九州大学学術情報リポジトリ
Kyushu University Institutional Repository

A method for estimating vocal-tract shape from a target speech spectrum

Kaburagi, Tokihiko
Faculty of Design, Kyushu University

https://hdl.handle.net/2324/7178814

出版情報：Acoustical Science and Technology. 36 (5), pp.428-437, 2015. 日本音響学会
バージョン：
権利関係：© 2015 by The Acoustical Society of Japan

KYUSHU UNIVERSITY

Acoust. Sci. & Tech. **36**, 5 (2015) ©2015 The Acoustical Society of Japan

PAPER

A method for estimating vocal-tract shape from a target speech spectrum

Tokihiko Kaburagi*

Faculty of Design, Kyushu University,
4–9–1 Shiobaru, Minami-ku, Fukuoka, 815–8540 Japan

(*Received 4 February 2015, Accepted for publication 23 May 2015*)

Abstract: We present a method to simultaneously estimate the cross-sectional area and length of the vocal tract from a speech spectrum. An iterative procedure determines the vocal-tract shape by gradually optimizing the parameter values to produce the target speech spectrum. The vocal-tract shape is updated in each iteration using a sensitivity function representing the change in formant frequency caused by a slight perturbation of the vocal-tract shape. Our method effectively optimizes the vocal-tract shape when combined with the perturbation relationship between the speech spectrum parameters (i.e., cepstral parameters) and formants. The estimation accuracy is examined using area function data for 10 English vowels (Story and Titze, *J. Phon.*, **26**, 223–260, 1998). The resulting average errors are 0.36 cm^2 for the cross-sectional area and 0.21 cm for the vocal-tract length. This corresponds to a 17.6% and 1.24% error, respectively. The formant frequency recovered from the estimated vocal-tract shape has an error of less than 4% for each of the first four formants. We also determine that the fundamental frequency of the target speech spectrum has an influence on the estimation accuracy.

Keywords: Vocal-tract area function, Vocal-tract length, Speech spectrum, Cepstral parameters, Inverse estimation

PACS number: 43.70.Bk, 43.72.Ct [doi:10.1250/ast.36.428]

1. INTRODUCTION

The acoustic mechanism of human speech production has been researched extensively and physical models have been developed for the vocal folds and vocal tract [1–5]. An example of the application of these models is investigating the interaction between the voice-source system and the vocal-tract filter [6–8]. When producing speech, the transfer function of the vocal tract (and hence formants) is determined by the geometry (length and cross-sectional area) of the vocal tract. Therefore, precise vocal-tract geometry information is required to synthesize speech using a physical model. Magnetic resonance imaging (MRI) can be used to obtain the morphological data of the vocal tract [9]. However, this is typically limited to sustained utterances because of the time required to scan volumetric images.

Estimating the area function from speech without using MRI is valuable for controlling an acoustic tube model of the vocal tract and generating speech signals. To estimate an area function from speech, Schroeder [10] and Mermelstein [11] proposed a perturbation theory based on the wave equation. They represented the vocal-tract shape using the Fourier expansion of the logarithm of the area function. They estimated the cross-sectional area using formants, assuming that the vocal tract is asymmetrical around the center point or is represented by low-order Fourier components. In line with the perturbation theory, Story [12] used a sensitivity function [13,14] representing the change in formant frequency caused by a small perturbation of the vocal-tract shape, and determined the area function from the formants. Story's method was subsequently improved by employing a least squares technique [15]. The length of the vocal tract was estimated simultaneously with the area parameters using the sensitivity function [14,16].

Besides the formant-based methods, the speech spectrum was also used as the acoustic target of vocal-tract estimation. This estimation method is much more convenient than the formant-based ones because the spectral information is readily computed from speech signals, whereas reliable extraction of formant frequencies is still under investigation [17]. However, the relationship between the vocal-tract shape and the speech spectrum is highly nonlinear and complex. Therefore, previous studies did not consider the underlying physical relationship

---

*e-mail: kabu@design.kyushu-u.ac.jp

between the vocal-tract shape and speech spectrum. Instead, the vocal-tract parameter was related to the acoustic parameter using an articulatory-acoustic database [18,19] or a linear [20] or nonlinear [21] mapping method. In the database search approach, a fine discretization of every vocal-tract parameter is needed to relate the articulatory and acoustic parameters accurately. Such discretization, however, increases exponentially both the size of the database and computational cost of the database search. The mapping method requires training of the mapping function using a set of articulatory-acoustic data samples; the determined mapping function is accordingly most effective within the training dataset.

The accuracy and applicability of the database search and mapping methods are highly dependent on the training dataset and the training process itself. To mitigate this problem and construct an estimation method without training the direct relationship between the articulatory and acoustic parameters, we present a method that takes into account the physical constraints during the speech production process. To estimate the vocal-tract shape from the spectral information of the target speech, we use two types of perturbation relationships to combine explicitly the speech spectrum and vocal-tract shape. The spectral information is represented using cepstral parameters that are easily computed from speech signals. The output parameters are the cross-sectional area and length of the vocal tract. We also use formant frequency as an intermediate parameter between the vocal-tract and spectrum parameters.

The first perturbation relationship is the sensitivity function relating small variations in the area and length parameters of the vocal tract to the change in formant frequency. The second perturbation relationship connects a change in the formant frequency with a change in the cepstral parameters [17]. By combining both perturbation relationships, we iteratively update the area and length parameters to decrease the spectral distance between the target speech spectrum and the vocal-tract spectrum calculated using the area and length parameters. Both perturbation relationships are linear and convenient to use in the parameter optimization scheme. However, they only hold for a local region in the parameter space.

The global estimation procedure requires an iterative optimization process starting from the given initial values of the vocal-tract parameters. In each iteration, the sensitivity function and Jacobian matrix representing the first and second perturbation relationships are recalculated. The updates of the formant frequencies and vocal-tract parameters are determined by minimizing an explicit cost function. We confirm the accuracy of our method experimentally on the morphological data for 10 English vowels [9].

This paper is organized as follows. Section 2 provides a description and mathematical explanation of our inverse estimation method. The numerical results are presented in Sect. 3. Section 4 provides the discussion and conclusions of our work.

2. ESTIMATION METHOD

In this section, we present our method for determining vocal-tract shape from a target speech spectrum. First, the cross-sectional area function of the vocal tract is represented as a sum of mode functions. The vocal-tract length is represented as a function of the first two mode coefficients. These mode coefficients are the unknown parameters to be determined in our method. We then determine the optimal values of these parameters by minimizing a cost function (i.e., the cepstral distance between the target speech spectrum and the cepstrum calculated from the current mode coefficients). This optimization is achieved by using two types of perturbation relationships: the relationship between the area function and the formant frequencies of the vocal tract and the relationship between the formant frequencies and the cepstral parameters.

