Estimating area function of the vocal tract from formants using a sensitivity function and least-squares

Kaburagi, Tokihiko
Faculty of Design, Kyushu University

Takano, Tetsuro
Human Techno System Co. Ltd.

Sakamoto, Yuki
Graduate School of Design, Kyushu University

https://hdl.handle.net/2324/7178813

出版情報：Acoustical Science and Technology. 34 (5), pp.301-310, 2013. 日本音響学会
バージョン：

Acoust. Sci. & Tech. **34**, 5 (2013)

PAPER

Estimating area function of the vocal tract from formants using a sensitivity function and least-squares

Tokihiko Kaburagi[1,*], Tetsuro Takano[2] and Yuki Sakamoto[3]

[1]*Faculty of Design, Kyushu University,*
4–9–1 Shiobaru, Minami-ku, Fukuoka, 815–8540 Japan
[2]*Human Techno System Co. Ltd.,*
Kawasho Bldg., 1–11–2 Nihon-bashi, Ningyo-cho, Chuo-ku, Tokyo, 103–0013 Japan
[3]*Graduate School of Design, Kyushu University,*
4–9–1 Shiobaru, Minami-ku, Fukuoka, 815–8540 Japan

(*Received 3 September 2012, Accepted for publication 24 January 2013*)

Abstract: We present methods for estimating the cross-sectional area function of the vocal tract from formant frequencies. They extend the work of Story (*J. Acoust. Soc. Am.*, **119**, 715–718, 1996) based on a sensitivity function representing the change in the formant frequency due to a perturbation of the cross-sectional area. In Method I, the area function is estimated through an iterative procedure that uses the sensitivity function as the basis function to optimize the area function that produces the target frequencies. In Method II, a mode function linearly expands the area function. The estimation is performed by optimizing the value of each mode coefficient, where the sensitivity function is used as a constraint in the optimization. As a specific feature, the summing weight of sensitivity functions in Method I and mode functions in Method II is determined by minimizing an objective function representing the frequency error of every formant. By using existing area function data for English vowels, we compare the performance of each method with respect to the estimation accuracy and convergence speed. The results show that our methods can effectively reduce degrees of freedom of the area function and quickly obtain the optimal solution with fair accuracy.

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

PACS number: 43.70.Bk [doi:10.1250/ast.34.301]

1. INTRODUCTION

In producing speech, movements of the articulatory organs such as the jaw, lips, and tongue control the shape of the vocal tract. In addition, the transfer function of the vocal tract, and hence the formants, are essentially determined by specific information of the vocal-tract geometry; i.e., the cross-sectional area function. The inverse problem of speech production has attracted much attention thus far [1,2] in an attempt to establish ultimately an effective method to estimate the hidden, invisible state of the vocal tract from speech without special observational tools such as X-ray imaging. One approach to this inversion problem is the conversion of acoustic features of speech to an area function for the vocal tract. Another is the conversion of speech acoustics to parameter values representing the position and shape of the articulatory organs.

The latter approach is advantageous in reducing the degrees-of-freedom in the inversion, because the number of articulatory parameters is in general smaller than the number of vocal-tract sections for which the area function is determined. Schroeter and Sondhi [3], for example, used an 11-parameter model of the articulators and estimated their values from speech with a constraint on the smoothness of articulatory movements, where the forward relation between the articulatory parameters and the speech acoustics was computed via the area function and an acoustic tube model of the vocal tract [4]. Such a structure-based framework was also used in other studies [5–7]. Alternatively, with the use of data collected by measuring the motion of articulatory organs with speech, the forward relation was more directly represented using statistical methods and was applied to the inversion [8–11].

The inversion problem according to the former approach was first investigated by Schroeder [12] and

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

Mermelstein [13] using perturbation theory 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 using the Fourier expansion of the logarithm of the area function, and showed how the Fourier coefficient is related to the poles and zeros of the vocal-tract impedance at the lips. Actually, accurate determination of the input impedance is a hard task and requires acoustical input-output measurement [12,14,15]. They therefore assumed the vocal-tract shape to be antisymmetric around the center point or be represented by low-order Fourier components, and estimated the cross-sectional area from formants. Sondhi and Resnick [16] went further and presented an inverse problem of the sound wave propagation in the tract for both lossless and lossy cases with the acoustic measurement of the vocal tract. An inversion method was also studied in which the direct relationship between the vocal-tract shape and the speech spectrum was determined by using a data set of the area function and a statistical model [17].

In line with perturbation theory, Story [18] used the sensitivity function and proposed a method for estimating the area function from specified target frequencies of the formants. The sensitivity function is derived based on a theorem by Ehrenfest and represents the change in the formant frequency due to a small perturbation of the cross-sectional area of the vocal tract [19–21]. Starting from an initial value, the area function is optimized iteratively to minimize the formant error, where the sensitivity function is used as a basis function by which the area function is updated. Here, the number of target formants is the same as the number of basis sensitivity functions. The scaling factor of each sensitivity function is determined uniquely from the formant and hence the so-called one-to-many relation in the inversion [5,12] does not take place in this framework, although the estimation result may depend on the initial value. Story's method is superior to the perturbation methods [12,13] in that the antisymmetric assumption imposed on the vocal-tract shape is no longer necessary.

In Story's method, however, the scaling factor is empirically determined and the method does not guarantee the simultaneous convergence of all formants. Because of these drawbacks, many iterations are needed to converge on the solution. To overcome the problem, we present below a method that determines the weighting parameter through a least-squares technique. Using the explicit criterion on the formant error, our method improves the convergence property, and the error of every formant decreases effectively with each iteration. We examine the accuracy of the method quantitatively using the actual data of the cross-sectional area function for ten English vowels obtained by MRI measurements [22].

