Vocal Tract Warping for Normalizing Inter-Speaker Differences in Vocal Tract Transfer Functions

Tatsuya Kitamura*, Hironori Takemoto† and Seiji Adachi‡
* Konan University, 8-9-1 Okamoto, Higashinada, Kobe, Hyogo 658-8501, Japan
E-mail: t-kitamu@konan-u.ac.jp Tel: +81-78-435-2535
† National Institute of Information and Communications Technology,
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan
E-mail: takemoto@nict.go.jp Tel: +81-774-95-2644
‡ Fraunhofer Institute for Building Physics, Nobelstrasse 12, 70569 Stuttgart, Germany
E-mail: seiji.adachi@ibp.fraunhofer.de Tel: +49-711-970-3437

***Abstract*—Vocal tract warping functions for normalizing vocal tract transfer functions of seven male subjects were calculated based on a vocal tract deformation method based on the vocal tract length sensitivity function. Vocal tract area functions for the five Japanese vowels of six subjects were tuned for their first four formant frequencies to be close to those of a target subject. The vocal tract warping functions were obtained as relationship between the original and deformed area functions. The results indicate that (1) the warping functions are not linear functions, (2) the vocal tract length of the deformed area functions are different from that of the target subject, and (3) the shape of the warping functions of the five vowels are not constant for each subject.**

I. Introduction

The shape of the vocal tract differs from person to person, and the differences cause the speaker individualities of speech sounds. The speaker individualities have been a major impediment to progress of speaker independent speech recognition. To overcome the speaker individualities, vocal tract length normalization have been studied (for example, [1]); however, the warping functions for vocal tract length normalization have not been calculated from actual vocal tract shape. In the present study, we thus estimate the warping functions from the vocal tract area functions measured from magnetic resonance imaging (MRI) data.

Yang and Kasuya[2] demonstrated that uniform and non-uniform normalization of the length of the vocal tract between male, female, and child subjects. In the uniform scaling, the vocal tract length were uniformly extended. In the non-uniform scaling, on the other hand, each length of the oral, pharyngeal, and laryngeal sections was normalized between the subject. In both method, the maximum cross-sectional area is also normalized. They reported that differences in the first two formant frequencies by the two methods, and concluded that the overall vocal tract length mainly contributes to the normalization of the vocal tract.

Recently, Adachi *et al*.[3] proposed a vocal tract deformation method based on the area and length sensitivity functions. These functions are equations for finding a change in formant frequency due to area and length perturbation of the vocal tract, respectively. Using the method, they demonstrated a male-female vocal tract shape conversion. In the present study, in order to obtain the vocal tract warping functions, we used only the length sensitivity function for deformation to normalize inter-speaker differences in vocal tract transfer functions.

This study was supported by SCOPE (071705001) of the Ministry of Internal Affairs and Communications, Japan, and Kakenhi (21300071, 21500184, and 21330170).

II. Materials

A. MRI Data

MRI data of seven Japanese male subjects AN, HT, KH, SA, SH, TI, and YT were obtained during production of the five Japanese vowels (/a/, /e/, /i/, /o/, and /u/) with a Shimadzu-Marconi ECLIPSE 1.5T Power Drive 250 at the ATR Brain Activity Imaging Center. Each subject was positioned to lie supine on the platform of the MRI unit. A head-neck coil was then positioned over the subject's head and neck region. The imaging sequence was a sagittal fast spin echo series with 2.0-mm slice thickness, no slice gap, no averaging, a 256×256-mm field of view, a 512×512-pixel image size, 41 or 51 slices, 90° flip angle, 11-ms echo time, and 3,000-ms repetition time.

The MRI data that show blur due to motion artifact were excluded from further analyses.

B. Vocal tract area functions

The teeth are imaged with low signal intensity as well as air by the MRI, and it is thus difficult to identify the boundary between them on MRI data. Volume data of the upper and lower jaws were measured and then superimposed on the MRI data by the method of Takemoto *et al*.[4] prior to measuring vocal tract area functions.

Cross-sectional areas of the vocal tract along its midline were measured at 2.5-mm intervals from the MRI data by the method of Takemoto *et al*.[5]. The bilateral piriform fossa cavities were excluded from the area functions in this study. In this study, a vocal tract area function is represented by a succession of truncated cones, not by a succession of cylindrical tubes. Figure 1 illustrates the extracted area functions and Table I lists the vocal tract length for the subjects.

