Determining the length and cross-sectional area of the vocal tract jointly from formants using acoustic sensitivity function

Kaburagi, Tokihiko
Faculty of Design, Kyushu University

出版情報：Acoustical Science and Technology. 35 (6), pp.290-299, 2014. 日本音響学会
バージョン：

Acoust. Sci. & Tech. **35**, 6 (2014)

PAPER

Determining the length and cross-sectional area of the vocal tract jointly from formants using acoustic sensitivity function

Tokihiko Kaburagi*

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

(*Received 6 January 2014, Accepted for publication 14 May 2014*)

Abstract: A method for determining the length and cross-sectional area of the vocal tract from target formants is presented. The area function was approximated by summing several mode functions to reduce the number of degrees of freedom among the area parameters. Furthermore, the tract length was represented as a function of the coefficients for two principal modes. The estimation was made based on the perturbation relationship, i.e., a sensitivity function was used to represent the change in formant frequency due to a small perturbation of the vocal-tract shape. Starting from initial values, the vocal-tract parameters were optimized iteratively, and the sensitivity functions were used as linear constraints to update the parameter values. The estimation accuracy was examined using area function data for 10 English vowels (Story and Titze, *J. Phon.*, **26**, 223–260, 1998). The results showed that the method is capable of determining vocal-tract shape with a satisfactory degree of accuracy, though the estimation accuracy strongly depends on the type of vowel. The dependency of the estimation error on the initial values of the parameters was also investigated.

Keywords: Vocal-tract area function, Vocal-tract length, Formant, Sensitivity function, Inverse estimation

PACS number: 43.70.Bk, 43.72.Ct [doi:10.1250/ast.35.290]

1. INTRODUCTION

The inverse problem of speech production has attracted considerable attention [1,2] as part of efforts to establish an effective method for determining the condition of the vocal tract during speech without the use of such observational tools as X-ray imaging and magnetic resonance imaging (MRI). An approach to the speech inversion problem is converting speech acoustics to an area function of the vocal tract. With this approach, analytic methods have been investigated, based on the physical relationship between the area function and transfer function of the vocal tract. Schroeder [3] and Mermelstein [4] proposed the perturbation theory based on the wave equation, where for simplicity the vocal-tract wall was assumed to be hard and lossless. They represented the vocal-tract shape by Fourier expansion of the logarithm of the area function and showed how the Fourier coefficient was related to the poles and zeros of the vocal-tract impedance at the lips. Accurate determination of the input impedance is in fact a difficult task and requires acoustic input-output measurement. Schroeder and Mermelstein therefore assumed the vocal-tract shape to be asymmetrical around the center point or be represented by low-order Fourier components, and they estimated the cross-sectional area using formants. Sondhi and Resnick [5] later presented an inverse problem of the sound-wave propagation in the vocal tract for both lossless and lossy cases by means of acoustic measurement of the vocal tract. Mokhtari *et al.* [6] determined the area function from speech acoustics based on a statistical mapping.

In line with perturbation theory, the vocal-tract area function has been estimated from formants by using the sensitivity function [7,8]. The derivation of the sensitivity function is based on a theorem by Ehrenfest [9], and it represents the change in formant frequency due to a small perturbation in the vocal-tract shape. The sensitivity function for the cross-sectional area was studied by Fant and Pauli [10]. Fant [11] and Adachi *et al.* [8] subsequently demonstrated the sensitivity function with respect to vocal-tract length. Story [7] and Adachi *et al.* [8] showed that the initial vocal-tract area function can be tuned iteratively so that each formant frequency calculated from the area function approaches the target frequency, where the sensitivity function was used as the basis for changing

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

the area function. Their methods do not require the asymmetrical assumption imposed on the vocal-tract shape in the perturbation theory. Kaburagi *et al.* [12] showed that the convergence speed of Story's optimization method improves using an explicit criterion for the formant error and the least-squares technique.

In the present study, we expand upon the works mentioned above [7,8,12]; we propose a method for determining simultaneously the length and cross-sectional area of the vocal tract from target formants. The area function is usually composed of 20 to 40 sections; however, only three or four formants are typically found for an adult male speaker, when considering the frequency range up to 4 kHz. This indicates that the available number of formants is very limited compared to the number of vocal-tract parameters. With respect to degrees of freedom of vocal-tract parameters, area parameters for total sections of the vocal tract are well represented by the sum of a number of mode functions without noticeable loss of accuracy [13]. In addition, the length of the vocal tract is represented as a function of the coefficients for the first and second modes [13]. This length prediction function forms a curved surface over the two mode coefficients, so that length data for the vocal tract of several vowels are interpolated.

Starting from initial values for the area and length parameters of the vocal tract, the unknown parameters, i.e., the coefficients of mode functions, are optimized iteratively to obtain the target formant frequencies. In the present study, we use the perturbation relation, i.e., the sensitivity functions for the cross-sectional area [10] and vocal-tract length [8,11]. We confirm the accuracy of our method using morphological data for 10 English vowels obtained by MRI measurements [13]. We also examine the dependency of the estimation accuracy on the vowel type and initial values of the parameters.

In terms of the mode representation, Mokhtari *et al.* [6] and Story [14] treated the sectional length as a component of the data vector in addition to the cross-sectional areas of all vocal-tract sections; then, they performed the principal component analysis (PCA). In the speech inversion study performed by Mokhtari *et al.* [6], the vocal-tract length was estimated by predicting the values of the first two components from formants by linear regression. In our method, the PCA was performed only for the cross-sectional areas, because the effect of the cross-sectional area and the effect of the vocal-tract length on formant frequencies are independently considered.

This paper is organized as follows. Section 2 provides a mathematical explanation of our inverse estimating method, including a brief clarification of the sensitivity function. The estimation procedure is also provided. Numerical results are presented in Sect. 3. Finally, the conclusions of this work are given in Sect. 4.

2. ESTIMATION METHOD

In this section, we present a method for determining the length and cross-sectional area of the vocal tract from target formant frequencies. First, the area function is represented as a sum of mode functions. The vocal-tract length is also represented as a function of two mode coefficients. These indicate that the mode coefficients are the unknown parameters to be determined in our estimation method. Thereafter, two types of the sensitivity function—one is related to a perturbation of the cross-sectional area and the other to a perturbation of the vocal-tract length—are explained briefly. Finally, the estimation method is described in subsequent subsections. The values of the mode coefficients are determined by least squares; in this process, the sensitivity function is used as a constraint to relate the change in area and length to the change in formant frequency.