2. ESTIMATION METHOD

In this section, we present methods for determining the cross-sectional area function of the vocal tract from specified values of the formant frequencies. First, the definition of the sensitivity function is explained and an overview of the estimation method proposed by Story [18] is given. Two methods, designated Method I and II, are then described in the succeeding subsections. Method I is based on the sensitivity function and the least-squares technique. Just as in Story's method, Method I uses the sensitivity function as the basis function to gradually optimize the area function that produces the target formant frequencies. In Method II, the area function is expressed by overlapping a number of mode functions. The value of the summing coefficient in this expression is determined from the formants through least-squares, where the sensitivity function is used as a constraint in the optimization to relate the change in the area function to the change in the formant frequency. We did not estimate the length of the vocal tract in this study.

2.1. Sensitivity Function and Story's Inverse Estimation Method

Using the sensitivity function derived by Fant and Pauli [20], the change in formant frequencies for a small perturbation of the cross-sectional area is represented by

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

where $S(n,i)$ is the sensitivity function related to the nth formant frequency and ith tube section of the vocal tract, F_n is the formant frequency ($n = 1, 2, \ldots, N$), $A(i)$ is the cross-sectional area of each tube section ($i = 1, 2, \ldots, L$), and L is the total number of sections. $S(n,i)$ can be calculated via

$$S(n,i) = \frac{E_{\mathrm{K}}(n,i) - E_{\mathrm{P}}(n,i)}{E(n)}, \tag{2}$$

where $E_{\mathrm{K}}(n,i)$ and $E_{\mathrm{P}}(n,i)$ are kinetic and potential energies defined as

$$E_{\mathrm{K}}(n,i) = \frac{1}{2}\frac{\rho l(i)}{A(i)}\,|U(n,i)|^2 \tag{3}$$

and

$$E_{\mathrm{P}}(n,i) = \frac{1}{2}\frac{A(i)l(i)}{\rho c^2}\,|P(n,i)|^2. \tag{4}$$

$U(n,i)$ and $P(n,i)$ are the Fourier transforms of the volume velocity and pressure at the frequency of the nth formant, ρ is the air density, and $l(i)$ the length of the tube section. $E(n)$ is the total energy given by

$$E(n) = \sum_{i=1}^{L} \{E_P(n,i) + E_K(n,i)\}. \tag{5}$$

The relationship between the area function and formant frequency is in principle nonlinear, but is linearized in Eq. (1) assuming small changes in the cross-sectional area and formant.

Story proposed a method for estimating the area function from specified formant frequencies [18]. Starting from an initial value, $A^0(i)$, the area function is iteratively tuned such that

$$A^{k+1}(i) = A^k(i) + \alpha \sum_{n=1}^{N} z_n S(n,i) \tag{6}$$

for $i = 1, 2, \ldots, L$, where k is the index of iteration and N is the number of formants. z_n is the relative error of the formant

$$z_n = \frac{\hat{F}_n - F_n}{F_n}, \tag{7}$$

where $\hat{F}_n$ is the target formant frequency and F_n is the formant calculated from the area function, $A^k(i)$, by using an acoustic tube model of the vocal tract. Note that $S(n,i)$ is also calculated from $A^k(i)$ in a similar manner from Eqs. (2) through (5).

The method is based on the idea that the sensitivity function can be regarded as a linear basis function in determining the updating term for the cross-sectional area. If $S(n,i)$ is positive for specific values of n and i, the nth formant frequency would increase (decrease) from Eq. (1) by increasing (decreasing) the cross-sectional area of ith tube section. Therefore, it is expected that the area function is effectively optimized using the iterative procedure. The actual updating term, $z_n S(n,i)$, is summed for every formant in the procedure. In addition, an unknown weighting parameter, α, scales this term.

2.2. Estimation Method Using Least-Squares Optimization (Method I)

Story's method has a drawback in that the value of the parameter (α) should be determined empirically; it is actually set to 10 in the literature [18]. In addition, the method does not guarantee that every formant simultaneously converges to the target. In general, a large number of iterations are needed to obtain a solution. To overcome the problem, we now show that least-squares optimization can be used to determine the updating term of the area function. We set $\beta_n = \alpha z_n$ and rewrite the updating equation as

$$A^{k+1}(i) = A^k(i) + \sum_{n=1}^{N} \beta_n S(n,i). \tag{8}$$

The optimal value of β_n is then found as follows.

To minimize the formant error, we use an explicit criterion. First, the formant frequency after updating, F_n^{k+1}, can be written as follows using the relationship given in Eq. (1):

$$\begin{aligned} F_n^{k+1} &= F_n^k + \Delta F_n \\ &= \left\{ 1 + \sum_{i=1}^{L} S(n,i) \frac{\Delta A(i)}{A^k(i)} \right\} F_n^k, \end{aligned} \tag{9}$$

where ΔF_n is the difference in the formant frequency before and after updating. In addition, the updating term of the cross-sectional area, $\Delta A(i)$, is defined from Eq. (8) as

$$\begin{aligned} \Delta A(i) &= A^{k+1}(i) - A^k(i) \\ &= \sum_{n=1}^{N} \beta_n S(n,i). \end{aligned} \tag{10}$$

The right side of Eq. (1) is then rewritten as

$$\begin{aligned} \sum_{i=1}^{L} S(n,i) \frac{\Delta A(i)}{A^k(i)} &= \sum_{p=1}^{N} \beta_p \sum_{i=1}^{L} \frac{S(n,i)S(p,i)}{A^k(i)} \\ &= \sum_{p=1}^{N} \beta_p T_{np}, \end{aligned} \tag{11}$$

where

$$T_{np} = \sum_{i=1}^{L} \frac{S(n,i)S(p,i)}{A^k(i)}. \tag{12}$$

Note that the actual value of T_{np} can be determined, because $A^k(i)$ is the area function before updating and $S(n,i)$ can be calculated from $A^k(i)$. From Eqs. (9) and (11), we then obtain the following relation:

$$F_n^{k+1} = \left\{ 1 + \sum_{p=1}^{N} \beta_p T_{np} \right\} F_n^k. \tag{13}$$