2.1. Overview of our Estimation Framework

We use mode functions to represent the area parameters $A(i)$, thereby reducing the number of degrees of freedom of the vocal-tract parameters with no noticeable loss of accuracy [9].

$$A(i) = A_0(i) + \sum_{m=1}^{M} \gamma_m \phi_m(i) \quad (i = 1, 2, \ldots, N_\mathrm{a}) \tag{1}$$

where $A_0(i)$ is the average area, γ_m is the mode coefficient, and $\phi_m(i)$ is the mode function obtained through principal component analysis. N_a is the number of vocal-tract sections and M is the number of mode functions.

The section length of the vocal tract is represented by a function of the first and second mode coefficients (γ_1 and γ_2) [9].

$$L_S(\gamma_1, \gamma_2) = \sum_{p=1}^{P} w_p(\gamma_1, \gamma_2) L_p, \tag{2}$$

$$w_p(\gamma_1, \gamma_2) = \frac{d_p(\gamma_1, \gamma_2)}{\sum_{p=1}^{P} d_p(\gamma_1, \gamma_2)}, \tag{3}$$

and

$$d_p(\gamma_1, \gamma_2) = [\{\gamma_{1p} - \gamma_1\}^2 + \{\gamma_{2p} - \gamma_2\}^2]^{-1}, \tag{4}$$

where L_p is the section length and γ_{1p} and γ_{2p} are the mode coefficients for the vowel p. $L_S(\gamma_1, \gamma_2)$ forms a curved surface interpolating the given data for the vowels $p = 1, 2, \ldots, P$. This prevents a possible complementary relationship between the length and area of the vocal tract when determining its acoustic properties, thus resulting in an improved estimation accuracy as shown in our previous study [16]. As explained in Sect. 2.4, a set of vocal-tract area function data for 10 English vowels [9] is used in this study to calculate the mean area $A_0(i)$, the mode functions $\phi_m(i)$, and the prediction function L_S.

Because of the nonlinear relationship between vocal-tract shape and the speech spectrum, the vocal-tract parameters are determined using an iterative optimization procedure starting with initial values.

$$A^{k+1}(i) = A^k(i) + \Delta A^k(i) \tag{5}$$

and

$$L^{k+1} = L^k + \Delta L^k, \tag{6}$$

where k is the iteration index. The vocal-tract shape is reconstructed from the M mode coefficients, γ_m, using Eqs. (1) and (2). Then, the update quantities are

$$\Delta A^k(i) = \sum_{m=1}^{M} \Delta\gamma_m \phi_m(i) \tag{7}$$

and

$$\begin{aligned} \Delta L^k &= L_S(\gamma_1^{k+1}, \gamma_2^{k+1}) - L_S(\gamma_1^k, \gamma_2^k) \qquad (8) \\ &= L_S(\gamma_1^k + \Delta\gamma_1, \gamma_2^k + \Delta\gamma_2) - L_S(\gamma_1^k, \gamma_2^k), \end{aligned}$$

where $\Delta\gamma_m = \gamma_m^{k+1} - \gamma_m^k$ is the update quantity of each mode coefficient.

The target speech spectrum is represented using cepstral parameters $\hat{c}(j)$ for $j = 1, 2, \ldots, N_c$, where N_c is the number of cepstral parameters. The values of $\Delta\gamma_m$ are determined in each iteration by minimizing the following cost function:

$$d_c = \sum_{j=1}^{N_c} w_c(j)\{c^{k+1}(j) - \hat{c}(j)\}^2, \tag{9}$$

where $c^{k+1}(j)$ is the cepstral parameter of the vocal-tract transfer function calculated from the cross-sectional area, $A^{k+1}(i)$, and the section length, L^{k+1}. $w_c(j)$ is the weighting parameter for the jth cepstral parameter.

2.2. Determining the Update Quantities

To determine the value of $\Delta\gamma_m$, we use the formant frequency as the intermediate parameter between the vocal-tract shape and the speech spectrum. Based on a perturbation relationship between cepstral parameters and formants,

$$\boldsymbol{\Delta c} = J\boldsymbol{\Delta f}, \tag{10}$$

and the matrix form representation, the cost function in Eq. (9) is rewritten as follows:

$$\begin{aligned} d_c &= \{\boldsymbol{c}^{k+1} - \hat{\boldsymbol{c}}\}^t W_c \{\boldsymbol{c}^{k+1} - \hat{\boldsymbol{c}}\} \\ &= \{\boldsymbol{c}^k - \hat{\boldsymbol{c}} + \boldsymbol{\Delta c}\}^t W_c \{\boldsymbol{c}^k - \hat{\boldsymbol{c}} + \boldsymbol{\Delta c}\} \qquad (11) \\ &= \{\boldsymbol{e}_c + J\boldsymbol{\Delta f}\}^t W_c \{\boldsymbol{e}_c + J\boldsymbol{\Delta f}\}, \end{aligned}$$

where

$$\hat{\boldsymbol{c}} = \{\hat{c}(1), \hat{c}(2), \ldots, \hat{c}(N_c)\}^t, \tag{12}$$

$$\boldsymbol{c}^k = \{c^k(1), c^k(2), \ldots, c^k(N_c)\}^t, \tag{13}$$

$$W_c = \mathrm{diag}\{w_c(1), w_c(2), \ldots, w_c(N_c)\}, \tag{14}$$

and t represents matrix transposition.

$$\boldsymbol{\Delta c} = \{\Delta c(1), \Delta c(2), \ldots, \Delta c(N_c)\}^t \tag{15}$$

is the change in cepstral parameters caused by the parameter update.

$$\boldsymbol{e}_c = \boldsymbol{c}^k - \hat{\boldsymbol{c}} \tag{16}$$

is the cepstral error before updating, and

$$\boldsymbol{\Delta f} = \{\Delta f_1, \Delta f_2, \ldots, \Delta f_{N_f}\}^t \tag{17}$$

is the change in the formant frequencies. N_f denotes the number of formants.