2.1. Mode Decomposition of the Area Function and Prediction of the Vocal-tract Length

To reduce the degrees of freedom of the parameters, vocal-tract areas for the total sections are represented using mode functions as [6,13]

$$A(i) = \overline{A(i)} + \sum_{m=1}^{M} \gamma_m \phi_m(i) \quad (i = 1, 2, \ldots, N_A) \qquad (1)$$

where $A(i)$ is the cross-sectional area for the ith vocal-tract section, $\overline{A(i)}$ is the average area taken over different phonemes in the data set, $\phi_m(i)$ is the mode function obtained through principal component analysis, γ_m is the mode coefficient, M is the number of mode functions used to reconstruct the areas, and N_A is the number of vocal-tract sections.

In addition, we suppose that the vocal-tract length is related to those areas and that the sectional length, $L(i)$, is identical for every vocal-tract section. Following the interpolation method of Story and Titze [13], the sectional length is determined by a function of the first and second mode coefficients:

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

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

and

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

where L_p is the sectional length and γ_{1p} and γ_{2p} are mode coefficients for the vowel p. $L_S(\gamma_1, \gamma_2)$ forms a curved surface over the γ_1–γ_2 plane so that given sectional lengths for vowels $p = 1, 2, \ldots, P$ are interpolated.

Based on this mode decomposition of the area parameters and the prediction of the sectional length, all the vocal-tract parameters (cross-sectional areas and sectional length) are represented by M mode coefficients, γ_m. In our method, γ_m is the unknown parameter to be determined from target formants.

2.2. Sensitivity Function between Formants and Vocal-tract Parameters

The change in formant frequencies for a small perturbation in the cross-sectional area was first represented as a sensitivity function, $S(n,i)$, by Fant and Pauli [10]. Later, Fant [11] and Adachi *et al.* [8] derived a sensitivity function, $\hat{S}(n,i)$, with respect to the vocal-tract length. With sensitivity functions, the perturbation relation is written as

$$\frac{\Delta F_n}{F_n} = \sum_{i=1}^{N_A} S(n,i)\frac{\Delta A(i)}{A(i)} \tag{5}$$

and

$$\frac{\Delta \hat{F}_n}{F_n} = \sum_{i=1}^{N_A} \hat{S}(n,i)\frac{\Delta L(i)}{L(i)}, \tag{6}$$

where $\Delta F_n/F_n$ and $\Delta \hat{F}_n/F_n$ are the relative change in the nth formant frequency, $\Delta A(i)/A(i)$ is the relative change in the cross-sectional area for the ith section, and $\Delta L(i)/L(i)$ is the relative change in the sectional length of the vocal tract for the ith section.

$S(n,i)$ and $\hat{S}(n,i)$ can be calculated via

$$S(n,i) = \frac{E_K(n,i) - E_P(n,i)}{E_T(n)} \tag{7}$$

and

$$\hat{S}(n,i) = -\frac{E_K(n,i) + E_P(n,i)}{E_T(n)}, \tag{8}$$

where $E_K(n,i)$ and $E_P(n,i)$ are kinetic and potential energies defined as

$$E_K(n,i) = \frac{1}{2}\frac{\rho L(i)}{A(i)}|U(n,i)|^2 \tag{9}$$

and

$$E_P(n,i) = \frac{1}{2}\frac{A(i)L(i)}{\rho c^2}|P(n,i)|^2. \tag{10}$$

$|U(n,i)|$ and $|P(n,i)|$ are the volume velocity and pressure amplitudes at the frequency of the nth formant, ρ is the air density, and c is the speed of sound. $E_T(n)$ is the total energy given by

$$E_T(n) = \sum_{i=1}^{N_A}\{E_P(n,i) + E_K(n,i)\}. \tag{11}$$

2.3. Joint Estimation of the Vocal-tract Parameters

The optimal values of the vocal-tract parameters are determined iteratively starting from initial values, so that the formant frequencies calculated from the parameters agree with the respective target frequencies. Peaks in the transfer function of the vocal tract are frequently referred to as vocal-tract resonances, but we also designate vocal-tract resonances as formants. The initial cross-sectional area is here denoted by $A^0(i)$ and the sectional length by L^0. The section index, i, is omitted for the length parameter because it is the same for all sections. The updating equations in the iterative optimization are thus written as

$$A^{k+1}(i) = A^k(i) + \Delta A^k(i) \tag{12}$$

and

$$L^{k+1} = L^k + \Delta L^k, \tag{13}$$

where k is the index of iterations.

If the area parameters are represented by the mode functions [Eq. (1)] and the sectional length is predicted from two mode coefficients [Eq. (2)], the updating quantities can be written as

$$\begin{aligned}\Delta A^k(i) &= A^{k+1}(i) - A^k(i)\\ &= \sum_{m=1}^{M}\gamma_m^{k+1}\phi_m(i) - \sum_{m=1}^{M}\gamma_m^k\phi_m(i)\\ &= \sum_{m=1}^{M}\Delta\gamma_m\phi_m(i)\end{aligned} \tag{14}$$

and

$$\Delta L^k = L_S(\gamma_1^{k+1},\gamma_2^{k+1}) - L_S(\gamma_1^k,\gamma_2^k), \tag{15}$$

where

$$\Delta\gamma_m = \gamma_m^{k+1} - \gamma_m^k \tag{16}$$

is the updating value for the mth mode coefficient. Taylor expansion of L_S gives

$$\begin{aligned}&L_S(\gamma_1^{k+1},\gamma_2^{k+1}) - L_S(\gamma_1^k,\gamma_2^k)\\ &\quad= \left(\Delta\gamma_1\frac{\partial}{\partial\gamma_1} + \Delta\gamma_2\frac{\partial}{\partial\gamma_2}\right)L_S(\gamma_1^k,\gamma_2^k),\end{aligned} \tag{17}$$

where the higher, nonlinear terms are omitted. Equation (15) can then be rewritten as

$$\Delta L^k = \sum_{m=1}^{2}\Delta\gamma_m\hat{\phi}_m \tag{18}$$

by setting

$$\hat{\phi}_m = \frac{\partial}{\partial\gamma_m}L_S(\gamma_1^k,\gamma_2^k). \tag{19}$$

Equations (14) and (18) indicate that the problem concerns how the value of $\Delta\gamma_m$ should be determined to

update adequately the area and length parameters. The optimal value of $\Delta\gamma_m$ is determined here by minimizing the following cost function

$$C = \sum_{n=1}^{N} \left\{ \frac{F_n^{k+1} - \mathcal{F}_n}{F_n^k} \right\}^2 \tag{20}$$

representing the squared error between the target frequency, $\mathcal{F}_n$, and the frequency after updating, F_n^{k+1}, normalized by the frequency before updating, F_n^k, for each formant. N is the number of target formants. To solve the problem, we should consider the dependence of the formant frequencies on $\Delta\gamma_m$.