We now obtain the error of the formant frequency, E_n, after updating using Eqs. (7) and (13):

$$\begin{aligned} E_n &= F_n^{k+1} - \hat{F}_n \\ &= \left\{ 1 + \sum_{p=1}^{N} \beta_p T_{np} \right\} F_n^k - (z_n + 1) F_n^k \\ &= \left\{ -z_n + \sum_{p=1}^{N} \beta_p T_{np} \right\} F_n^k. \end{aligned} \tag{14}$$

Here, we consider the relative error E_n / F_n^k and minimize it using least-squares. The total cost function is defined as

$$\begin{aligned} C &= \sum_{n=1}^{N} \left\{ \frac{E_n}{F_n^k} \right\}^2 \\ &= \sum_{n=1}^{N} \left\{ -z_n + \sum_{p=1}^{N} \beta_p T_{np} \right\}^2. \end{aligned} \tag{15}$$

C is a quadratic function of the unknown parameter, β_p ($p = 1, 2, \ldots, N$), and the optimal value of β_p is determined from the condition $\partial C / \partial \beta_p = 0$, giving the relation

$$\sum_{p=1}^{N} T_{np}\beta_p = z_n \quad (n = 1, 2, \ldots, N). \tag{16}$$

This represents a set of linear simultaneous equations, and can be rewritten in matrix form as

$$T\boldsymbol{\beta} = \boldsymbol{z}. \tag{17}$$

T is a $N \times N$ symmetrical coefficient matrix where the (n, p) component is given by T_{np} $(= T_{pn})$. $\boldsymbol{\beta}$ is the parameter vector

$$\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \cdots \ \beta_N]^{\mathrm{t}} \tag{18}$$

and $\boldsymbol{z}$ is the error coefficient vector

$$\boldsymbol{z} = [z_1 \ z_2 \ \cdots \ z_N]^{\mathrm{t}}. \tag{19}$$

The superscript t signifies the transposition. Finally, the optimal value of the coefficient $\boldsymbol{\beta}$ can be determined via

$$\boldsymbol{\beta} = T^{-1}\boldsymbol{z} \tag{20}$$

and the area function is updated from Eq. (8).

2.3. Estimation Method Using a Mode Decomposition of the Area Function (Method II)

Story's method and our Method I both estimate the cross-sectional area function iteratively by adding sensitivity functions at each step of the procedure. However, the sensitivity function by definition represents the acoustic sensitivity for a small change in the area function, and simple addition of the sensitivity functions may result in an unnatural vocal-tract configuration. We therefore present another method, Method II, by using a mode decomposition of the area function [17,22] in addition to the least-squares technique.

Story and Titze [22] measured 3D vocal-tract shapes for ten English vowels using MRI and obtained their area functions, where the number of vocal-tract sections was the same for all vowels. They also performed a principal component analysis and computed mode functions such that

$$A(i) = A_0(i) + \sum_{m=1}^{M} \gamma_m \phi_m(i), \tag{21}$$

where $A_0(i)$ is the mean cross-sectional area of the ith section, γ_m is the coefficient for the mth component, and $\phi_m(i)$ is the mth mode function. M is the number of mode functions reconstructing the area function. If the value of γ_m is determined properly from given formant frequencies, we can expect that this alternative method is capable of estimating the area function within physiological tolerances.

To derive the estimation algorithm, the difference in the cross-sectional area before and after updating is represented as

$$\begin{aligned} \Delta A(i) &= A^{k+1}(i) - A^k(i) \\ &= \sum_{m=1}^{M} \Delta\gamma_m \phi_m(i), \end{aligned} \tag{22}$$

where

$$\Delta\gamma_m = \gamma_m^{k+1} - \gamma_m^k \tag{23}$$

is the difference in the coefficient. Equation (1) can then be rewritten as

$$\begin{aligned} \frac{\Delta F_n}{F_n} &= \sum_{i=1}^{L} \frac{S(n,i)}{A^k(i)} \sum_{m=1}^{M} \Delta\gamma_m \phi_m(i) \\ &= \sum_{m=1}^{M} \Delta\gamma_m T_{nm}, \end{aligned} \tag{24}$$

where

$$T_{nm} = \sum_{i=1}^{L} \frac{S(n,i)}{A^k(i)} \phi_m(i). \tag{25}$$

The definition of T_{nm} is different from that in Method I [see Eq. (12)], but the value of T_{nm} can again be determined before updating.