J in Eq. (10) is the Jacobian matrix, and the (j, l) component of J can be given as [17]

$$J_{jl} = \frac{\partial c(j)}{\partial f_l} = -\frac{4\pi}{f_s}\exp\left(-\frac{\pi j}{f_s}b_l^k\right)\sin\left(\frac{2\pi j}{f_s}f_l^k\right), \tag{18}$$

where f_l^k and b_l^k are the frequency and band width of the lth formant determined from the vocal-tract parameters $A^k(i)$ and L^k. f_s is the sampling frequency. The value of $\boldsymbol{\Delta f}$ is then determined by setting $\partial d_c/\partial\boldsymbol{\Delta f} = 0$, which leads to

$$\boldsymbol{\Delta f} = -(J^t W_c J)^{-1} J^t W_c \boldsymbol{e}_c. \tag{19}$$

Next, we determine the value of $\Delta\gamma_m$ from the change in the formant frequency, Δf_l, based on the minimization of the following cost function:

$$d_f = \sum_{l=1}^{N_f}\left\{\frac{f_l^{k+1} - (f_l^k + \Delta f_l)}{f_l^k}\right\}^2. \tag{20}$$

In our previous study [15,16], we showed that f_l^{k+1} can be predicted as a function of $\Delta\gamma_m$ using

$$f_l^{k+1} = \left\{1 + \sum_{m=1}^{M} T_{lm}\Delta\gamma_m\right\} f_l^k, \tag{21}$$

where T_{lm} is a given coefficient that is explained in the next subsection. The cost function can then be rewritten as

$$d_f = \sum_{l=1}^{N_f}\left\{-z_l + \sum_{m=1}^{M} T_{lm}\Delta\gamma_m\right\}^2, \tag{22}$$

where $z_l = \Delta f_l / f_l^k$ is the change in formant frequency normalized by the pre-update formant frequency. By setting $\partial d_f/\partial\Delta\gamma_m = 0$, the optimal value of $\Delta\gamma_m$ is determined by simultaneously solving the following equations:

$$\sum_{m=1}^{M} T_{lm}\Delta\gamma_m = z_l \quad (l = 1, 2, \ldots, N_f). \tag{23}$$

Equation (23) can be expressed in matrix form as $T\Delta\boldsymbol{\gamma} = z$ and the value of $\Delta\gamma_m$ is determined as

$$\Delta\boldsymbol{\gamma} = T^{+}z, \tag{24}$$

where T is a coefficient matrix composed of T_{lm} and $+$ is the Moore–Penrose inverse matrix. Finally, the parameter values for the vocal tract are updated using

$$A^{k+1}(i) = A^{k}(i) + \sum_{m=1}^{M} \Delta\gamma_m \phi_m(i) \tag{25}$$

and

$$L^{k+1} = L_S(\gamma_1^k + \Delta\gamma_1, \gamma_2^k + \Delta\gamma_2). \tag{26}$$

2.3. Determining the Linear Coefficients T_{lm}

The Taylor expansion of L_S in Eq. (8) gives

$$\begin{aligned} &L_S(\gamma_1^{k+1}, \gamma_2^{k+1}) - L_S(\gamma_1^k, \gamma_2^k) \\ &\quad = \left(\Delta\gamma_1 \frac{\partial}{\partial\gamma_1} + \Delta\gamma_2 \frac{\partial}{\partial\gamma_2}\right) L_S(\gamma_1^k, \gamma_2^k), \end{aligned} \tag{27}$$

where the higher, nonlinear terms have been omitted. Equation (8) can then be rewritten as

$$\Delta L^k = \sum_{m=1}^{2} \Delta\gamma_m \tilde{\phi}_m \tag{28}$$

by setting $\tilde{\phi}_m = \frac{\partial}{\partial\gamma_m} L_S(\gamma_1^k, \gamma_2^k)$ for $m = 1$ and 2.

The change in formant frequencies for a small perturbation in the cross-sectional area was represented by a sensitivity function $\bar{S}(n,i)$ by Fant and Pauli [13]. Later, Adachi *et al.* derived a sensitivity function, $\tilde{S}(n,i)$, with respect to the vocal-tract length [14]. Using these sensitivity functions, the perturbation relationships can be written as

$$\frac{\Delta\bar{f}_l}{f_l} = \sum_{i=1}^{N_a} \bar{S}(l,i) \frac{\Delta A(i)}{A(i)} \tag{29}$$

and

$$\frac{\Delta\tilde{f}_l}{f_l} = \sum_{i=1}^{N_a} \tilde{S}(l,i) \frac{\Delta L(i)}{L(i)}, \tag{30}$$

where $\Delta\bar{f}_l/f_l$ and $\Delta\tilde{f}_l/f_l$ are the relative changes in the lth formant frequency, $\Delta A(i)/A(i)$ is the relative change in the cross-sectional area, and $\Delta L(i)/L(i)$ is the relative change in the section length of the vocal tract.

Substituting Eqs. (7) and (28) into the perturbation relationships in Eqs. (29) and (30) gives [15,16]:

$$\frac{\Delta\bar{f}_l}{f_l^k} = \sum_{i=1}^{N_a} \frac{\bar{S}(l,i)}{A^k(i)} \sum_{m=1}^{M} \Delta\gamma_m \phi_m(i) = \sum_{m=1}^{M} \bar{T}_{lm} \Delta\gamma_m \tag{31}$$

and

$$\frac{\Delta\tilde{f}_l}{f_l^k} = \sum_{i=1}^{N_a} \frac{\tilde{S}(l,i)}{L^k} \sum_{m=1}^{2} \Delta\gamma_m \tilde{\phi}_m = \sum_{m=1}^{2} \tilde{T}_{lm} \Delta\gamma_m, \tag{32}$$

where

$$\bar{T}_{lm} = \sum_{i=1}^{N_a} \frac{\bar{S}(l,i)}{A^k(i)} \phi_m(i) \tag{33}$$

and

$$\tilde{T}_{lm} = \sum_{i=1}^{N_a} \frac{\tilde{S}(l,i)}{L^k} \tilde{\phi}_m(i). \tag{34}$$

The values of $\bar{T}_{lm}$ and $\tilde{T}_{lm}$ are known since they are calculated from the vocal-tract parameters before updating.

The updated formant frequencies are obtained by summing the pre-update frequencies and the frequency changes caused by the update of the cross-sectional area and the sectional length:

$$f_l^{k+1} = f_l^k + \Delta\bar{f}_l + \Delta\tilde{f}_l = \left\{1 + \sum_{m=1}^{M} T_{lm} \Delta\gamma_m\right\} f_l^k \tag{35}$$

where $T_{lm} = \bar{T}_{lm} + \tilde{T}_{lm}$ and $\tilde{T}_{lm} = 0$ for $m > 2$. This equation is equivalent to Eq. (21).

2.4. Estimation Procedure

The estimation settings specify the number of cepstral parameters, N_c, the number of formants, N_f, and the number of mode functions, M. Figure 1 shows the estimation procedure steps:

(1) A set of vocal-tract area function data for 10 English vowels obtained from static MRI measurements (the subject was a male native speaker of English) [9] was used to calculate the mean area $A_0(i)$, the mode functions $\phi_m(i)$, and the prediction function L_S for the section length. The mode coefficients (γ_m) and the vocal-tract parameters ($A^0(i), L^0$) were initialized and the index was set to $k = 0$.

(2) The sensitivity functions ($\bar{S}(l,i), \tilde{S}(n,i)$), the frequency response of the vocal tract ($H(\omega)$), the formant frequencies and band widths (f_l^k, b_l^k), and the cepstral parameters ($c^k(j)$) were calculated from the vocal-tract parameters ($A^k(i), L^k$), using an acoustic model [2].