Substitution of Eqs. (14) and (18) into the perturbation relation in Eqs. (5) and (6) gives

$$\begin{aligned} \frac{\Delta F_n}{F_n^k} &= \sum_{i=1}^{N_A} \frac{S(n,i)}{A^k(i)} \sum_{m=1}^{M} \Delta\gamma_m \phi_m(i) \\ &= \sum_{m=1}^{M} \Delta\gamma_m T_{nm} \end{aligned} \tag{21}$$

and

$$\begin{aligned} \frac{\Delta \hat{F}_n}{F_n^k} &= \sum_{i=1}^{N_A} \frac{\hat{S}(n,i)}{L^k} \sum_{m=1}^{2} \Delta\gamma_m \hat{\phi}_m \\ &= \sum_{m=1}^{2} \Delta\gamma_m \hat{T}_{nm}, \end{aligned} \tag{22}$$

where

$$T_{nm} = \sum_{i=1}^{N_A} \frac{S(n,i)}{A^k(i)} \phi_m(i) \tag{23}$$

and

$$\hat{T}_{nm} = \sum_{i=1}^{N_A} \frac{\hat{S}(n,i)}{L^k} \hat{\phi}_m. \tag{24}$$

It should be noted that the values of T_{nm} and $\hat{T}_{nm}$ are known; this is because $A^k(i)$ and L^k are parameters before updating, and the sensitivity functions are calculated from them. $\phi_m(i)$ is provided beforehand, and $\hat{\phi}_m$ is calculated from the mode coefficients before updating. In addition, $L^k = L_S(\gamma_1^k, \gamma_2^k)$.

From Eqs. (21) and (22), the formant frequencies after updating are now derived by summing the frequencies before updating and the frequency changes due to the updating of the vocal-tract parameters, i.e.,

$$\begin{aligned} F_n^{k+1} &= F_n^k + \Delta F_n + \Delta \hat{F}_n \\ &= \left\{ 1 + \sum_{m=1}^{M} \Delta\gamma_m \tilde{T}_{nm} \right\} F_n^k, \end{aligned} \tag{25}$$

where

$$\tilde{T}_{nm} = T_{nm} + \hat{T}_{nm} \tag{26}$$

and $\hat{T}_{nm} = 0$ for $m > 2$. We then define z_n as the frequency difference between the target formant, $\mathcal{F}_n$, and the formant before updating, F_n^k:

$$z_n = \frac{\mathcal{F}_n - F_n^k}{F_n^k}. \tag{27}$$

The value of z_n is also known. The formant error after updating can now be written as

$$\begin{aligned} F_n^{k+1} - \mathcal{F}_n &= \left\{ 1 + \sum_{m=1}^{M} \Delta\gamma_m \tilde{T}_{nm} \right\} F_n^k - (1 + z_n) F_n^k \\ &= \left\{ -z_n + \sum_{m=1}^{M} \Delta\gamma_m \tilde{T}_{nm} \right\} F_n^k \end{aligned} \tag{28}$$

and the cost function is finally written as

$$\begin{aligned} C &= \sum_{n=1}^{N} \left\{ \frac{F_n^{k+1} - \mathcal{F}_n}{F_n^k} \right\}^2 \\ &= \sum_{n=1}^{N} \left\{ -z_n + \sum_{m=1}^{M} \Delta\gamma_m \tilde{T}_{nm} \right\}^2. \end{aligned} \tag{29}$$

Because the cost function is a quadratic function of $\Delta\gamma_m$, the optimality condition, $\partial C / \partial \Delta\gamma_m = 0$, leads to a set of linear equations

$$\sum_{m=1}^{M} \Delta\gamma_m \tilde{T}_{nm} = z_n \quad (n = 1, 2, \ldots, N) \tag{30}$$

from which the values of $\Delta\gamma_m$ are obtained. These linear equations are written in matrix form as

$$\tilde{\boldsymbol{T}} \boldsymbol{\Delta\gamma} = \boldsymbol{z}, \tag{31}$$

where $\tilde{\boldsymbol{T}}$ is a coefficient matrix comprising $\tilde{T}_{nm}$ for $1 \leq n \leq N$ and $1 \leq m \leq M$. $\boldsymbol{\Delta\gamma} = [\Delta\gamma_1\ \Delta\gamma_2\ \cdots\ \Delta\gamma_M]^{\mathrm{t}}$ is the parameter vector, and $\boldsymbol{z} = [z_1\ z_2\ \cdots\ z_N]^{\mathrm{t}}$ is the error coefficient vector. t represents the transposition. Finally, the value of $\boldsymbol{\Delta\gamma}$ is determined as

$$\boldsymbol{\Delta\gamma} = \tilde{\boldsymbol{T}}^{+} \boldsymbol{z}, \tag{32}$$

where $+$ represents the Moore-Penrose pseudo-inverse matrix [15]. The parameter values are updated using $\Delta\gamma_m$ as

$$A^{k+1}(i) = A^k(i) + \sum_{m=1}^{M} \Delta\gamma_m \phi_m(i) \tag{33}$$

and

$$L^{k+1} = L_S(\gamma_1^k + \Delta\gamma_1, \gamma_2^k + \Delta\gamma_2). \tag{34}$$

Equation (31) represents a set of linear equations with respect to the unknown parameters, $\Delta\gamma_m$. The number of the parameters, M, is greater or equal to the number of linear equations, N, that is given by the number of target formants such as $N \leq M$ in our estimation. When $N < M$, Eq. (32) indicates that the value of $\Delta\gamma_m$ is determined

under linear equations given in Eq. (31) so that the norm of the parameter vector, $\Delta\gamma^{\mathrm{t}}\Delta\gamma$, is minimized, because there are an infinite number of solutions to the equations. It is necessary to recall here that γ_m is the mode coefficient. The mode functions in Eq. (1) are determined through principal component analysis; hence, the first mode is the most important and has the largest variance. The variance associated with each mode decreases as the mode number increases.

