minimized error was expressed as the following equation:

$$\text{Power of the minimized error} = \frac{1}{M}\sum e(n)^2, \qquad (12)$$

where M was the number of samples in $e(n)$. The typical vowel /a/ was considered for the filter as follows; the value of the first peak frequency of the filter, F_{R1}, was set to 969 Hz, and the value of the second peak frequency of the filter, F_{R2}, was set to 1184 Hz. The sampling frequency of the simulated singing voices was 44.1 kHz.

Table I shows the average percentage of the error of each of the ARX-LF model parameter values by the previous method [9] in each value of f0. Table II shows the average percentage of the error of each of the ARX-LF model parameter values by the proposed method in each value of f0. Comparing Table I and II, it is indicated that the values of percentage of error of O_q, F_{R1}, and F_{R2} by the proposed method are sufficiently smaller than those by the previous method in any datum. The values of percentage of error of the others are broadly small using the proposed method. As a result, the obtained results estimated using the proposed method are considered reasonable in the ARX model parameters and the part of the LF model parameters.

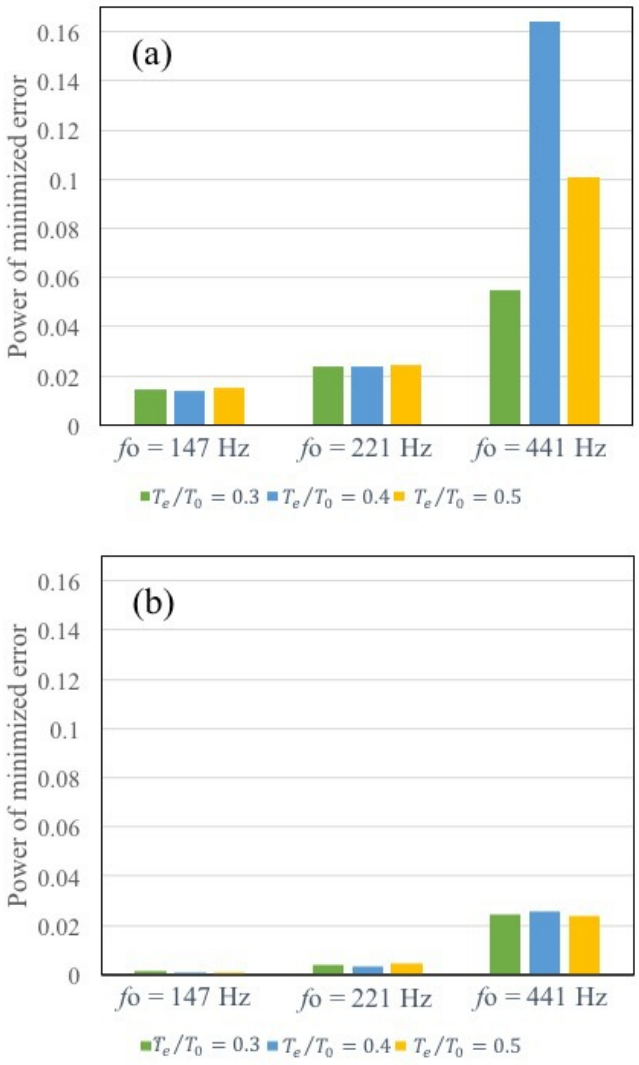
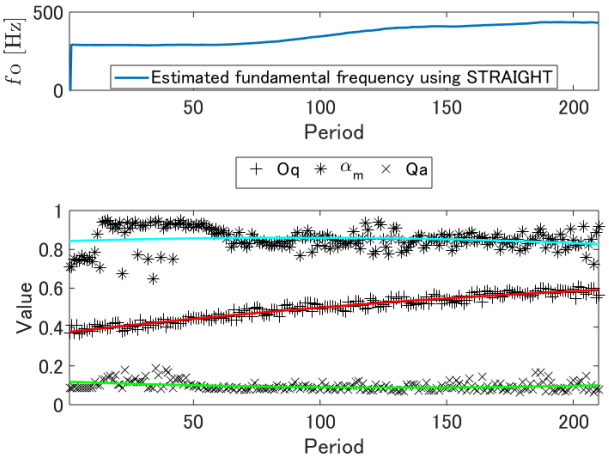
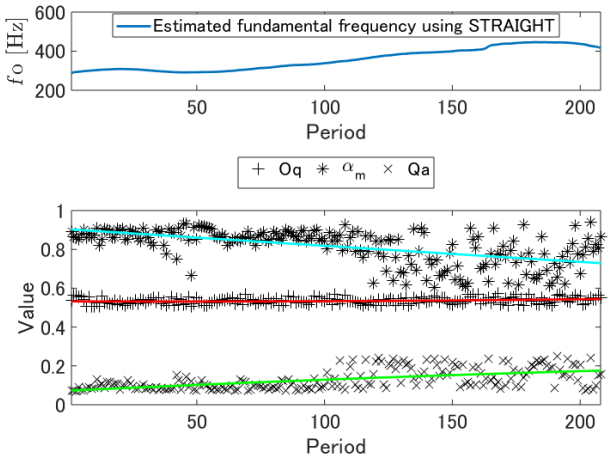