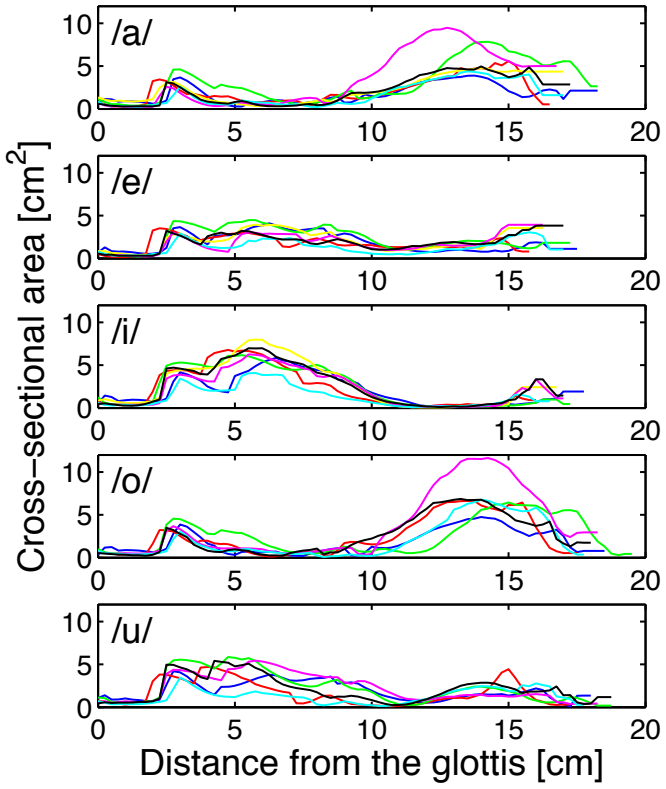

Fig. 1. Vocal tract area functions of the five Japanese vowels from seven subjects (blue line: AN, red line: HT, green line: KH, yellow line: SA, magenta line: SH, cyan line: TI, black line: YT).

TABLE I
VOCAL TRACT LENGTH OF THE FIVE JAPANESE VOWELS FROM SEVEN SUBJECTS IN CM.

Subject	Vowel				
	/a/	/e/	/i/	/o/	/u/
AN	18.25	17.50	17.75	18.50	18.00
HT	16.50	15.75	16.00	17.50	17.25
KH	18.25	17.25	17.25	19.50	18.75
SA	17.00	16.25	16.75		
SH	16.75	16.25	17.00	18.25	18.25
TI	17.00	17.00	16.50	17.75	17.75
YT	17.25	17.00	17.00	18.00	18.75
Mean	17.29	16.71	16.89	18.25	18.13

III. METHODS

A. Calculating vocal tract transfer functions

Calculation of velocity-to-velocity transfer functions of the vocal tract area functions was based on a transmission line model[6]. The transfer functions were calculated for the frequency region up to 5 kHz considering the radiation impedance at the mouth and assuming the glottal area is zero.

The radiation impedance of the vocal tract Z_R was approximated by the following equation suggested by Caussé *et al.*[7]:

$$\frac{Z_R}{\rho c} = \frac{z^2}{4} + 0.0127z^4 + 0.082z^4 \ln z - 0.023z^6 + j(0.6133z - 0.036z^3 + 0.034z^3 \ln z - 0.0187z^5), \quad (1)$$

$$z = kr, \quad (2)$$

where ρ is the air density, c is the speed of sound, k is the wave number, and r is the radius of the open end. We assumed $\rho = 1.15\mathrm{kg/m}^3$ and $c = 349.3.0$ m/sec. It should be noted that Eq. (1) is valid for a frequency region satisfying $kr < 1.5$.

In addition to the losses above, the model includes losses due to heat conduction, viscous friction, and vibration at the vocal tract wall.

B. Vocal tract length sensitivity function

The length sensitivity function is an equation for finding a change in formant frequency due to longitudinal perturbation of the vocal tract[3].

By assuming a planar wave propagation in the vocal tract, we can represent the vocal tract by area function $A(x)$, where x is the distance from the glottis. This planar wave is characterized by sound pressure $p(x,t)$ and volume velocity $U(x,t)$. Because the flow does not pass through the vocal tract wall, the radiation pressure on the vocal tract wall when the nth resonance mode of the vocal tract is generated is

$$P^{(n)}(x) = \mathrm{PE}^{(n)}(x) - \mathrm{KE}^{(n)}(x), \quad (3)$$

where $\mathrm{PE}^{(n)}(x)$ and $\mathrm{KE}^{(n)}(x)$ are the time averages of the potential energy density $\mathrm{PE}^{(n)}(x,t)$ and the kinetic energy density $\mathrm{KE}^{(n)}(x,t)$. These energy densities are defined as