The formant frequency after updating is then given as

$$\begin{aligned} F_n^{k+1} &= F_n^k + \Delta F_n \\ &= \left\{ 1 + \sum_{m=1}^{M} \Delta\gamma_m T_{nm} \right\} F_n^k, \end{aligned} \tag{26}$$

and the formant error after updating as

$$\begin{aligned} E_n &= F_n^{k+1} - \hat{F}_n \\ &= \left\{ 1 + \sum_{m=1}^{M} \Delta\gamma_m T_{nm} \right\} F_n^k - (z_n + 1) F_n^k \\ &= \left\{ -z_n + \sum_{m=1}^{M} \Delta\gamma_m T_{nm} \right\} F_n^k. \end{aligned} \tag{27}$$

The total cost function is now obtained as

$$\begin{aligned} C &= \sum_{n=1}^{N} \left\{ \frac{E_n}{F_n^k} \right\}^2 \\ &= \sum_{n=1}^{N} \left\{ -z_n + \sum_{m=1}^{M} \Delta\gamma_m T_{nm} \right\}^2. \end{aligned} \tag{28}$$

Finally, the optimization problem is formulated as a set of linear simultaneous equations as in the previous method such that

$$T\boldsymbol{\gamma} = \boldsymbol{z}. \tag{29}$$

T is a $N \times M$ coefficient matrix where the (n, m) component is given by T_{nm} in Eq. (25). $\boldsymbol{\gamma}$ is the following parameter vector

$$\boldsymbol{\gamma} = [\Delta\gamma_1 \ \Delta\gamma_2 \ \cdots \ \Delta\gamma_M]^{\mathrm{t}} \tag{30}$$

and $\boldsymbol{z}$ is the coefficient vector

$$z = [z_1 \ z_2 \ \cdots \ z_N]^{\mathrm{t}}. \tag{31}$$

We assume the condition $N \leq M$ indicating that the number of mode functions is equal to or greater than the number of formants to obtain a proper solution of the estimation problem. The optimal value of the parameter vector, $\boldsymbol{\gamma}$, is determined via

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

where the superscript $+$ signifies the pseudo-inverse of the matrix [23]. The area function is then updated as

$$\begin{aligned} A^{k+1}(i) &= A_0(i) + \sum_{m=1}^{M} \gamma_m^{k+1}\phi_m(i) \\ &= A_0(i) + \sum_{m=1}^{M} (\gamma_m^k + \Delta\gamma_m)\phi_m(i). \end{aligned} \tag{33}$$

2.4. Estimation Procedure

Figure 1 shows the procedure for estimating the vocal-tract cross-sectional area function from specified formant frequencies, $\hat{F}_n$ $(n = 1, 2, \ldots, N)$, and the initial area function, $A^0(i)$ $(i = 1, 2, \ldots, L)$. In Method II, mode functions, $\phi_m(i)$ $(m = 1, 2, \ldots, M)$, and the mean area function, $A_0(i)$, are also needed. The index k is set to 0.

First, we apply the area function, $A^k(i)$, to an acoustic tube model of the vocal tract. We use the frequency-domain model proposed by Sondhi and Schroeter [4]. The formant frequency, F_n^k, is then estimated by finding the peaks of the transfer function

$$H(\omega) = \frac{1}{A_{\mathrm{T}}(\omega) - C_{\mathrm{T}}(\omega)Z_{\mathrm{R}}(\omega)}, \tag{34}$$

where $Z_{\mathrm{R}}(\omega)$ is the radiation impedance at the lip opening. $A_{\mathrm{T}}(\omega)$ and $C_{\mathrm{T}}(\omega)$ are the elements of the transmission matrix for the whole vocal tract [4]. We also calculate the sensitivity function, $S(n, i)$, from the volume velocity and acoustic pressure at each tube section of the vocal tract, as given in Eqs. (2) through (5).

Next, we calculate the coefficients of the linear equations T_{np} given in Eq. (12) in Method I or T_{nm} given in Eq. (25) in Method II as well as the error coefficient z_n in Eq. (7). The optimal value of β_n in Method I or $\Delta\gamma_m$ in Method II is then determined by solving the simultaneous equations in Eq. (17) or Eq. (29). The area function is then updated as given in Eq. (8) or Eq. (33).

The index k is set to $k + 1$, and we repeat the procedure until the mean absolute formant error

$$E_F = \frac{1}{N}\sum_{n=1}^{N} |\hat{F}_n - F_n| \tag{35}$$

falls below a threshold condition such as $E_F < 1$ (Hz). In addition, the value of β_n or $\Delta\gamma_m$ is scaled using a factor common for the index, n or m, so that the maximum change of the cross-sectional area, $\Delta A(i)/A(i)$, is less than 10% in each iteration, because Eq. (1) holds for a small change in the cross-sectional area and formant frequency.

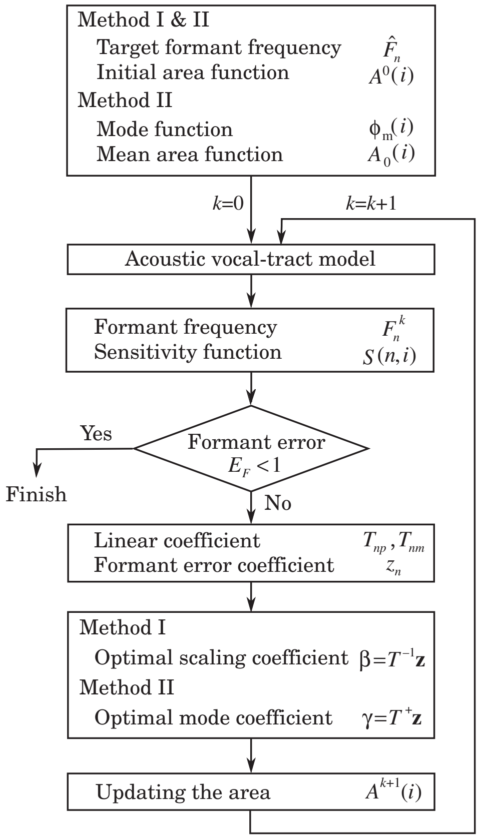
Method I & II
Target formant frequency $\hat{F}_n$
Initial area function $A^0(i)$
Method II
Mode function $\phi_m(i)$
Mean area function $A_0(i)$
k=0
k=k+1
Acoustic vocal-tract model
Formant frequency F_n^k
Sensitivity function $S(n,i)$
Yes
Formant error E_F <1
Finish
No
Linear coefficient T_{np}, T_{nm}
Formant error coefficient z_n
Method I
Optimal scaling coefficient $\beta = T^{-1}\mathbf{z}$
Method II
Optimal mode coefficient $\gamma = T^{+}\mathbf{z}$
Updating the area $A^{k+1}(i)$