Fig. 9. Average power of minimized error of simulation data, (a) by previous method and (b) by proposed method

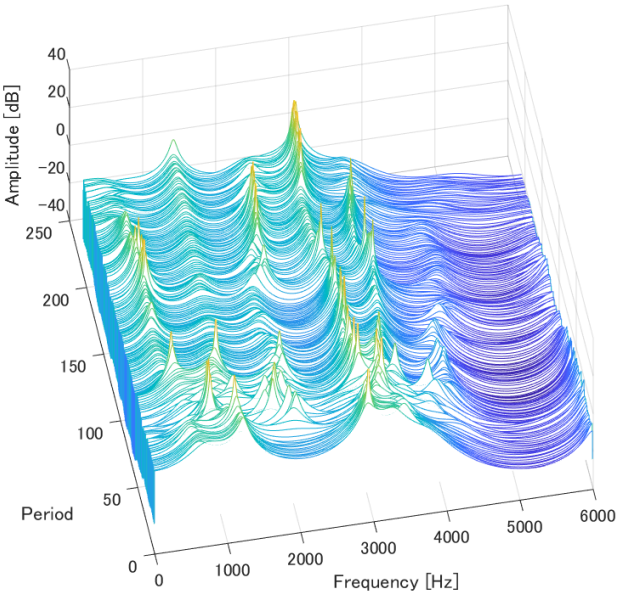
Figure 9(a) shows the average power of the minimized error by the previous method [9] in each dataset. Figure 9(b) shows the average power of the minimized error by the proposed method in each dataset. As Fig. 9(b), the power of the minimized error was almost 0 in each condition, especially the results of data with f0 as 147 Hz and 221 Hz. Comparing Figs. 9(a) and (b), the power of the minimized error by the proposed method was smaller than that by the previous method: decreasing by 91.8% for 147 Hz, decreasing by 84.2% for 221 Hz, and decreasing by 71.9% for 441 Hz. As a result, accurate estimation of singing voices in wide range f0 is achieved using the proposed method.



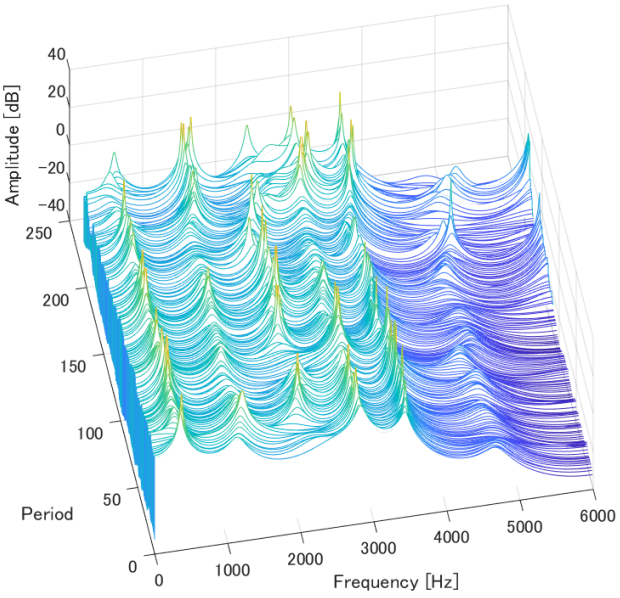
(a) Estimated results of time fluctuations of fundamental frequency and glottal source waveform in tenor singing voice A



(b) Estimated results of time fluctuations of fundamental frequency and glottal source waveform in tenor singing voice B



(c) Estimated results of time fluctuation of vocal tract shape in tenor singing voice A



(d) Estimated results of time fluctuation of vocal tract shape in tenor singing voice B

Fig. 10. Estimated results of glottal source waveform and vocal tract shape for 2 sets of real singing voice data by proposed method

C. Experiment using the real singing voice with wide frequency range

Two real singing voices A and B with a wide frequency range were tenor singing voices sung with the vowel /a/. These data were offered by Prof. Tsuzaki, Kyoto City University of Arts, and included singing sound waves and EGG signals. The first rows of Fig. 10(a) and (b) show the time fluctuations of f0 estimated by STRAIGHT [15] in these data. The sampling frequency of the singing voice data was 44.1 kHz.