$$\mathrm{PE}^{(n)}(x,t) = \frac{1}{2}\frac{1}{\rho c^2}p_n^2(x,t), \quad (4)$$

$$\mathrm{KE}^{(n)}(x,t) = \frac{1}{2}\rho\left(\frac{U_n(x,t)}{A(x)}\right)^2, \quad (5)$$

where $p_n(x,t)$ and $U_n(x,t)$ are the pressure and volume velocity when the nth mode is generated. The total energy of the nth mode E_n then can be calculated from the potential and kinetic energy densities:

$$E_n = \int_0^L \left\{\mathrm{PE}^{(n)}(x) + \mathrm{KE}^{(n)}(x)\right\} A(x)dx. \quad (6)$$

Next, we consider longitudinal deformation of the vocal tract[3]. The deformation can be expressed by letting the cross-sectional area at distance x be displaced along the length axis by $\delta x(x)$ ($\delta x(0)$ is set to zero). The local expansion or contraction ratio at x is denoted as $\Delta(x) \equiv \frac{\delta x(x)}{dx}$. A new distance is defined as $x' = x + \delta x(x)$ and the area function after the deformation is $A(x') \equiv A(x)$. In this case, the vocal tract length sensitivity function of the nth mode derived by Adachi *et al.*[3] is

$$S^{(n)}(x) = -\frac{\left\{\mathrm{PE}^{(n)}(x) + \mathrm{KE}^{(n)}(x)\right\} A(x)}{E_n}. \quad (7)$$

When we represent the area function as a succession of truncated cones or piecewise linear function, we can represent it as a set of nodes (x_s, A_s) for $s = 0, \ldots, N_s$, where s is the node index, x_s is the distance from the glottis, and A_s is the cross-sectional area. In this case, we have the length sensitivity function in discrete form:

$$S_s^{(n)} = -\frac{\Delta x_s}{2E_n}\left(\mathrm{PKE}_s^{(n)} A_s + \mathrm{PKE}_{s-1}^{(n)} A_{s-1}\right), \quad (8)$$

where $\mathrm{PKE}_s^{(n)} = \mathrm{PE}_s^{(n)} + \mathrm{KE}_s^{(n)}$, and N_s is the number of sections of the area function.

C. Calculating vocal tract warping functions

Adachi *et al*.[3] also proposed a vocal tract shape deformation method based on the area and length sensitivity functions. In this study, we used only the length sensitivity function for deformation in order to normalize the transfer functions only by local expansion or contraction of the area functions, and to obtain the warping functions.

Let the nth target formant frequency be T_n and that of a given vocal tract area function (x_s, A_s), be f_n. The difference between these formant frequencies normalized by f_n is given by $z_n = \frac{T_n - f_n}{f_n}$. The deformation is performed iteratively using the following update rule:

$$x_s^{new} = x_{s-1}^{new} + \Delta x_s \left(1 + \beta \sum_{n=1}^{N_f} z_n S_s^n\right) \quad \text{for } s = 1, \ldots, N_s \text{ with } x_0^{new} = x_0, \quad (9)$$

where N_f is the number of formants to tune and β is a coefficient to control the perturbation amplitude. We set N_f to 4 and β to 2.0.

To prevent the laryngeal tube from being overly elongated, we applied an additional rule at each iteration step:

$$x_s^{new} = \min\left\{x_{s-1}^{new} + \Delta x_{max}, x_s^{new}\right\} \quad (10)$$

where Δx_{max} was set to 6.5 mm in the present study.

The above iteration rules given in Eqs. (9) and (10) were applied until each z_n for $n = 1, \ldots, 4$ were less than 0.01 or the number of iteration reached 1,000. A vocal tract warping function is lastly obtained as correspondence relationship between x_s and x_s^{new}.

The target subject was AN, whose vocal tract length is closest to the mean of that for all the subjects, and the area function of the other subjects were deformed for the first four formant frequencies to be close to those of the target subject. It should be noted that the method does not guarantee the optimum deformation of area functions to move its formant frequencies closer to the target ones.