(3) The cepstral error was calculated for the target spectrum. If a cepstral distance value ($\sqrt{d_c/\sum_j w_c(j)}$) was below a threshold, $A^k(i)$ and L^k were used as the estimation results and the procedure was terminated.

(4) The formant-change Jacobian matrix was calculated from the formant frequencies and band widths (Eq. (18)). The formant change was then calculated from the cepstral error and the Jacobian matrix (Eq. (19)). In addition, the update quantity of the mode coefficients ($\Delta\gamma_m$) was calculated from the relative formant change and linear coefficients (Eq. (24)). The method for calculating the linear coefficients is described in Sect. 2.3.

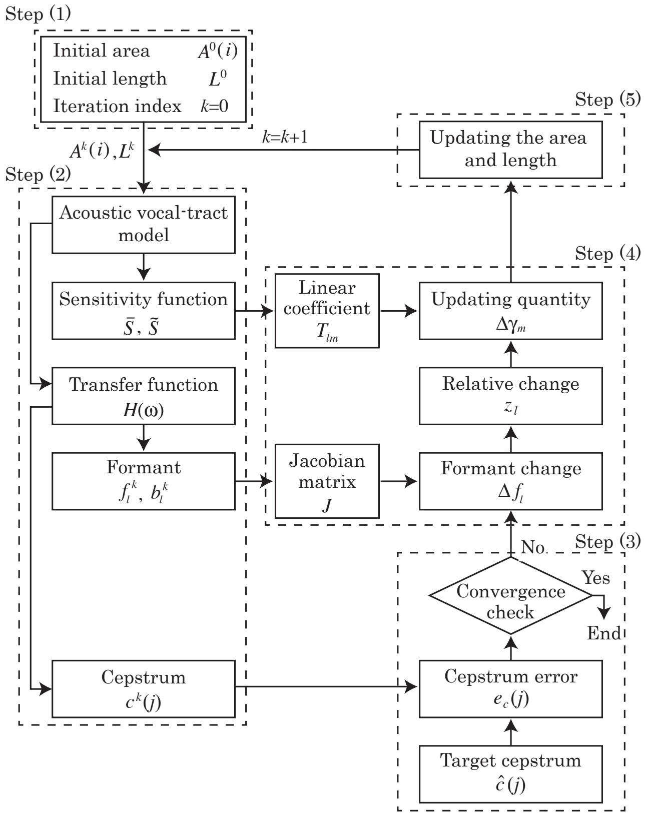

Fig. 1 Procedure for estimating the vocal-tract shape from the target speech spectrum.

(5) The vocal-tract parameters were updated to $A^{k+1}(i)$ and L^{k+1} from $\Delta\gamma_m$ (Eqs. (25) and (26)). The index was set to $k = k + 1$. The procedure was repeated from Step 2.

In the first step of the procedure, the initial values of the parameters were determined as follows. Candidates of the initial vocal-tract shape were calculated from every combination of seven values for the first mode coefficient ($\gamma_1 = -7.5, -5.0, -2.5, 0, 2.5, 5.0, 7.5$) and three values for the second mode coefficient ($\gamma_2 = -2.5, 0, 2.5$). The other mode coefficients were set to zero. For each of the 21 combinations, the vocal-tract shape was reconstructed using Eqs. (1) and (2). The frequency response of the vocal tract and the cepstral parameters were also calculated. Next, the optimal candidate among the 21 vocal-tract shapes (i.e., the reconstruction corresponding to the minimum cepstral distance from the target spectrum) determined $A^0(i)$ and L^0.

In the second step, the frequency-domain acoustic tube model [2] is based on wave propagation in a lossy vocal tract. The model includes the effects of vocal-tract wall vibration, viscous friction loss, and heat conduction loss on the surface of the vocal-tract wall. The radiation impedance at the lips was set according to [22], and we assume that the glottis is closed. In the third step of the procedure, the convergence of the iterative procedure was judged using a threshold value of 0.08 (determined empirically).

Note that the vocal-tract spectrum is represented using the complex cepstrum throughout the method including the spectral distance in Eq. (9) and the perturbation relationship in Eq. (10). To calculate the value of $c^k(j)$ in the second step, the ordinal real cepstrum was first obtained as the inverse Fourier transform of the logarithm of the magnitude response of the vocal tract, $\ln|H(\omega)|$, then converted to the complex cepstrum with the minimum phase property [23]. For this study, $H(\omega)$ was computed over the frequency range 0 to 5000 Hz at 2 Hz intervals. The target spectrum, $\hat{c}(j)$, was conveniently obtained from the speech through a linear predictive analysis by converting the linear predictive parameters to complex cepstrum parameters using a recursive formula [24].

3. ESTIMATION RESULTS

3.1. Experimental Conditions

The estimation accuracy was quantitatively examined using area function data for 10 English vowels collected from a male speaker using magnetic resonance imaging [9]. First, speech signals were generated from the area function data using a two-mass model of the vocal folds [5] and an acoustic tube model of the vocal tract [2]. The model [5] uses a linear spring and damper to represent the viscoelastic property of the vocal fold. Furthermore, the separation of glottal flow from the surface of the vocal fold is considered, allowing accurate estimation of the volume flow rate and pressure distribution along the vocal fold.

The parameters of the synthetic model were set as follows. We used the same mechanical constants for the upper and lower masses of each vocal fold. The mass was $m = 0.1$ g, the spring constant was $k = m(2\pi \times 100)^2$ dyn/cm, the spring constant connecting both masses was $0.6k$ dyn/cm, and the damper parameter was set to $0.2\sqrt{km}$ dyn·s/cm. During contact between the left and right vocal folds, the spring constant was increased to $4k$ dyn/cm and the damper parameter was set to $2.2\sqrt{km}$ dyn·s/cm. The length of the vocal fold was 1.4 cm, the subglottal pressure was 8 cmH$_2$O, and the separation constant of the glottal flow was 1.2. We used a sampling frequency of 10 kHz. The length of each speech signal was 200 ms.