To take this mode property into account, we determine the inverse matrix $\tilde{\boldsymbol{T}}^{+}$ such that a weighted norm, $\boldsymbol{\Delta\gamma}^{\mathrm{t}}\boldsymbol{W}\boldsymbol{\Delta\gamma}$, is minimized, where $\boldsymbol{W}$ is a positive definite matrix [16]. This problem can be solved using the Lagrange multiplier method. We define a cost function $d = \boldsymbol{\Delta\gamma}^{\mathrm{t}}\boldsymbol{W}\boldsymbol{\Delta\gamma} - (\tilde{\boldsymbol{T}}\boldsymbol{\Delta\gamma} - \boldsymbol{z})^{\mathrm{t}}\lambda$, where λ is the unknown multiplier. To minimize the cost function, we set $\partial d/\partial\boldsymbol{\Delta\gamma} = 2\boldsymbol{W}\boldsymbol{\Delta\gamma} - \tilde{\boldsymbol{T}}^{\mathrm{t}}\lambda = 0$, and obtain the relationship $2\boldsymbol{W}\boldsymbol{\Delta\gamma} = \tilde{\boldsymbol{T}}^{\mathrm{t}}\lambda$. Substitution of $\boldsymbol{\Delta\gamma} = 0.5\boldsymbol{W}^{-1}\tilde{\boldsymbol{T}}^{\mathrm{t}}\lambda$ into $\tilde{\boldsymbol{T}}\boldsymbol{\Delta\gamma} = \boldsymbol{z}$ gives $0.5\tilde{\boldsymbol{T}}\boldsymbol{W}^{-1}\tilde{\boldsymbol{T}}^{\mathrm{t}}\lambda = \boldsymbol{z}$, and the value of the multiplier is determined as $\lambda = 2(\tilde{\boldsymbol{T}}\boldsymbol{W}^{-1}\tilde{\boldsymbol{T}}^{\mathrm{t}})^{-1}\boldsymbol{z}$. Finally, the solution is given as $\boldsymbol{\Delta\gamma} = 0.5\boldsymbol{W}^{-1}\tilde{\boldsymbol{T}}^{\mathrm{t}}\lambda = \boldsymbol{W}^{-1}\tilde{\boldsymbol{T}}^{\mathrm{t}}(\tilde{\boldsymbol{T}}\boldsymbol{W}^{-1}\tilde{\boldsymbol{T}}^{\mathrm{t}})^{-1}\boldsymbol{z}$, and it indicates that the inverse matrix in our estimation method is given as

$$\tilde{\boldsymbol{T}}^{+} = \boldsymbol{W}^{-1}\tilde{\boldsymbol{T}}^{\mathrm{t}}(\tilde{\boldsymbol{T}}\boldsymbol{W}^{-1}\tilde{\boldsymbol{T}}^{\mathrm{t}})^{-1}. \tag{35}$$

The value of $\Delta\gamma_m$ is scaled after the solution is obtained, so that the maximum change of the cross-sectional area, $\Delta A(i)/A^k(i)$ $(= \sum_{m=1}^{M}\Delta\gamma_m\phi_m(i)/A^k(i))$, is less than 10%. This is because Eqs. (5) and (6) apply for a small change in the vocal-tract parameters and formant frequency. The scaling factor is determined by dividing 0.1 by the maximum value of $\Delta A(i)/A^k(i)$ for $i = 1, 2, \ldots, N_A$, and the scaled version of $\Delta\gamma_m$ is used to update the area and length in Eqs. (33) and (34). If the cross-sectional area is negative for a specific vocal-tract section after updating, the area for that section is changed to a small positive value, such that $A^{k+1}(i) = \max(A^{k+1}(i), A_{\min})$, where $A_{\min}$ is the minimum area value.

2.4. Estimation Procedure

In this study, a set of vocal-tract area function data for 10 English vowels [13] was used to calculate the mean area $\overline{A(i)}$ and mode functions $\phi_m(i)$, where $i = 1, 2, \ldots, N_A$ and $m = 1, 2, \ldots, M$. Figure 1 shows the following steps of the estimation procedure. The estimation condition is the number of target formants, N, and the number of mode functions, M.

(1) Given the target formant frequencies, $\mathcal{F}_n$, for $n = 1, 2, \ldots, N$, the initial value of the area parameter was set as $A^0(i) = \overline{A(i)}$. This implies that the initial values of the mode coefficients were zero, and the initial value of the sectional length was set as $L^0 = L_S(0, 0)$. The iteration index was set to $k = -1$.

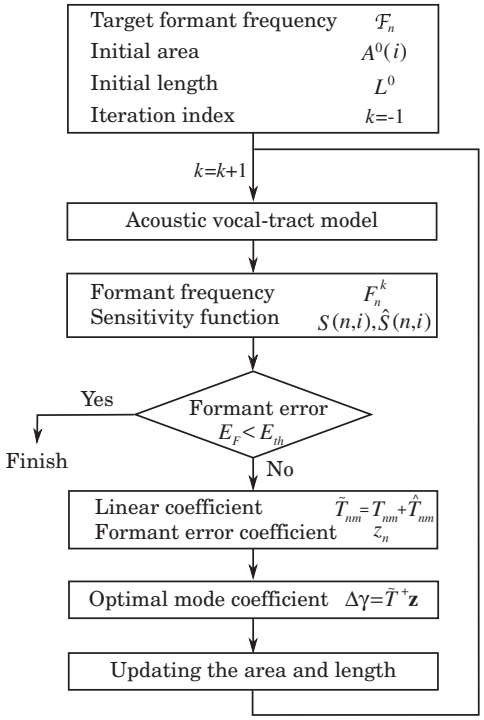

Fig. 1 Procedure for estimating the cross-sectional area and length of the vocal tract from specified formant frequencies.

(2) The iteration index was changed to $k = k + 1$. The frequency response of the vocal tract was calculated from $A^k(i)$ and L^k using an acoustic tube model of the vocal tract [17]. The formant frequency, F_n^k, was determined from an adequate peak of the vocal-tract spectrum. The sensitivity functions—$S(n, i)$ and $\hat{S}(n, i)$ in Eqs. (7) and (8)—were also calculated for the frequency of each formant.