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

3. ESTIMATION RESULTS

To evaluate the accuracy of the estimation methods quantitatively, we performed experiments using area function data of the ten English vowels, /i/, /ɪ/, /ɛ/, /æ/, /ʌ/, /ɑ/, /ɔ/, /o/, /ʊ/, and /u/, collected using a MRI measurement for a male native speaker [22]. Using the cross-sectional area data of each vowel and an acoustic tube model of the vocal tract [4], we first calculated the transfer function and resonance frequencies of the vocal tract. These resonance frequencies were then used as target frequencies, $\hat{F}_n$, and the cross-sectional area data as the right answer to the inverse estimation problem.

The mean cross-sectional area, $A_0(i)$, and the mode function, $\phi_m(i)$, were also computed from the area function data for these ten vowels using principal component analysis. The mean cross-sectional area was used as the initial value of the area function, $A^0(i)$, in each estimation. Depending on the estimation condition, the number of target formants was changed to $N = 2$, 3, and 4 in each

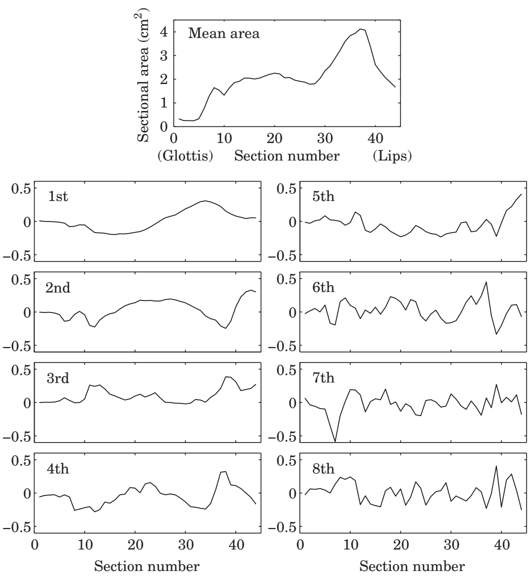

Fig. 2 Mean cross-sectional area and mode functions from first to eighth calculated using area function data of ten English vowels [22] and principal component analysis.

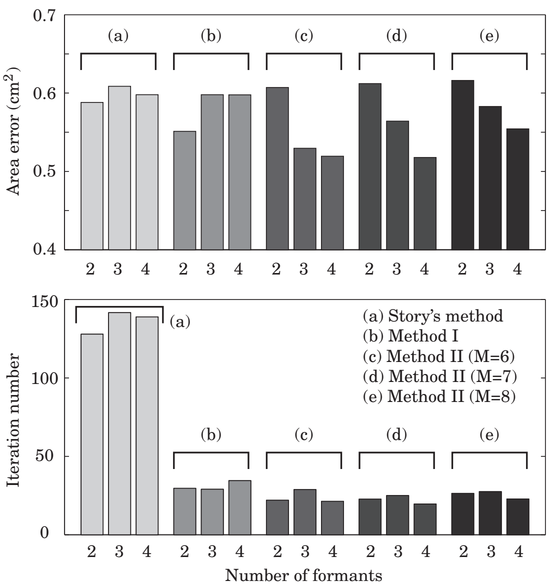

Fig. 3 Estimation error of the area function (top) and iteration number (bottom) as a function of the number of target formant frequencies, N. In each plot, (a) shows the result of Story's method, (b) is for Method I, and (c), (d), and (e) for Method II, where the number of mode functions, M, is 6 in (c), 7 in (d), and 8 in (e). Each plot shows the mean of the ten vowels.

estimation method and the number of mode functions to $M = 4$, 5, 6, 7, and 8 in Method II. Figure 2 shows the calculated mean cross-sectional area and mode functions as a function of the section number of the vocal tract. The number of tube sections, L, was 44 for all vowels. The first section corresponds to the glottis and 44th to the lips. The cumulative proportion increases as the number of mode functions increases; specifically, 97.2%, 98.7%, 99.2%, 99.7%, and 99.9% for M equal to 4, 5, 6, 7, and 8, respectively. We did not estimate the length of the vocal tract but used a measured value given in the literature [22].

3.1. Comparison of Estimation Methods

First, we compare the estimation accuracy of Story's method [18] and our Methods I and II. The top graph in Fig. 3 shows the estimation error of the area function averaged for the ten vowels as a function of the number of target formants, N. The target formants were selected in ascending order such that the frequencies for the first and second formants were constrained when $N = 2$, for example. The plots for (a) show the results from Story's method and (b) those in Method I. The plots (c), (d), and (e) show the results from Method II, where the number of mode functions, M, was set to 6 in (c), 7 in (d), and 8 in (e), respectively. The lower graph shows the mean number of iterations needed to satisfy the convergence criterion.

When N was 2, the estimation error in Method I was $0.55\,\text{cm}^2$ and smaller than the error of $0.58\,\text{cm}^2$ in Story's method. Furthermore, the iteration number significantly decreased from 128 in Story's method to 30 in Method I indicating the effectiveness of the least-squares technique. When N increased from 2 to 3 or 4, on the other hand, the estimation error increased unexpectedly in both methods. In addition to the number of targets, N specifies the number of sensitivity functions used to recover the area function [Eq. (8)]. The sensitivity function expresses the acoustical effect caused by a change in the cross-sectional area. In contrast to the mode function used in Method II, it does not necessarily reflect a physiologically relevant deformation pattern of the vocal tract. The actual reason for the increase in the error when N is 3 or 4 is not clear, but the sensitivity functions for the third and fourth formants, added to recover the area function, could have degraded the estimation accuracy.