Next, we analyzed the synthesized speech using 16th order linear predictive analysis. The whole samples of each speech signal were preprocessed using a hamming window having 200-ms length and a single-order FIR high-pass filer before the analysis. The linear predictive parameters were then converted to cepstral parameters [24] and used as the target spectrum. Finally, the vocal-tract parameters (cross-sectional area and length) were estimated from the target spectrum and the result was compared with the actual area

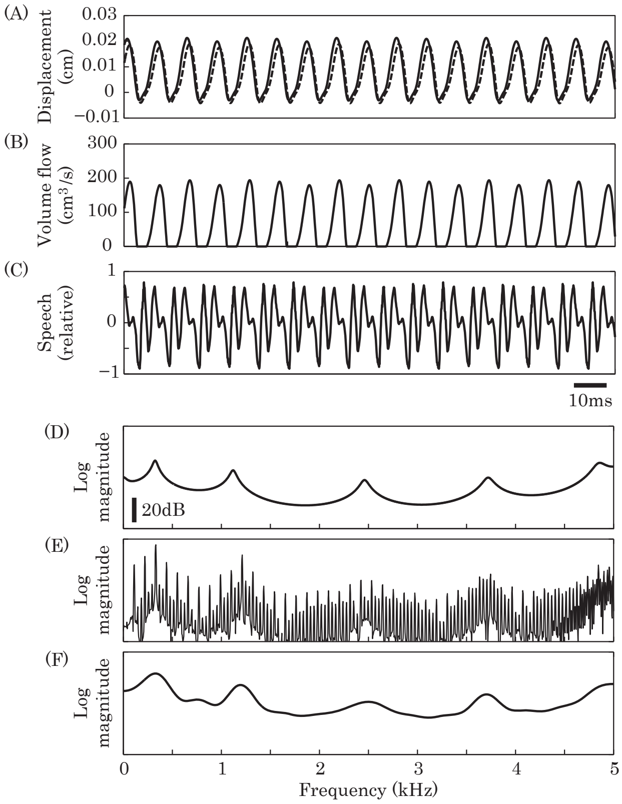

Fig. 2 Speech production simulation result for the vowel /u/. The waveforms show the temporal patterns of (A) the displacement corresponding to the lower mass (solid line) and upper mass (dashed line) of the vocal fold, (B) the glottal volume flow rate, and (C) the synthesized speech. The spectra show (D) the acoustic characteristics of the vocal tract calculated from the area function data, (E) the spectrum of the synthesized speech, and (F) the spectral envelope of the synthesized speech (used as the target spectrum for the inverse estimation) obtained through cepstral analysis.

function data. The number of vocal-tract sections was $N_a = 44$. The number of cepstral parameters was $N_c = 30$ and the weight included in the cepstral distance (d_c) was set to $w_c(j) = j^{0.4}$. We set the number of formants to four ($N_f = 4$) and the number of mode functions to seven ($M = 7$) according to published experimental results [15].

Figure 2 shows the result of the speech production simulation and cepstral analysis for the vowel /u/. The waveforms for the vocal-fold displacement, glottal volume flow, and synthesized speech are taken from a stable portion of the simulation result. Characteristic features of human speech production—such as the phase difference between the upper and lower portions of the vocal fold during self-oscillation and the tilt of the glottal flow waveform—are observed in the figure. The spectrum labeled (F) is the cepstrum-based spectral envelope calculated from the synthesized speech. It was used as the target for estimating the vocal-tract parameters. The figure shows that the spectrum in (F) adequately approximates the vocal-tract spectrum in (D) calculated using the area function data with the acoustic tube model [2]. Sharp peaks in (D) are rounded in (F) because of the fundamental properties of cepstral analysis.

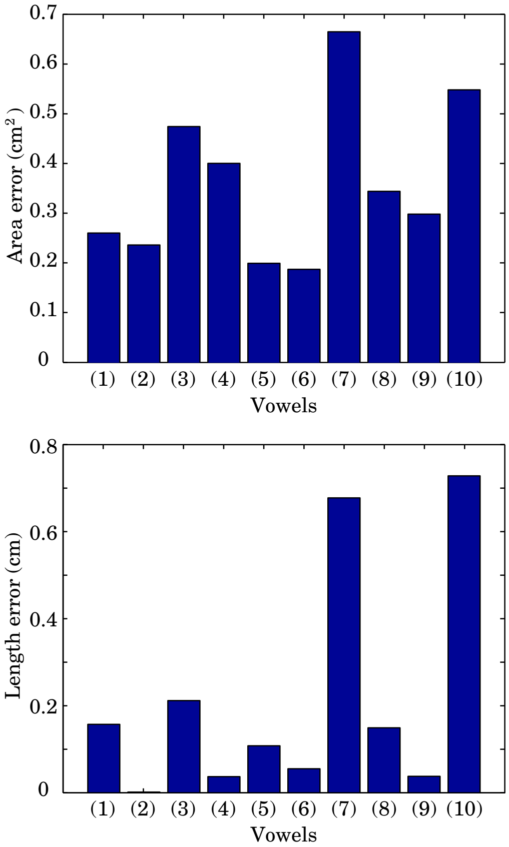

Fig. 3 Estimation error of the vocal-tract cross-sectional area (top) and length (bottom) as a function of vowels (1) /i/, (2) /ɪ/, (3) /ɛ/, (4) /æ/, (5) /ʌ/, (6) /ɑ/, (7) /ɔ/, (8) /o/, (9) /ʊ/, and (10) /u/. The area error was calculated by averaging the errors for all vocal-tract sections for each vowel. The vocal-tract length was calculated by multiplying the section length by the number of vocal-tract sections.

3.2. Morphological and Acoustic Evaluation of the Estimation Accuracy

The upper plot of Fig. 3 shows the cross-sectional area estimation error calculated by $\frac{1}{N_a}\sum_{i=1}^{N_a}|A_e(i) - A_c(i)|$, where $A_e(i)$ and $A_c(i)$ represent the estimated and actual area values, respectively. The area error varies depending on the type of vowel. The minimum and maximum area error values were $0.18\,\mathrm{cm}^2$ ((6) /ɑ/) and $0.66\,\mathrm{cm}^2$ ((7) /ɔ/), respectively. The mean area error across all of the vowels was $0.36\,\mathrm{cm}^2$. The average actual cross-sectional area was about $2.05\,\mathrm{cm}^2$, thus, the estimation error was approximately 17.6% of the actual area. As shown in the lower part of Fig. 3, the estimation results for the vocal-tract length produced low errors with the exception of (7) /ɔ/