The second rows of Fig. 10(a) and (b) show the estimated parameter values of the glottal source waveform in each period. The polynomial curves are fitted to each set of estimated O_q, α_m, and Q_a in each dataset. Figure 10(c) and (d) show the estimated results of frequency responses of the vocal tracts in each period. In Fig. 10, the values of O_q of singing voice A and the values of Q_a of singing voice B increase, and the values of α_m decrease. The scatter of the values of α_m and Q_a in the latter half seems to be reliable in 10(c) and (d) because the time fluctuation of estimated vocal tract shape is an almost contiguous change. As a result, the time fluctuation of the glottal source waveform and vocal tract shape is observed in the singing voice with a wide frequency range by the proposed method.

VI. Discussion

The previous methods [8], [9] had two problems of estimation of glottal source waveforms and vocal tract shape for singing voices with wide frequency range: the inaccuracy of estimating GCI and the difficulty of estimating the ARX model parameter values in singing voices with high f0. These problems lead to the inaccurate estimation of glottal source waveforms and vocal tract shape in singing voices with high f0.

In the evaluation, the following results were confirmed. In the experiment using the real singing voice, the accuracy of estimation of GCI was indicated in Fig. 8. In the experiment using the simulated singing voice, the reasonable estimated results of ARX model parameter values in the simulated singing voice with high f0 were shown in Table II. From the experiment using the simulated singing voice and the real singing voice with a wide frequency range, Fig. 9 and 10 indicate the accurate and reliable estimation of singing voices in wide frequency range. Therefore, it is indicated that the problems of the previous methods [8], [9] are solved by the proposed method.

From Fig. 10, interestingly, the different time-fluctuations are observed in each parameter O_q, α_m, and Q_a. The previous studies indicated the relationship of O_q, α_m, and Q_a between modal and falsetto as Equation (13) [13], [16], [17].

$$O_q : \quad \text{modal} < \text{falsetto} \tag{13a}$$

$$\alpha_m : \quad \text{modal} > \text{falsetto} \tag{13b}$$

$$Q_a : \quad \text{modal} < \text{falsetto} \tag{13c}$$

However, the time fluctuation O_q, α_m, and Q_a in singing voices with f0 changing from modal to falsetto is different from Equation (13) in Fig. 10. Additionally, Henrich *et al.* measured O_q of singing voices using EGG and reported that O_q was small when the glottis was tense and that O_q of modal singing voices was smaller than that of falsetto singing voices [16]. Sakakibara *et al.* reported that the glottis vibrated with unstable and inconstant opening and closure in speech with high f0 [13]. Thus, it is conjectured that the scatter results of α_m and Q_a in Fig. 10 suggest the time fluctuation condition of glottal tension and closure.

VII. Conclusions

This paper proposed a method of estimation of glottal source waveforms and vocal tract shape for singing voices with wide fundamental frequency range. In the analysis experiments for evaluation of the proposed method, estimation of GCI in real singing voices was proved to be accurate, and the estimated results of ARX model parameter values in simulated singing voices with high f0 were proved to be reasonable. This means that the problems of the previous methods [8], [9] are solved by the proposed method. Additionally, the reliable estimated results of singing voices with wide frequency range were obtained in analysis experiments using real singing voices with the time fluctuations of f0. As a result, the proposed method can estimate glottal source waveforms and vocal tract shape for singing voices with wide frequency range.

Acknowledgment

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

[3] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Acoustical Society of America Journal,* Vol. 87, pp. 820–857, 1990.

[4] J. D. Markel and A. H. Jr. Gray, "Linear Prediction of Speech (Communication and Cybernetics)," *Springer,* 2013.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech Conference,* 2009.

[13] K. Sakakibara, H. Imagawa, M. Kimura, H. Yokonishi, and N. Tayama, "Modal analysis of vocal fold vibrations using laryngotopography," in *Eleventh Annual Conference of the International Speech Communication Association,* 2010.

[14] 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.

[15] 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.

[16] N. Henrich, C. Alessandro, B. Doval and M. Castellengo, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," *The Journal of the Acoustical Society of America,* Vol. 117, No. 3, pp. 1417–1430, 2005.

[17] K. Sakakibara, "Production Mechanism of Voice Quality in Singing (Feature Articles; Phonetics and Speech Technology)," *Journal of the Phonetic Society of Japan*, Vol. 7, No. 3, pp. 27–39, 2003.