IV. Results and Discussions

The vocal tract transfer functions of the five Japanese vowels from the seven subjects are depicted in Fig. 2. The mean and standard deviation of the first four formant frequencies measured from the transfer functions are listed in Table II. The formants were identified by a peak-picking method. Individual

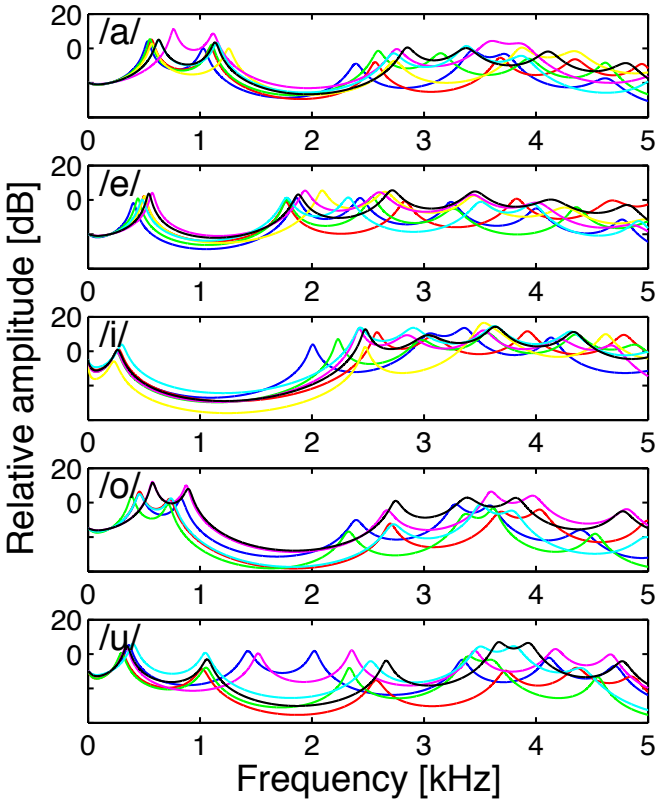

Fig. 2. Vocal tract transfer functions of the five Japanese vowels from seven subjects (blue line: AN, red line: HT, green line: KH, yellow line: SA, magenta line: SH, cyan line: TI, black line: YT).

TABLE II
Mean and standard deviation (SD) of the first, second, third, and fourth formant frequencies (F1, F2, F3, and F4) of vocal tract transfer functions of the five Japanese vowels from seven subjects in Hz.

	Vowel				
	/a/	/e/	/i/	/o/	/u/
F1 (Mean)	608	494	264	487	344
F1 (SD)	73	53	20	68	35
F2 (Mean)	1,127	1,870	2,374	788	1,189
F2 (SD)	64	109	179	72	204
F3 (Mean)	2,661	2,584	3,063	2,595	2,412
F3 (SD)	143	159	207	168	209
F4 (Mean)	3,501	3,454	3,753	3,473	3,517
F4 (SD)	221	179	388	141	140

variation in the transfer functions cannot be explained only by shift along the frequency axis, which is caused by differences in the vocal tract length between the subjects.

Figure 3 shows the vocal tract warping functions and Table III lists the norm of z_n. One or more of z_n $(n = 1, \ldots, 4)$ for the vowel /e/ of subjects HT, KH, SA, SH, and YT, and the vowel /u/ of subject SH were not less than the threshold 0.01 after the 1,000-iterations of the deformation of the area function.

The results showed that the warping functions are not linear functions. In addition, the vocal tract length after the deformation is different between the subjects and are different