(3) If the mean formant error $E_F = \frac{1}{N}\sum_{n=1}^{N}|F_n^k - \mathcal{F}_n|$ was smaller than the threshold, E_{th}, the values of the vocal-tract parameters was output as the estimation result and the procedure was terminated.

(4) The coefficient $\tilde{T}_{nm} = T_{nm} + \hat{T}_{nm}$ and the formant error z_n were calculated. The updating quantity of the mode coefficients, $\Delta\gamma_m$, was then determined as $\boldsymbol{\Delta\gamma} = \tilde{\boldsymbol{T}}^{+}\boldsymbol{z}$, where the pseudo-inverse matrix was calculated as indicated in Eq. (35).

(5) The vocal-tract parameters were updated from $\Delta\gamma_m$, as detailed in Eqs. (33) and (34). The procedure was repeated from step 2.

The frequency-domain acoustic tube model [17] is based on wave propagation in a lossy vocal tract. The model includes the effect of wall vibration, viscous friction loss, and heat conduction loss on the surface of the vocal-tract wall. The radiation impedance at the lips was given following the literature [18], and the glottis was supposed to be closed.

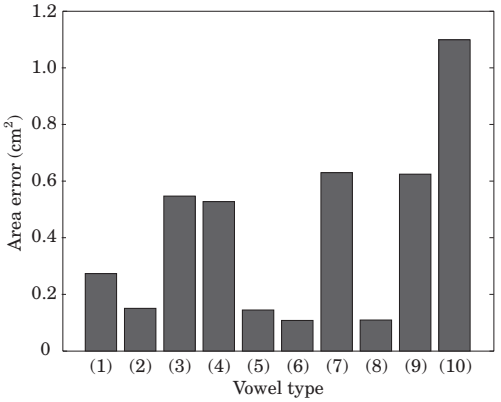

Fig. 2 Estimation error of the cross-sectional area as a function of vowels (1) /i/, (2) /ɪ/, (3) /ɛ/, (4) /æ/, (5) /ʌ/, (6) /ɑ/, (7) /ɔ/, (8) /o/, (9) /ʊ/, and (10) /u/. The error was calculated by taking the average of errors for all vocal-tract sections with each vowel.

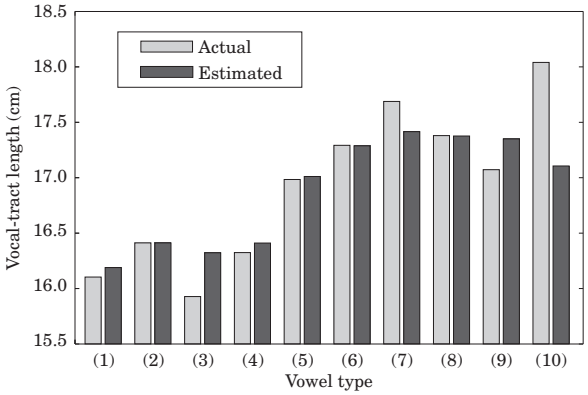

Fig. 3 Comparison of the actual and estimated length of the vocal tract. The light-gray bars indicate the actual length and the dark-gray bars the estimated length. The vocal-tract length was calculated by multiplying the sectional length by the number of vocal-tract sections.

3. ESTIMATION RESULTS

3.1. Estimation Conditions

We performed numerical experiments using area function data for 10 English vowels—/i/, /ɪ/, /ɛ/, /æ/, /ʌ/, /ɑ/, /ɔ/, /o/, /ʊ/, and /u/—by means of MRI measurements using a male native speaker [13]. The number of tube sections, N_A, was 44 for all vowels. We first calculated the transfer function and peak frequencies of the vocal tract from the data using an acoustic tube model [17]. These peak frequencies were used as the target frequencies, $\mathcal{F}_n$, and the area function data were taken as being correct for the inverse estimation. The mean cross-sectional area and mode functions were also computed from the data (these results are plotted in Fig. 2 of our previous paper [12]).

The termination condition on the formant error, E_{th}, was set to 1 Hz, and the maximum number of iterative computations was limited to 400. This limit was sufficiently large compared with the mean iteration number for convergence to be obtained. The weight for calculating the inverse matrix given in Eq. (35) was set as a diagonal matrix such that $W = \text{diag}\{w_1, w_2, \ldots, w_M\}$. The value of each component, w_m, was taken from the first M components of a sequence $\{0.0125, 0.0125, 0.067, 0.067, 0.25, 0.25, 1, 1\}$.

From the experimental results with Method II presented in the literature [12], the number of target formants was set to four ($N = 4$) and the number of mode functions was set to seven ($M = 7$) in this study.

3.2. Estimation Accuracy

Figure 2 shows the estimation error of the cross-sectional area for each vowel. The error was calculated from the cross-sectional area of each vocal-tract section such as $\frac{1}{N_A}\sum_{i=1}^{N_A}|A_e(i) - A_c(i)|$, where $A_e(i)$ and $A_c(i)$ represent the estimated and correct areas, respectively. The error was found to be small for (6) /ɑ/ and (8) /o/ (0.11 cm^2), and it was extremely large for (10) /u/ (1.09 cm^2). Thus, the estimation error depended strongly on the particular vowel. Figure 3 shows the total length of the vocal tract: the actual length is plotted using light-gray bars and the estimated length is indicated by dark-gray bars. The error was noticeable for (3) /ɛ/, (7) /ɔ/, (9) /ʊ/, and (10) /u/, and it should be noted that the area error was also large for these vowels.

Here, we show the mean estimation error calculated for all vowels. With regard to the cross-sectional area, the mean estimation error was 0.42 cm^2. The average actual cross-sectional area was about 2.05 cm^2; therefore, the estimation error was about 20.4% of the actual area. With regard to the vocal-tract length, the mean error was 0.21 cm. The average actual vocal-tract length was about 16.9 cm; therefore, the estimation error was about 1.23% of the actual length.

It was also proved that the area error for the present method (0.42 cm^2) was only slightly larger than the estimation error of 0.38 cm^2 calculated for the same vowels using our previous method (Method II in the literature [12]). To compare the estimation accuracy, the inverse matrix in the previous method (Eq. (32) in the literature [12]) was calculated by Eq. (35) in this study using the same weighting matrix. Note that the vocal-tract length was not estimated and fixed to the correct value in the previous method.