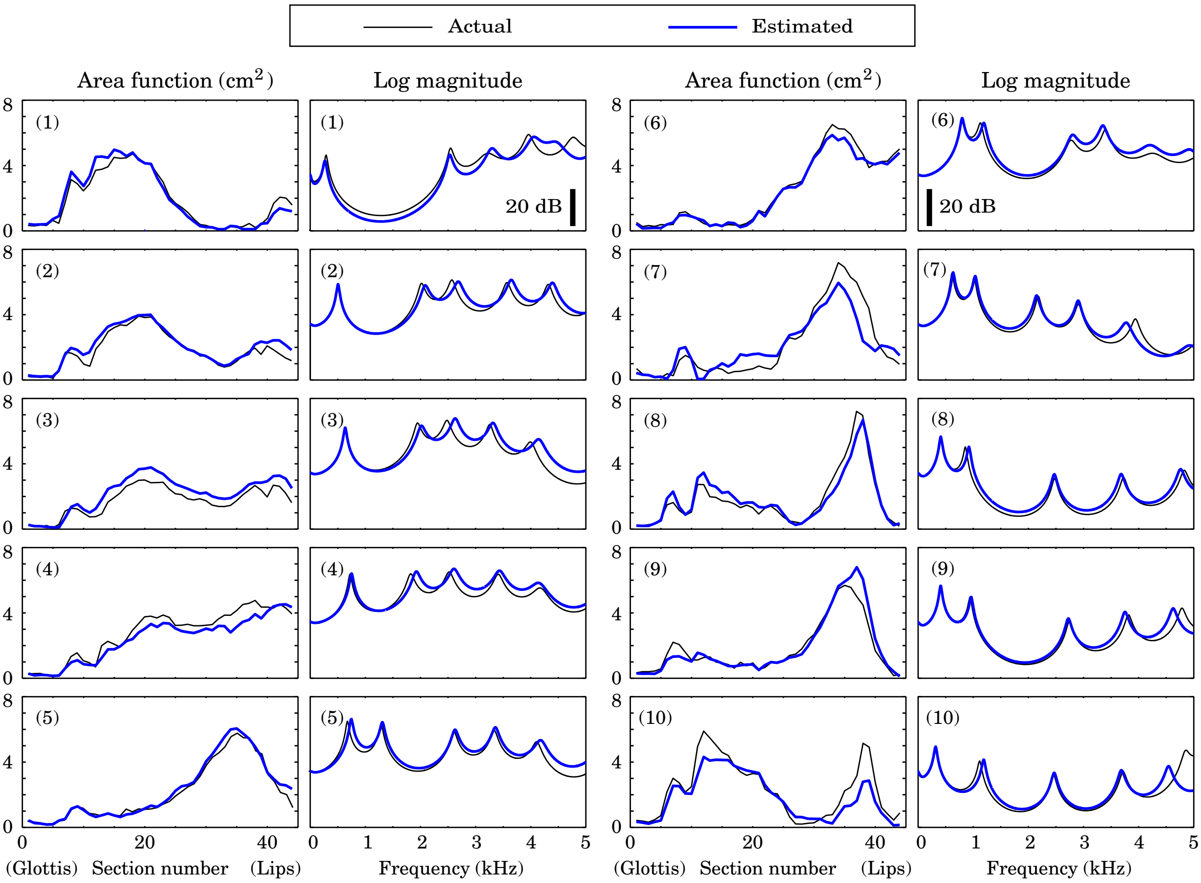

Fig. 4 Estimated area functions and frequency responses for (1) /i/, (2) /ɪ/, (3) /ɛ/, (4) /æ/, (5) /ʌ/, (6) /ɑ/, (7) /ɔ/, (8) /o/, (9) /ʊ/, and (10) /u/. The thin lines show the data based on MRI measurements (Story and Titze, 1998) from which the target speech spectra were generated. The thick lines show the estimation results.

and (10) /u/. The mean length error across all of the vowels was 0.21 cm, which corresponds to 1.24% of the actual mean length of 16.9 cm.

Regarding the large estimation error of /ɔ/ and /u/, a similar result was obtained when the vocal-tract shape was estimated from the formant frequencies [16]. In this previous study, it was shown that multiple vocal-tract shapes produce the same frequencies for the first four formants. This result was interpreted as a typical example of a one-to-many relationship between the acoustic and vocal-tract parameters. This problem of non-uniqueness can substantially degrade the estimation accuracy of the vocal-tract shape, as found for specific vowels such as /ɔ/ and /u/.

Figure 4 shows the area function estimated from the target speech spectrum and frequency response of the vocal tract. The estimated area function plotted on the left (thick lines) corresponds closely to the actual values (thin lines). The estimates for (5) /ʌ/ and (6) /ɑ/ were particularly close while (3) /ɛ/, (4) /æ/, (7) /ɔ/, and (10) /u/ produced noticeable errors. The transfer function shown on the right was calculated using the actual and estimated area function. The figure shows that the estimated spectrum was relatively accurate up to the frequency of the fourth formant.

The frequency for the first formant was accurately recovered for each vowel from the estimated vocal-tract shape, as indicated in the figure. The average absolute frequency error was 15.2 Hz for the first formant and 52.2 Hz for the second formant. The relative error obtained by normalizing the frequency error by the actual formant frequency was 2.93% for the first formant and 3.95% for the second formant. The average relative error for the third and fourth formants was 2.00% and 1.26%, respectively. These results indicate that the characteristic formant frequencies were satisfactorily recovered for each vowel in the estimated frequency response of the vocal tract.

Because formant frequencies are hidden, intermediate parameters, it is useful to examine estimation accuracy when the vocal-tract shape has been directly estimated from formants (as in our previous method [16]). The resulting area and length errors were, on average, 0.33 cm^2 and 0.19 cm, respectively, for the same English vowels when using four formant frequencies as the target. The initial vocal-tract shape was selected from 21 candidates, as in the present method. It is clear that the increase in

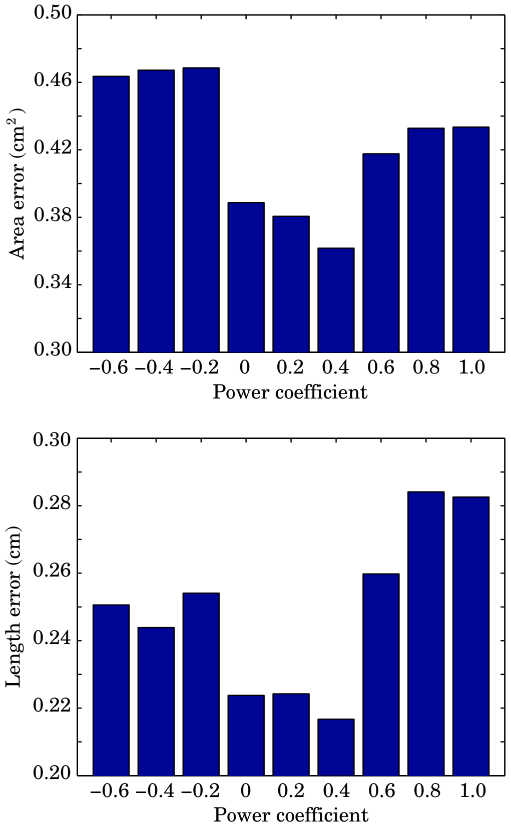

Fig. 5 Mean estimation error of the vocal-tract cross-sectional area (top) and length (bottom) as a function of the power coefficient.