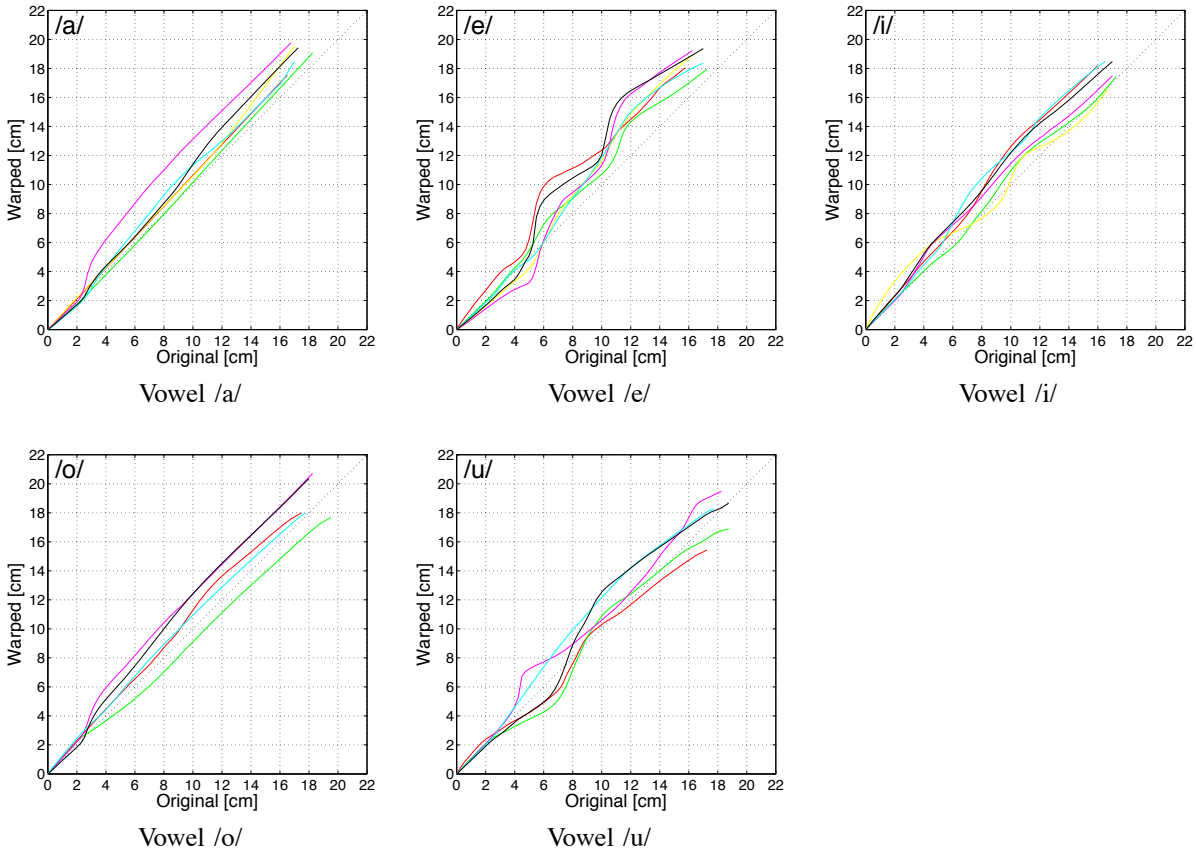

Fig. 3. Vocal tract warping functions for the five Japanese vowels for six subjects (red line: HT, green line: KH, yellow line: SA, magenta line: SH, cyan line: TI, black line: YT).

TABLE III
NORM OF z_n, THE DIFFERENCE BETWEEN TARGET (T_n) AND RESULTANT (f_n) FORMANT FREQUENCIES NORMALIZED BY f_n.

Subject	Vowel				
	/a/	/e/	/i/	/o/	/u/
HT	0.014	0.145	0.013	0.013	0.012
KH	0.014	0.028	0.014	0.013	0.013
SA	0.013	0.019	0.012		
SH	0.013	0.077	0.016	0.013	0.051
TI	0.012	0.014	0.013	0.014	0.013
YT	0.015	0.081	0.014	0.016	0.013

from that of the target subject AN. These results imply that individual differences in the vocal tract shape are the dominant factor for the individual variation in the formant frequencies rather than the differences in the vocal tract length.

The warping functions of each subject are different among the vowels, implying that it could be difficult to set a single warping function for each speaker in vocal tract length normalization methods.

V. CONCLUSIONS

In this study, the vocal tract warping functions were calculated by the vocal tract deformation method based on the vocal tract length sensitivity function[3]. The area functions of the six subjects were tuned by local expansion or contraction. The resultant warping functions were non-linear ones that are different between the subjects and vowels.

ACKNOWLEDGMENT

The MRI data analyzed in this study were measured as part of "Research on Human Communication" with funding from the National Institute of Information and Communications Technology, Japan.

REFERENCES

[1] Q. Lin and C. Che, Normalizing the vocal tract length form speaker independent speech recognition, *IEEE Signal Processing Letters*, 2, 201–203 (1995).
[2] C.-S. Yang and H. Kasuya, Uniform and non-uniform normalization of vocal tracts measured by MRI across male, female and child subjects, *IEICE Trans. Inf. & Syst.*, E78-D, 732–737 (1995).
[3] 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).
[4] H. Takemoto, T. Kitamura, H. Nishimoto and K. Honda, A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions, *Acoust. Sci. & Tech.*, 25, 468–474 (2004).
[5] H. Takemoto, K. Honda, S. Masaki, Y. Shimada and I. Fujimoto, Measurement of temporal changes in vocal tract area function from 3D cine-MRI data, *J. Acoust. Soc. Am.*, 119, 1037–1049 (2006).
[6] S. Adachi and M. Yamada, An acoustical study of sound production in biphonic singing Xöömij, *J. Acoust. Soc. Am.*, 105, 2920–2932 (1999).
[7] R. Caussé, J. Kergomard and X. Lurton, Input impedance of brass musical instruments – comparison between experiment and numerical models, *J. Acoust. Soc. Am.*, 75, 241–254 (1984).