Here, it is important to compare our estimation method with others and point out its fundamental feature. Inverse estimation methods such as those of Refs. 3, 5, 8–11, and 17 use a set of articulatory-acoustic data consisting of information about the vocal-tract shape or the state of the articulatory organs and corresponding vocal-tract spectrum or formant frequencies. The mapping relation between the articulatory and acoustic parameters is then trained from the data set. In general, the interrelation between both parameters would be more precise when the number of explanatory variables (formant frequencies, for example) in the mapping increases, and the estimation error then decreases. On the other hand, neither Story's method nor ours uses such an explicit relationship. The tendency of the estimation error in Fig. 3 may be partly due to this fundamental difference in the estimation principle.

In Method II, the estimation error decreased as the number of formants increased. Furthermore, the optimal value of the number of mode functions was found to be $M = 7$. When $N = 4$ and $M = 7$, the mean estimation error was $0.51\,\mathrm{cm}^2$. This error was the smallest among all the methods and estimation conditions on N and M. The mean iteration number was 19 and smaller than that in Method I. Therefore, we can conclude that Method II is superior to other methods in both estimation accuracy and computational efficiency.

Here, the result for the condition $M = 4$ is not shown, as the iterative procedure did not converge for a number of vowels when $N = 2$, 3, and 4. For $M = 5$, convergence was obtained for all vowels only for $N = 2$. The reason convergence was not obtained is possibly related to the above discussion about the property of the sensitivity function and mode function. The mode function is capable of effectively representing the deformation pattern of the vocal tract. For some vowels, however, the area function constructed by combining four or five mode functions was possibly acoustically insensitive to achieve the target formant frequencies. As a result, convergence was not obtained during optimization irrespective of the large number of iterations of more than a thousand.

The figure also shows that the estimation error tends to increase when the value of N is fixed to 2 or 3 and the value of M increases from 6 to 8. When N is 4, the estimation error for $M = 8$ was slightly larger than that for $M = 7$. Possibly this is because the estimation problem was more ambiguous when we used more mode functions to recover the area function from the same number of targets. Unlike Method I, Method II determines the value of unknown parameters, $\boldsymbol{\gamma}$, from Eq. (32) by using the pseudo-inverse of the coefficient matrix, T, because the number of parameters (M) is greater than the number of linear equations (N) that constrain the parameter values. We then determine the parameter values so that the norm of the parameter vector is the minimum under the linear constraints. When the number of target formants is 2 or 3, the number of linear constraints is also 2 or 3. On the other hand, the number of unknown parameters increases as the number of mode functions increases. Therefore, for higher M, the parameters have more unsolved degrees of freedom that are not constrained by the linear equations. The optimization problem then becomes more ambiguous, which can explain the tendency of the estimation error observed in Fig. 3.

3.2. Comparison of Vowels

Figure 4 shows the estimation error of the area function for each vowel, where the results from Story's method ($N = 2$) are plotted in (a), those from Method I ($N = 2$) in (b), and those from Method II ($N = 4$ and $M = 7$) in (c). The estimation conditions on N and M were determined so that the mean estimation error was the minimum in each method. The residual error when the actual area function was approximated by seven mode functions is also plotted in (c) for each vowel, represented by the white bar, and this approximation error was much smaller than the estimation error, represented by the black bar.

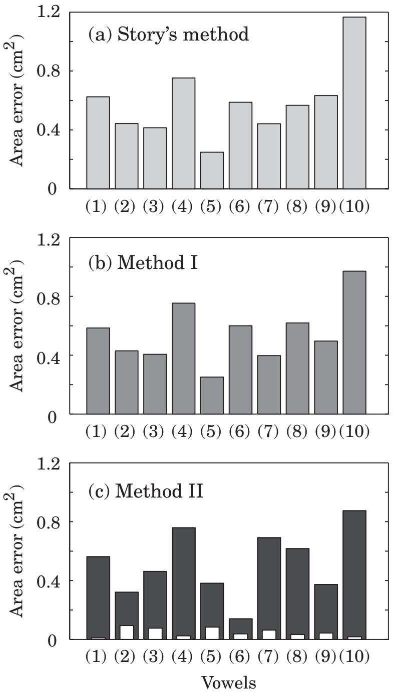

Fig. 4 Estimation error of the area function as a function of vowels (1) /i/, (2) /ɪ/, (3) /ɛ/, (4) /æ/, (5) /ʌ/, (6) /ɑ/, (7) /ɔ/, (8) /o/, (9) /ʊ/, and (10) /u/ for (a) Story's method ($N = 2$), (b) Method I ($N = 2$), and (c) Method II ($N = 4$, $M = 7$). The white bar in (C) represents the residual error when the actual area function is approximated by using seven mode functions.

The vowel-dependent patterns for the estimation error are quite similar between Story's method and Method I. The error was found to be minimum for /ʌ/ and maximum for /u/. For Method II, in contrast, the error was minimum for /ɑ/ and maximum for /u/. The difference between the maximum and minimum errors was $0.91\,\mathrm{cm}^2$ in Story's method, $0.71\,\mathrm{cm}^2$ in Method I, and $0.73\,\mathrm{cm}^2$ in Method II. The estimation error is therefore largely dependent of vowel type, when compared to the dependence on target number or the estimation method itself.