3.3. Estimated Area Function and Vocal-tract Spectrum

Figure 4 shows the area function estimated from the formants and frequency response of the vocal tract. The

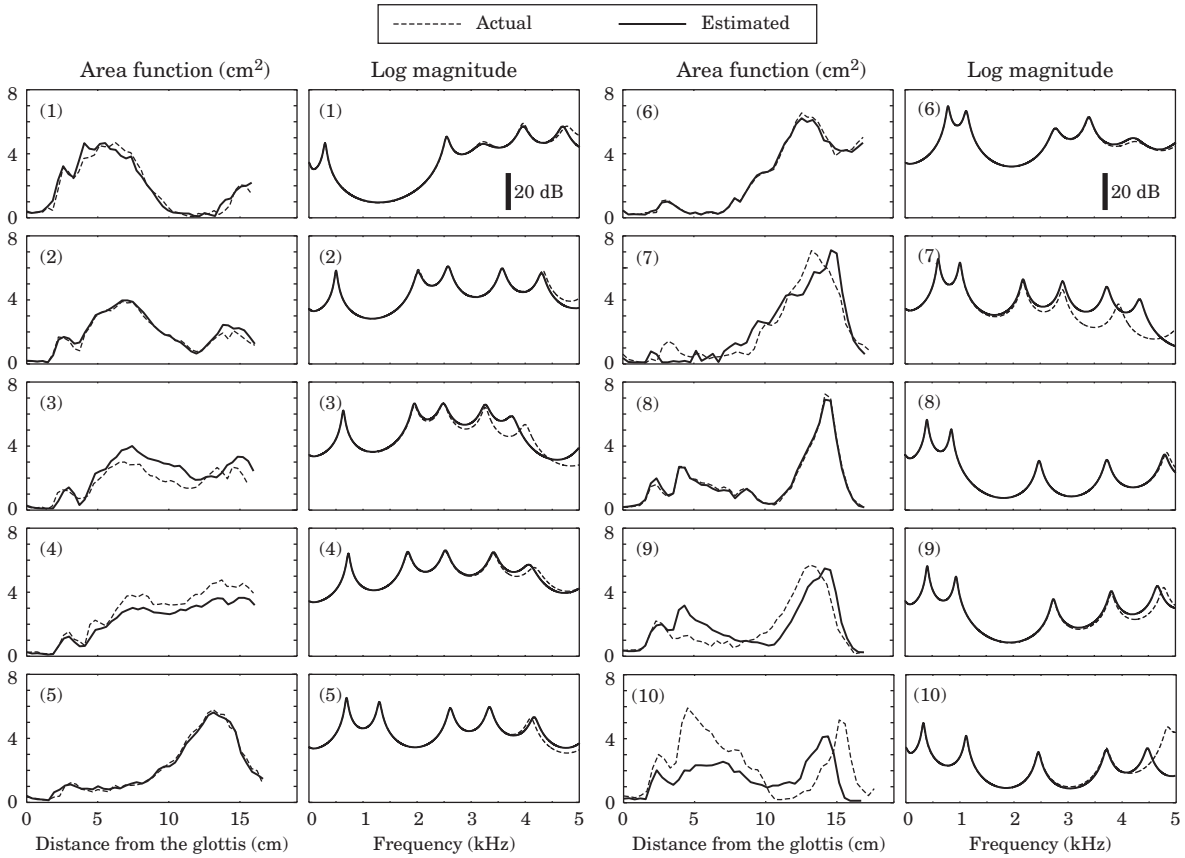

Fig. 4 Estimated area functions and frequency responses for (1) /i/, (2) /ɪ/, (3) /ɛ/, (4) /æ/, (5) /ʌ/, (6) /ɑ/, (7) /ɔ/, (8) /o/, (9) /ʊ/, and (10) /u/. The broken lines indicate the data based on MRI measurements (Story and Titze, 1998) from which the target formant frequencies were specified. The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

cross-sectional area was plotted as a function of the distance from the glottis. The estimated area function (solid lines) agrees well with the actual values (broken lines); this was especially the case with (2) /ɪ/, (5) /ʌ/, (6) /ɑ/, and (8) /o/. For (3) /ɛ/, (4) /æ/, (9) /ʊ/, and (10) /u/, the error is noticeable and the error was distributed over the whole vocal-tract region. This discrepancy was also evident with (7) /ɔ/, where an unexpected constriction of the area function can be observed. There, the cross-sectional area was fixed to a small positive value of 0.1 cm^2 for those sections; that was because the area for those sections became negative as a result of the iterative calculation of the mode coefficients.

The frequency response of the vocal tract was calculated from the actual and estimated area function data, and it is indicated in Fig. 4 by broken and solid lines, respectively. The frequencies for four formants were specified as the target, and the estimated frequency response proved accurate over the frequency range up to 5 kHz except for (3) /ɛ/, (7) /ɔ/, and (10) /u/. For (4) /æ/, the estimated frequency response was mostly accurate up to 5 kHz despite the relatively large estimation error of the cross-sectional area. These numerical results for /æ/ suggest the existence of multiple vocal-tract shapes, which produce the same formant frequencies. This one-to-many relationship is a possible cause of the estimation error, and we investigate this problem in the next subsection with regard to the dependency on the initial value of the cross-sectional area.

3.4. Dependency on Initial Value of Cross-sectional Area

To examine the one-to-many relationship between the acoustic and vocal-tract parameters, the initial value of the cross-sectional area was set as

$$A^0(i) = (1-\alpha)\cdot\overline{A(i)} + \alpha\cdot A_T(i), \tag{36}$$

where α is the interpolation rate between the mean cross-sectional area, $\overline{A(i)}$, and the actual cross-sectional area for the test vowel, $A_T(i)$, and $i = 1, 2, \ldots, N_A$. The initial value gradually approached the actual value as α increased from 0 to 0.9 in 0.05 increments. The initial value of the sectional length was set to $L^0 = (\gamma_1, \gamma_2)$, where γ_1 and γ_2 are mode coefficients computed from the value of $A^0(i)$. Figure 5 shows the estimation error for (4) /æ/, (9) /ʊ/, and (10) /u/, for which the area error was relatively large

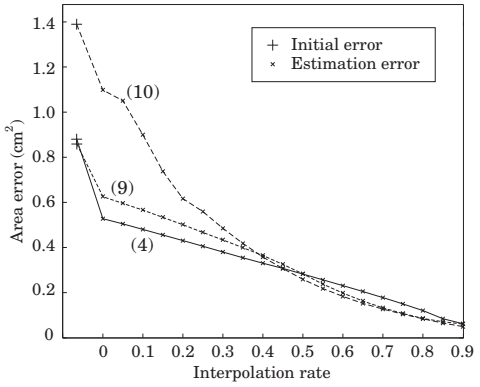

Fig. 5 Estimation error for (4) /æ/, (9) /ʊ/, and (10) /u/ as a function of the interpolation rate of the initial value.

