Acoust. Sci & Tech **26**, 1 (2005)

PAPER

Individual variation of the hypopharyngeal cavities and its acoustic effects

Tatsuya Kitamura*, Kiyoshi Honda and Hironori Takemoto

ATR Human Information Science Laboratories,
2–2–2 Hikaridai, "Keihanna Science City," Kyoto, 619–0288 Japan

(*Received 23 January 2004, Accepted for publication 28 July 2004*)

Abstract: Morphological measurements of the hypopharynx are conducted to investigate the correlation between fine structures of the vocal tract and speaker characteristics. The hypopharynx includes the laryngeal tube and bilateral cavities of the piriform fossa. MRI data during sustained phonation of the five Japanese vowels by four subjects are obtained to analyze intra- and inter-speaker variation of the hypopharynx. Morphological analysis on the mid-sagittal and transverse planes revealed that the shape of the hypopharynx was relatively stable, regardless of vowel type, in contrast to relatively large inter-speaker variation, and these results are confirmed quantitatively by a simple similarity method. The small intra-speaker variation of the hypopharynx is confirmed by further morphological analysis using high-quality MRI data for one of the subjects, obtained by using the "phonation-synchronized method" and "custom laryngeal coil." Furthermore, acoustical effects of the individual variation of the hypopharynx are estimated by using a transmission line model. Vocal tract area function of one of the subjects above the hypopharynx is combined with the hypopharyngeal cavities of other subjects, and their transfer functions are calculated. The results show that the inter-speaker variation of the hypopharynx affects spectra in the frequency range beyond approximately 2 5 kHz

Keywords: Speaker characteristics, MRI, Vocal tract, Hypopharynx, Intra- and inter-speaker variation

PACS number: 43.70.Aj, 43.70.Gr [DOI. 10.1250/ast.26 16]

1. INTRODUCTION

Human speech does not only convey linguistic and paralinguistic information but also transmits nonlinguistic information. The latter can be better described as biological information because it delivers speaker characteristics arising from the individual variation in body geometry. Each speaker demonstrates unique characteristics in speech signals just as each person has a characteristic face, and both speech sounds and facial images facilitate communication before establishing verbal and expressive understanding Speaker characteristics also provide the framework for recognizing phonemes through information regarding age, gender, and physical conditions of the speaker. These facts provide evidence that speaker characteristics play a critical role in speech communication.

Speaker characteristics consist of two major elements: physical variations of speech organs and behavioral variabilities, such as talking styles or dialects. The former is thought to have two components: individual characteristics of the laryngeal source, and of the supralaryngeal vocal tract. While physical properties of the vocal folds mainly determine the gross difference in speaker age and gender, variations in vocal tract shape produce strong differences in speech spectra that distinguish one speaker from another.

The relationship between vocal tract shape in vowels and vowel spectra (e.g. formant frequencies) has been investigated by many researchers since Chiba and Kajiyama [1]. In contrast, the actual causal factors of speaker characteristics in the vocal tract remain unknown. Investigation into this topic is important not only for confirming results of psychoacoustical studies on perceptual clues of speaker characteristics but also for speech signal processing, such as voice quality control and speaker-independent speech recognition.

Imaging techniques such as magnetic resonance imaging (MRI) have made it possible to obtain the precise shape and area function of the vocal tract during phonation. A number of studies have shown differences in vocal tract shape across speakers by using those techniques. Baer *et al.* [2] employed the MRI technique to visualize three-dimensional vocal tract shape and to measure the area functions during sustained vowels for two speakers. Moore [3] also

*e-mail kitamura@atr.jp

used MRI to measure area functions for five speakers. Story *et al.* [4] and Saito *et al.* [5] compared area functions for vowels, and Narayanan *et al.* [6,7] and Alwan *et al.* [8] compared those for consonants among several speakers. While these studies have demonstrated the vowel-specific shapes of the vocal tract, they did not focus on the sources of speaker characteristics in the vocal tract.