Figure 5 shows the area function estimated from the formants and frequency response of the vocal tract calculated from the area function as given in Eq. (34). The estimation condition was the same as that in Fig. 4 in each method. The area function estimated by Method I

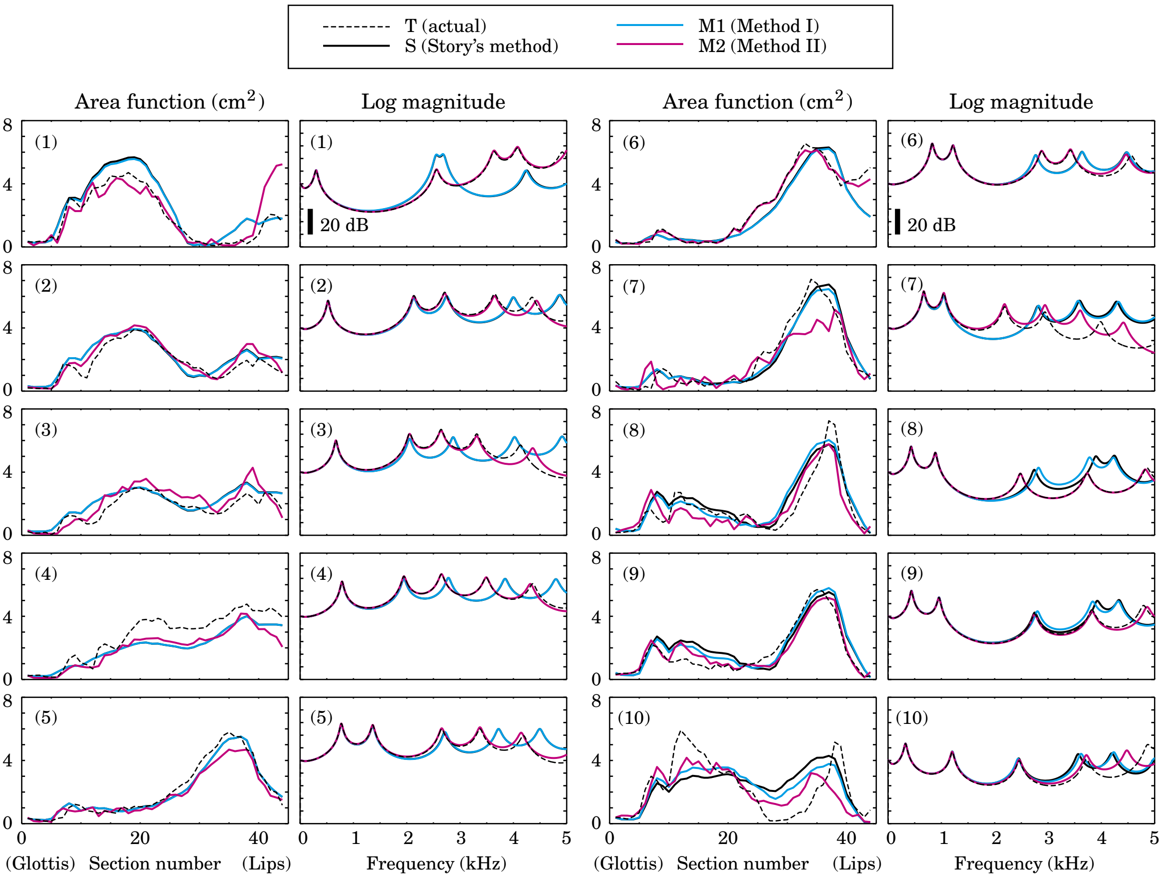

Fig. 5 Estimated area functions and frequency responses for (1) /i/, (2) /ɪ/, (3) /ɛ/, (4) /æ/, (5) /ʌ/, (6) /ɑ/, (7) /ɔ/, (8) /o/, (9) /ʊ/, and (10) /u/. Broken lines indicate the data based on the MRI measurement [22] from which the target formant frequencies were specified. Solid lines are the estimation results for Story's method ($N = 2$), Method I ($N = 2$), and Method II ($N = 4$, $M = 7$).

is very similar to the result of Story's method and therefore both plots are overlapped in the figure except for /o/, /ʊ/, and /u/. In Method II, the error is prominent for a specific region near the lip opening for /i/ and around the anterior part of the vocal tract for /ɔ/ and /o/. For /æ/ and /u/, the error is distributed over the whole vocal-tract region.

Except for /ɔ/ and /u/, the vocal-tract spectrum in Method II agrees well with that calculated from the actual area function over the frequency range up to 5 kHz. In Story's method and Method I, good agreement with the actual spectrum is obtained at least up to the second formant frequency, except for /i/.

3.3. Dependency on the Initial Value of the Area Function

Finally, we investigated the dependency of methods on the initial value of the area function. In addition to the mean of the area function data for all vowels, the mean of frontal vowels /i, ɪ, ɛ, æ/ and the mean of other vowels /ʌ, ɑ, ɔ, o, ʊ, u/ were used as the candidate for the initial value. When the target formant frequencies were specified, the best was selected so that the mean difference between the target frequency and the corresponding formant frequency calculated from each candidate area function was minimum. The initial area function might be closer to the actual one in this way and we expected an improvement in the estimation accuracy and convergence speed.

The estimation condition was set to $N = 2$ in Story's method and Method I, and $N = 4$ and $M = 7$ in Method II. As a result, the iteration number decreased on average from 128 to 121 in Story's method, from 30 to 25 in Method I, and from 19 to 14 in Method II. However, the estimation error remained the same or was even slightly worse: 0.60 cm^2 for Story's method, 0.57 cm^2 for Method I, and 0.48 cm^2 for Method II.