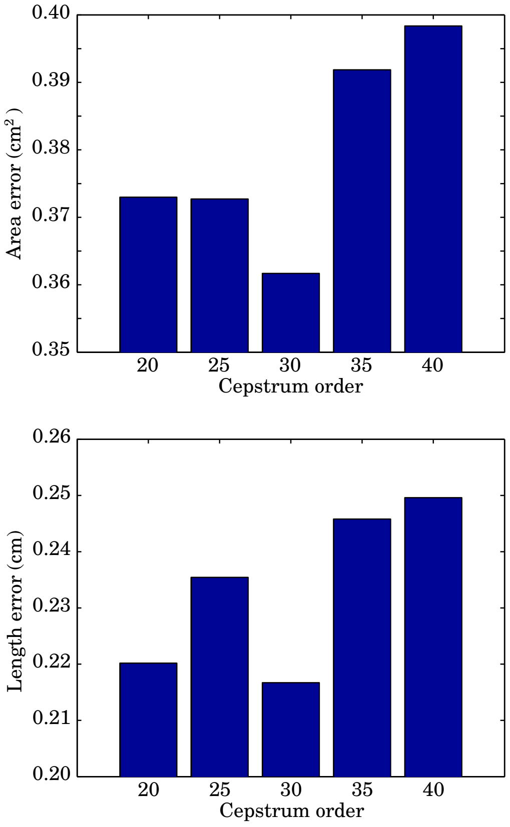

Fig. 6 Mean estimation error of the vocal-tract cross-sectional area (top) and length (bottom) as a function of the cepstrum order (N_c).

estimation error, i.e., 0.03 cm^2 for area and 0.02 cm for vocal-tract length, was very small when using the speech spectrum as the target in the present method. It is also valuable to compare the estimation results of our method with those of the linear mapping method [20]. The estimation results reported in the literature [20] show that the area error is almost the same (0.367 cm^2) and the length error is slightly better (0.150 cm) than those for our method. However, it should be noted that the dataset used in the linear mapping method and the number of principal components differed from those in our study.

3.3. Influence of the Parameters of the Cost Function

Next, we examined the estimation error with respect to the parameters of the cost function. First, the cepstrum order was fixed to $N_c = 30$, and the power coefficient, α, of the cost weight ($w_c(j) = j^\alpha$) was varied from −0.6 to 1.0. The mean estimation error for the area (top) and length (bottom) for 10 vowels is plotted in Fig. 5. We obtained a minimal error with a power coefficient of 0.4 for both the area and length parameters. We then varied the cepstrum order from 20 to 40 while the power coefficient was fixed at 0.4. The results plotted in Fig. 6 show a minimum estimation error with a cepstrum order of 30. The experiments described in Sect. 3.2 (Figs. 3 and 4) were performed using these optimal cost function parameter values, i.e., a cepstrum order of 30 and a power coefficient of 0.4.

Note that the target spectrum $\hat{c}(j)$ in Eq. (9) calculated from the speech signal may include the frequency characteristics of the voice source and the lip opening acoustic radiation in addition to the frequency response of the vocal tract. However, the cepstral parameters denoted by $c^k(j)$ are calculated from the frequency response of the vocal tract only. Even when speech signals are processed by a pre-emphasis filter prior to the cepstral analysis, it is difficult to completely remove the frequency components resulting from the voice source and acoustic radiation.

Therefore, these other frequency components included in the target spectrum can degrade the estimation accuracy, and the cost function in Eq. (9) should be robust against this problem. Figure 5 indicates an optimal power coefficient of 0.4 and suggests that the cepstral error should be weighted strongly when the index of cepstral parameters, j, increases. This cost weighting is somewhat consistent with the root-power sums weighted distance measure used for speech recognition techniques [25]. The root-power sums technique emphasizes spectral peaks when comparing a pair of speech spectra. It is expected that our cost function

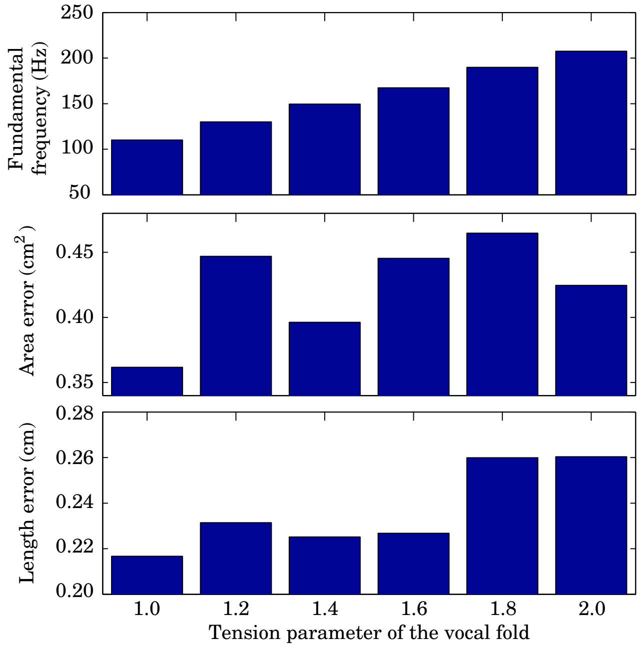

Fig. 7 The fundamental frequency of the target speech spectrum averaged over 10 vowels (top), the mean vocal-tract cross-sectional area estimation error (middle), and the mean vocal-tract length estimation error (bottom) as functions of the tension parameter.

has a similar effect in heavily weighting spectral peaks when comparing the target speech spectrum and the vocal-tract spectrum. As a result, our estimation method is accurate despite the potential presence of the voice source and radiation characteristics in the target speech spectrum.

3.4. Influence of the Fundamental Frequency of the Target Speech Spectrum

The target spectrum in our experiment is obtained through the cepstral analysis of speech signals. Speech analysis is generally influenced by the fundamental frequency. Therefore, we performed an additional experiment to evaluate this influence by changing the fundamental frequency of the target speech spectrum. In the production simulation of the target speech spectrum, the fundamental frequency was altered by dividing the mass parameters of the vocal fold model by a tension parameter, Q, and by multiplying the spring constants by Q. The natural frequency of the mass-spring system of the model was thus changed in proportion to the tension parameter. The subglottal pressure was 8 cmH_2O when Q was between 1.0 and 1.6, and it was increased to 10 cmH_2O to maintain the self-sustained oscillation of the vocal folds when Q was 1.8 and 2.0.

Figure 7 shows the fundamental frequency of the target speech spectrum, the area error, and the length error averaged for 10 vowels. The fundamental frequency increased from 110.1 Hz to 207.5 Hz as Q increased from 1.0 to 2.0. The area error reached a maximum of 0.46 cm^2 when Q was 1.8. The minimum error was 0.36 cm^2 when Q was 1.0, so the maximum increase in error, $(0.46 - 0.36)/0.36$, was about 28%. The length error increased monotonically as the tension parameter increased. The minimum error was 0.21 cm when Q was 1.0 and reached a maximum of 0.26 cm when Q was 2.0. Thus, the maximum increase in error, $(0.26 - 0.21)/0.21$, was about 24%.