Several studies have revealed the relationship between individual differences in the global shape of the vocal tract and speech spectra. Yang and Kasuya [9–12] described individual differences of vocal tract area functions extracted from MRI data using a boy, adult female and adult male speakers. They divided the entire vocal tract into three sections. the oral, the pharyngeal, and the laryngeal, and described geometrical differences in those sections. Based on the results, they proposed a normalization method for the area function by treating the sections separately Apostol *et al* [13] suggested a similar segmentation of area functions by dividing the vocal tract into four cavities: the laryngeal cavity, the back cavity, the constriction cavity, and the front cavity. They also proposed a speaker transformation method based on this segmentation. Honda *et al.* [14] compared geometrical variation and variation in vowel articulation by using an X-ray microbeam system. Honda [15] also showed a correlation between geometrical measures and the lower formant frequencies, and Fitch *et al.* [16] reported a high positive correlation between vocal tract length and body size. Because vocal tract length tends to correlate with formant frequencies, the variation in speakers' body size can signal speaker characteristics in speech. The above studies have provided evidence that the global shape of the vocal tract is one of the sources of individual speaker characteristics. However, there seem to be more factors in addition to the global shape of the vocal tract. In particular, the role of individual variations in the fine structures of the vocal tract for speaker characteristics has been underexplored.

The anatomical sources of speaker characteristics can be investigated for as components that show large inter-speaker variation and small intra-speaker (i.e., inter-phoneme) variation because such speaker characteristics extend beyond segmental or syllabic spans in speech. Thus, if there are regions that show large inter-speaker variation in shape and size with small intra-speaker variation in the vocal tract, it is possible that those regions are important factors for speaker characteristics. One region potentially satisfying these conditions is the lower part of the vocal tract that includes the hypopharyngeal cavities, i.e., the laryngeal tube and the piriform fossa. According to Dang and Honda [17], the area function of the piriform fossa is different among their subjects and relatively stable during sustained production of different vowels. Takemoto *et al.* [18] reported that the area function of the laryngeal vestibule was almost constant throughout a continuous utterance of vowels /aiueo/, using the data for a male subject obtained by their 3D cine-MRI technique. If the above results can be generalized for speakers and the shape and size of those cavities show large inter-speaker variation, then the cavities are possible regions that are responsible for speaker characteristics. It is well known that variation near the closed end of the vocal tract influences a wide frequency range in speech spectra [19]. Therefore, the inter-speaker variation of the lower part of the vocal tract can cause large variation in speech spectra. Considering these vocal tract characteristics, in the present study we focus on the hypopharynx and investigate intra- and inter-speaker variation of the geometry during sustained vowels We analyze MRI data obtained from four subjects selected for this purpose and carry out morphological analyses on the mid-sagittal and transverse planes. In addition, we obtain high-quality MRI data for one of the subjects to conduct a further analysis of intra-speaker variation of the vocal tract shape. Finally, we perform an acoustical simulation employing a transmission line model to investigate the acoustical effects of inter-speaker variation of the hypopharyngeal cavity.

2. STRUCTURE OF LOWER VOCAL TRACT

Figure 1 shows a sketch of the vocal tract and the three-dimensional shape of the lower part of the vocal tract. This part includes the laryngeal tube and the piriform fossa. The laryngeal tube is divided into the laryngeal ventricle and the laryngeal vestibule. The laryngeal ventricle is a cavity located between the true and false vocal folds, while the laryngeal vestibule is a narrow tube located superior of the laryngeal ventricle and opening into the main pharynx. The piriform fossa are a pair of bilateral cavities located behind the laryngeal tube. These two cavities are in the shape of an inverted cone and open into the main pharynx.