To investigate the influence of the initial value in more detail, it was set as

$$A^0(i) = (1 - w) \cdot A_0(i) + w \cdot A_T(i), \tag{36}$$

where w is the interpolation rate between the mean cross-sectional area, $A_0(i)$, and the actual cross-sectional area for the test vowel, $A_T(i)$, and $i = 1, 2, \ldots, L$. The initial value gradually approached the actual value as w increased from

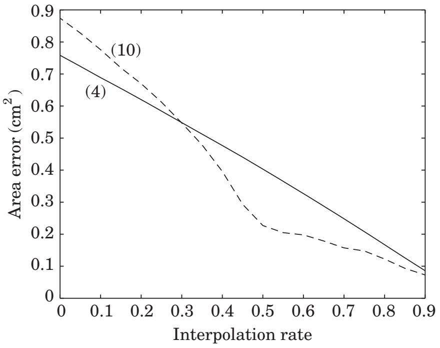

Fig. 6 Estimation error from Method II for (4) /æ/ and (10) /u/ as a function of the interpolation rate of the initial value. The estimation condition is $N = 4$ and $M = 7$.

0 to 0.9 in steps of 0.05. Figure 6 shows the estimation error for (4) /æ/ and (10) /u/, for which the error was relatively large among the vowels, in Method II ($N = 4$ and $M = 7$). The error decreased monotonically for each vowel as w increased, indicating that the estimated area function was slightly different for each value of w, although the frequencies of four formants were identical. The result implies the existence of a one-to-many relation in the inversion problem and possibly this was the chief cause of the large estimation error for these vowels.

4. SUMMARY AND DISCUSSION

We have presented two methods for estimating the vocal-tract cross-sectional area function from formants using the sensitivity function and least-squares optimization of the updating coefficient. In Method I, the sensitivity function was directly used as the basis function for updating the area function. In Method II, the area function is represented by summing linear orthogonal mode functions to take the actual deformation pattern of the vocal tract into account. The cost function was given as the error between the frequencies of target formants and those calculated from the area function in each iteration. Owing to the linearity of the updating equations for the area function, the values of the unknown weighting coefficients for the sensitivity functions in Method I and those for the mode functions in Method II were determined optimally so that the specified target formant frequencies were progressively achieved. Method performances were evaluated using the actual area function data for the ten English vowels [22], from which the initial value for the estimation procedure and mode functions were also determined.

The main results obtained by the evaluation tests are as follows.

(1) The estimation error was on average 0.55 cm^2 for Method I when two target formants were used. This error was slightly smaller than that in Story's method (0.58 cm^2). It was also shown that the iteration number decreased from 128 in Story's method to 30 in Method I, and the solution in Method I converged very quickly.

(2) The minimum of the estimation error, 0.51 cm^2, was obtained when four targets and seven mode functions were used in Method II. The iteration number was 19 on average and smaller than for Method I.

(3) The estimation accuracy significantly depended on the type of the vowel. The minimum-maximum range of the estimation error for ten vowels was 0.91 cm^2 from Story's method, 0.71 cm^2 from Method I, and 0.73 cm^2 from Method II.

(4) The estimation error did not decrease effectively when the number of candidates for the initial value of the area function increased from one to three.

The area function in general involves many sections required to model the vocal-tract shape with enough accuracy. The number of parameters representing the cross-sectional area and length of each section then becomes very large compared with, for example, the number of formants which convey linguistic information. A different shape of the vocal tract has been suggested as possibly producing similar acoustic characteristics such as the vocal-tract spectrum or resonance frequencies [5,12]. With respect to the non-uniqueness problem, numerical results presented in this paper showed that the use of the sensitivity function or mode function is to some extent 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 estimation error was prominent for some specific vowels. Our methods could not recover acoustically important features of the vocal tract such as the cross-sectional area near the vocal-tract constriction (typically for /u/) or the opening area of the lips (typically for /i/ in Method II). For /æ/, the estimation error was distributed over the entire vocal-tract region in both methods. From the numerical investigation, the existence of the one-to-many relation was suggested as a possible source for the prominent errors observed for these vowels.

Another problem was that the estimated area functions tended to give somewhat irregular results in Method II. The spatial pattern of the mode function is jagged along the vocal-tract sections for higher order modes, implying that such higher modes were overly weighted in the estimation process, and as a result, the smoothness of the area function was degraded. To overcome the problem, further investigation will be conducted to integrate a priori information about the variance of each mode (eigenvalue obtained in the principal component analysis) into the estimation process. In addition, drawbacks of the current method should be improved by estimating the length and cross-sectional area of the vocal tract simultaneously, and by taking the speaker dependence of empirical mode functions into account.

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

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

[5] 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).

[6] S. E. Levinson and C. E. Schmidt, "Adaptive computation of articulatory parameters from the speech signal," *J. Acoust. Soc. Am.*, **74**, 1145–1154 (1983).

[7] S. Panchapagesan and A. Alwan, "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model," *J. Acoust. Soc. Am.*, **129**, 2144–2162 (2011).

[8] 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).

[9] 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).

[10] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Trans. Speech Audio Process.*, **12**, 175–185 (2004).

[11] S. Hiroya and M. Honda, "Speaker adaptation method for acoustic-to-articulatory inversion using an HMM-based speech production model," *IEICE Trans. Inf. Syst.*, **E87-D**, 1071–1078 (2004).

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

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

[14] M. M. Sondhi and B. Gopinath, "Determination of vocal-tract shape from impulse response at the lips," *J. Acoust. Soc. Am.*, **49**, 1867–1873 (1971).

[15] T. Mochida and M. Honda, "Estimation of the vocal tract area function from the impulse response at the lips," *J. Acoust. Soc. Jpn.* (*J*), **55**, 147–155 (1999) (in Japanese).

[16] 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).

[17] 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).

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

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

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

[21] 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).

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

[23] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. (The Johns Hopkins University Press, Baltimore, 1996).