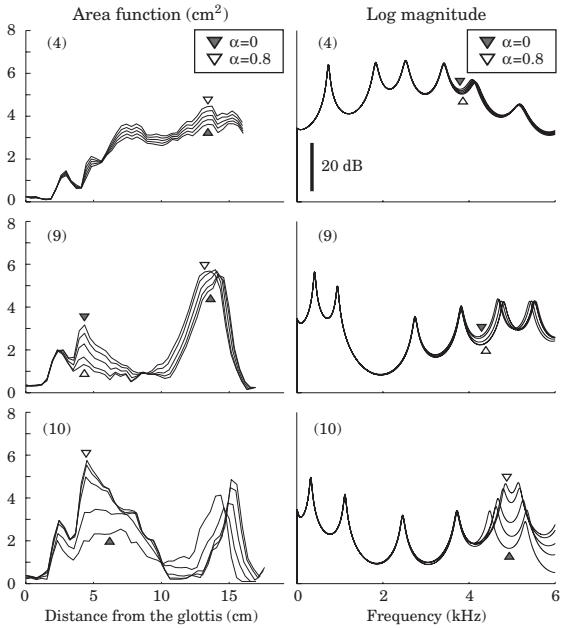

Fig. 6 Estimated area functions and frequency responses for (4) /æ/, (9) /ʊ/, and (10) /u/. The plots overlap for the interpolation rate (α) of 0, 0.2, 0.4, 0.6, and 0.8. The solid and open triangles indicate the results for $\alpha = 0$ and 0.8, respectively.

among the vowels. The plus mark shows the initial error before optimization, i.e., the difference between the initial area and the target area calculated as $\frac{1}{N_A}\sum_{i=1}^{N_A}|\overline{A(i)} - A_T(i)|$. The cross mark shows the estimation error after the optimization for each value of the interpolation rate. Figure 5 shows that the area error decreased monotonically as α increased. This also indicates that the estimated area function was different for each value of α.

Figure 6 shows the estimated cross-sectional area and the frequency response of the vocal tract. The values of the interpolation rate were 0, 0.2, 0.4, 0.6, and 0.8. The area function corresponding to $\alpha = 0.8$ was the most accurate and closest to the correct value. For each vowel, the frequencies for the first four formants were identical irrespective of the α value, which indicates that the requirement for the formants was satisfied. However, the estimated area functions depended on the α value, and the frequency response changed in the frequency range above the fourth formant.

3.5. Independent Treatment of the Length Parameter

Both the cross-sectional area and length of the vocal tract can affect formant frequencies in accordance with the perturbation relationships given in Eqs. (5) and (6). These relationships also imply that area and length parameters can be complementary in achieving target formants. In other words, there can be multiple solutions of the vocal-tract parameters resulting in the same formant frequencies. This redundancy was resolved in our method by predicting the sectional length as a function of coefficients for the first and second modes [Eq. (2)]. Here, we compare the estimation accuracy with that of an alternative method in which the length parameter is independent of the area parameters, and thus examine the effectiveness of our method.

In the alternative method, the change in the formant frequency, $\Delta\hat{F}_n$, due to a small change in the sectional length, ΔL^k, is obtained from the perturbation relationship [Eq. (6)] as

$$\Delta\hat{F}_n = \left\{\hat{T}_n \Delta L^k\right\} F_n^k, \tag{37}$$

where $\hat{T}_n = \sum_{i=1}^{N_A}\hat{S}(n,i)/L^k$. If we let $T_{n(M+1)} = \hat{T}_n$ and $\Delta\gamma_{M+1} = \Delta L^k$, Eq. (25) can be expressed as

$$\begin{aligned} F_n^{k+1} &= \left\{1 + \sum_{m=1}^{M}\Delta\gamma_m T_{nm} + \hat{T}_n\Delta L^k\right\}F_n^k \\ &= \left\{1 + \sum_{m=1}^{M+1}\Delta\gamma_m T_{nm}\right\}F_n^k. \end{aligned} \tag{38}$$

The optimal value of ΔL^k can then be determined from the cost function C as the $(M+1)$th component of $\boldsymbol{\Delta\gamma}$ in Eq. (32), and the length parameter is updated as $L^{k+1} = L^k + \Delta L^k$ in the iterative estimation procedure.

Estimation results for 10 English vowels showed that the mean estimation error of the cross-sectional area increased from 0.42 cm^2 for our method to 0.44 cm^2 for the alternative method. With regard to the vocal-tract length, the mean error increased significantly from 0.21 for our method to 0.91 cm for the alternative one. The estimation error conspicuously increased for (1) /i/, (2) /ɪ/, (6) /ɑ/, and (7) /ɔ/, and these results are plotted in Fig. 7. The estimated vocal-tract length was 1.43 cm shorter than the actual length for (1) /i/, 0.85 cm longer for (2) /ɪ/, and 0.89 cm shorter for (6) /ɑ/, respectively (see Fig. 4 also for the comparison of the estimated results). The error was

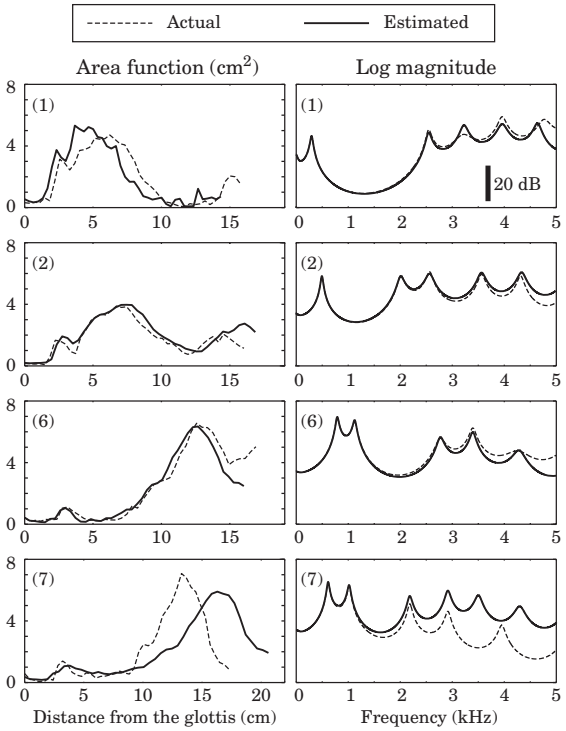

Fig. 7 Estimated area functions and frequency responses for (1) /i/, (2) /ɪ/, (6) /ɑ/, and (7) /ɔ/, where the sectional length of the vocal tract was treated as an independent parameter. The broken lines indicate the data based on MRI measurements (Story and Titze, 1998) and the solid lines show the estimation results.