These results confirm that the vocal-tract shape estimation accuracy is influenced by the fundamental frequency. When the target spectrum is calculated from speech, an increased fundamental frequency causes sparse frequency components, hindering the extraction of the vocal-tract acoustic characteristics. The target speech spectrum was likely erroneous (i.e., did not accurately represent the vocal-tract acoustic characteristics) for high-pitched speech signals, resulting in an increased vocal-tract shape estimation error.

4. CONCLUSIONS

We presented a method to accurately estimate vocal-tract shape (cross-sectional area and length) from a speech spectrum. The target spectrum was represented using cepstral parameters and the cross-sectional area and length of the vocal tract were determined simultaneously by minimizing cepstral distance using an iterative optimization procedure. In each iteration, a perturbation relationship between the formants and cepstral parameters was used to update the formant frequencies according to the cepstral error. In addition, a perturbation relationship between the vocal-tract shape and the formants (i.e., the acoustic sensitivity function) was used to update the area and length parameters. In contrast to our previous publications [15,16], the present estimation of the vocal-tract shape does not require the extraction of formant frequencies from speech signals.

We conducted numerical analyses to evaluate our estimation method. The resulting estimation error was 0.36 cm^2 for the vocal-tract cross-sectional area and 0.21 cm for the vocal-tract length. This corresponded to approximate errors of 17.6% of the true cross-sectional area and 1.24% of the true length. We also showed that the increase of the estimation error was very small compared with our previous method [16], for which formant frequencies were used as the acoustic parameters. The frequency error was less than 4% for each of the first four formants recovered from the estimated vocal-tract shape. From these morphological and acoustic evaluations, we conclude that our method estimates vocal-tract shape with a satisfactory degree of accuracy. When speech signals are regenerated from the estimated vocal-tract shapes using a model of human speech production, we expect a low frequency error for the lower formants leading to intelligible synthesized speech. Therefore, future work will include conducting a study to construct an analysis-synthesis framework of speech. In this framework, the control parameters of the speech production model are extracted from real speech and speech re-synthesis is performed based on the human speech production mechanism.

Our experiments also showed that the estimation accuracy is influenced by the fundamental frequency of the target speech. The frequency range used in the experiment (110 to 208 Hz) covers almost the standard adult male vocal range. The maximum error increase was 28% for the cross-sectional area and 24% for the length relative to the minimum respective errors obtained with a fundamental frequency of 110 Hz. These results indicate that the fundamental frequency requires consideration when the estimation method is used to analyze real speech. The target spectrum was represented using cepstral parameters which were obtained through a linear predictive analysis. Therefore, using a pitch-robust speech analysis method such as discrete all-pole modeling [26] or the weighted linear prediction method [27] could improve the estimation accuracy for high-pitched speech signals.

ACKNOWLEDGEMENT

This research was partly supported by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (Grant No. 23300071).

REFERENCES

[1] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.*, **51**, 1233–1268 (1972).

[2] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust. Speech Signal Process.*, **35**, 955–967 (1987).

[3] X. Pelorson, A. Hirschberg, R. R. van Hassel, A. P. J. Wijnands and Y. Auregan, "Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model," *J. Acoust. Soc. Am.*, **96**, 3416–3431 (1994).

[4] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," *J. Acoust. Soc. Am.*, **97**, 1249–1260 (1995).

[5] N. J. C. Lous, G. C. J. Hofmans, R. N. J. Veldhuis and A. Hirschberg, "A symmetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prosthesis design," *Acustica united with Acta Acustica*, **84**, 1135–1150 (1998).

[6] I. T. Tokuda, J. Horáček, J. G. Švec and H. Herzel, "Comparison of biomechanical modeling of register transitions and voice instabilities with excised larynx experiments," *J. Acoust. Soc. Am.*, **122**, 519–531 (2007).

[7] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *J. Acoust. Soc. Am.*, **123**, 2733–2749 (2008).

[8] T. Kaburagi, "Voice production model integrating boundary-layer analysis of glottal flow and source-filter coupling," *J. Acoust. Soc. Am.*, **129**, 1554–1567 (2011).

[9] B. H. Story and I. R. Titze, "Parameterization of vocal tract area functions by empirical orthogonal modes," *J. Phon.*, **26**, 223–260 (1998).

[10] M. R. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," *J. Acoust. Soc. Am.*, **41**, 1002–1010 (1967).

[11] P. Mermelstein, "Determination of the vocal-tract shape from measured formant frequencies," *J. Acoust. Soc. Am.*, **41**, 1283–1294 (1967).

[12] B. H. Story, "Technique for tuning vocal tract area functions based on acoustic sensitivity functions," *J. Acoust. Soc. Am.*, **119**, 715–718 (2006).

[13] G. Fant and S. Pauli, "Spatial characteristics of vocal tract resonance modes," *Proc. Speech Commun. Semin. Stockholm*, pp. 121–132 (1974).

[14] S. Adachi, H. Takemoto, T. Kitamura, P. Mokhtari and K. Honda, "Vocal tract length perturbation and its application to male-female vocal tract shape conversion," *J. Acoust. Soc. Am.*, **121**, 3874–3885 (2007).

[15] T. Kaburagi, T. Takano and Y. Sakamoto, "Estimating area function of the vocal tract from formants using a sensitivity function and least-squares," *Acoust. Sci. & Tech.*, **34**, 301–310 (2013).

[16] T. Kaburagi, "Determining the length and cross-sectional area of the vocal tract jointly from formants using acoustic sensitivity function," *Acoust. Sci. & Tech.*, **35**, 290–299 (2014).

[17] D. D. Mehta, D. Rudoy and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *J. Acoust. Soc. Am.*, **132**, 1732–1746 (2012).

[18] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. (Marcel Dekker, New York, 1992).

[19] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin and E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *J. Acoust. Soc. Am.*, **100**, 1819–1834 (1996).

[20] P. Mokhtari, T. Kitamura, H. Takemoto and K. Honda, "Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients," *J. Phon.*, **35**, 20–39 (2007).

[21] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *J. Acoust. Soc. Am.*, **92**, 688–700 (1992).

[22] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. (Springer Verlag, New York, 1972), pp. 36–38.

[23] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, Upper Saddle River, N.J., 1978), pp. 360–362.

[24] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, Upper Saddle River, N.J., 1978), p. 442.

[25] K. K. Paliwal, "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition," *Speech Commun.*, **1**, 151–154 (1982).

[26] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Process.*, **39**, 411–423 (1991).

[27] P. Alku, J. Pohjalainen, M. Vainio, A. Laukkanen and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *J. Acoust. Soc. Am.*, **134**, 1295–1313 (2013).