largest for (7) /ɔ/. The estimated vocal-tract length was 3.40 cm longer, and this error corresponded to 19.2% of the actual length. The estimated cross-sectional area function was quite different from the actual one, but the frequencies of the first four formants exactly agreed with those of the target formants.

4. SUMMARY AND DISCUSSION

We presented a method for estimating the length and cross-sectional area of the vocal tract jointly from formants. Because the relationship between the vocal-tract area function and formant frequency is highly nonlinear, an optimal solution was obtained based on an iterative procedure. In each iteration, a sensitivity function and least-squares optimization were used to update the parameters. The sensitivity function represents the change in the formant frequency due to a small perturbation of the area or length parameter. Therefore, this relationship allows the updating quantity of the parameters to be accurately determined. The cross-sectional areas for all vocal-tract sections were represented by the sum of several mode functions, and the sectional length was predicted from the values of the first and second mode coefficients. With this constraint, a change in the first or second mode coefficient makes a change in the area and also a change in the length, and then the values of area and length parameters were adjusted jointly.

We evaluated the basic accuracy of our method for 10 English vowels. As a result, the estimation error was largely dependent of vowel type. We also found that the estimation error decreased monotonically as the initial value approached the actual value. In combination with the spectral analysis of the vocal-tract transfer function, we observed that the estimated vocal-tract spectrum was in good agreement with the actual spectrum below the frequency for the fourth formant despite the change in the vocal-tract shape. The most prominent dependency and drastic change in the area and frequency response was found for /u/, for which a close alignment of the fifth and sixth formants was observed. It is likely that the estimation procedure was unable to reproduce such formant alignment unless the initial value was set very close to the correct cross-sectional areas.

Experimental result clearly shows that these are multiple vocal-tract shapes that produce the same frequencies for the first four formants. This result is interpreted as a typical example of a one-to-many relationship between the acoustic and vocal-tract parameters. The one-to-many relationship or non-uniqueness problem features prominently in speech inversion studies [3,19]. The results presented in this paper showed that our method is to some extent able to deal with the non-uniqueness problem: the use of the sensitivity function and mode function was effective in reducing the degrees of freedom of the area function and in determining the area function from formants with a satisfactory degree of accuracy. However, the inversion error was large for some particular vowels, and the experimental results suggested that the non-uniqueness problem can degrade the estimation accuracy to a high degree for these vowels.

The numerical result also indicates that the area error and length error were correlated such that both errors were large for the vowel /u/, for example. A similar trend was also found for /ɛ/, /ɔ/, and /ʊ/. The sectional length was determined using our method from the coefficients for the first and second modes of the area parameters. Therefore, the length error increased as the error for these modes increased, resulting in the correlation between the area and length errors for these vowels. On the other hand, it was questionable how the estimation accuracy and the relationship between the area and length parameters change when the sectional length was estimated independently of the area parameters. In Sect. 3.5, the estimation error increased when the sectional length was treated as an independent parameter. This result confirmed the effectiveness of using the predicting function of the sectional length in our method. In addition, it was suggested that there was

a complementary relationship between the area and length parameters in achieving target formant frequencies, and this complementary relationship caused the increase in the estimation error.

Further studies will be dedicated to examining (1) continuous estimation of the vocal-tract shape, (2) the problem of inter-speaker variability, and (3) the use of speech spectrum parameter as the acoustic target. The first issue is related to the estimation of dynamic changes in the vocal-tract shape from a temporal sequence of formant frequencies. We can then make use of the temporal smoothness of the vocal-tract parameters. It is important to examine if the effect of the non-uniqueness problem is reduced by considering smoothness. The second issue is also important, because the area-function data used in this experiment were taken from one subject, and the accuracy of our method should be examined with multiple speakers. With regard to the third issue, it is convenient if the vocal-tract shape is estimated using speech spectrum parameters without extracting formant frequencies from speech signals. Toward this end, we will further extend the estimation method by using the relationship between cepstrum parameters and formant frequencies in addition to the acoustic sensitivity function.

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] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. Speech Audio Process.*, **2**, 133–150 (1994).

[2] K. Iskarous, "Vowel constrictions are recoverable from formants," *J. Phonet.*, **38**, 375–387 (2010).

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

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

[5] M. M. Sondhi and J. R. Resnick, "The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis," *J. Acoust. Soc. Am.*, **73**, 985–1002 (1983).

[6] 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. Phonet.*, **35**, 20–39 (2007).

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

[8] 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).

[9] M. Greenspan, "Simple derivation of the Boltzmann-Ehrenfest adiabatic principle," *J. Acoust. Soc. Am.*, **27**, 34–35 (1955).

[10] G. Fant and S. Pauli, "Spatial characteristics of vocal tract resonance modes," *Speech Commun. Semin. Stockh.*, pp. 1–3 (1974).

[11] G. Fant, "Vocal-tract area and length perturbations," *Speech Transm. Lab. Q. Prog. Status Rep.*, **4**, 1–14 (1975).

[12] 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).

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

[14] B. H. Story, "Vocal tract modes based on multiple area function sets from one speaker," *J. Acoust. Soc. Am.*, **125**, EL141–EL147 (2009).

[15] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. (Johns Hopkins University Press, Baltimore, 1996), pp. 257–258.

[16] H. Yehia and F. Itakura, "A method to combine acoustic and morphological constraints in the speech production inverse problem," *Speech Comun.*, **18**, 151–174 (1996).

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

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

[19] B. S. Atal, J. J. Chang, M. V. Mathews and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *J. Acoust. Soc. Am.*, **63**, 1535–1555 (1978).