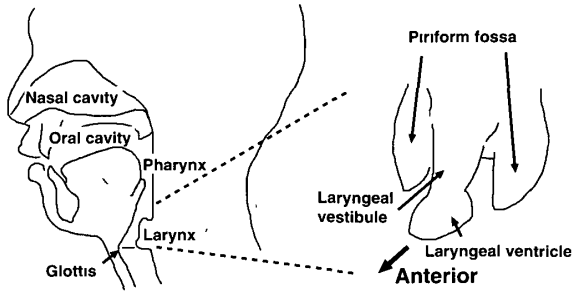


Fig. 1 Structure of the vocal tract on the mid-sagittal plane, and three-dimensional shape of the lower part of the vocal tract

3. MORPHOLOGICAL ANALYSIS OF VOCAL TRACT ON MID-SAGITTAL PLANE

3.1. MRI Acquisition

Magnetic resonance images of 11 Japanese male subjects producing the five Japanese vowels (/a/, /e/, /i/, /o/, and /u/) were obtained with a Shimadzu-Marconi ECLIPSE 1.5T Power Drive 250 at the ATR Brain Activity Imaging Center. Experienced radiologists conducted the examinations The imaging sequence was a sagittal Fast Spin Echo (FSE) series with 2.0-mm slice thickness, no slice gap, no averaging, a 256 × 256 mm field of view (FOV), a 512 × 512 pixel image size, 51 slices, 90° FA, 11-ms TE, and 3,000-ms TR. The total acquisition time was approximately 180 s.

Each subject was briefed on the experimental procedures prior to scanning, and each was positioned to lie supine on the platform of the MRI unit. A torso coil was then positioned over the subject's head and neck region. The subjects were instructed to maintain steady phonation and to breathe in gently through the mouth to hold the soft palate up during scanning. The subjects were not instructed on the pitch frequency of speech uttered during scanning

The sagittal MR images were transferred from the MRI system to a personal computer. The MR image quality was examined to exclude any data from the subjects that showed indistinct laryngeal structures caused by inhalation and/or swallowing saliva during scanning. Consequently, images of only four subjects were selected for further analysis.

3.2. Morphological Analysis

Outlines of the vocal tract for the five vowels were traced manually on the mid-sagittal slice and superimposed for each subject. The superimposed outlines shown in Fig. 2 indicate that the shapes of the lower part of the vocal tract show relatively small variation during production of the vowels while the glottis height changes, and these observations are consistent with the results of Takemoto *et al.* [18] mentioned above.

4. MORPHOLOGICAL ANALYSIS OF LOWER VOCAL TRACT ON TRANSVERSE PLANES

4.1. Cross-Sectional Shape Extraction

To clarify intra- and inter-speaker variation in the shape of the lower part of the vocal tract, we carried out a morphological analysis on transverse slices. The MR images mentioned in the previous section were interpolated using a trilinear algorithm to reconstruct volume data sets consisting of a cubic voxel of 0.5 × 0 5 × 0.5 mm. The transverse slices shown in Fig. 3 were extracted from the volume data sets for this analysis; the size of a transverse

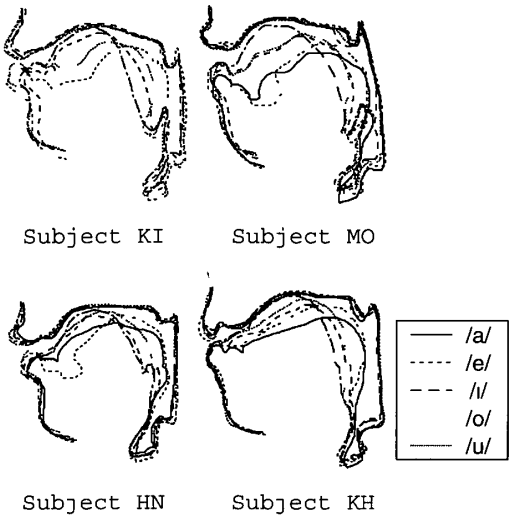


Fig. 2 Superimposed mid-sagittal vocal-tract outlines for /a/, /e/, /i/, /o/, and /u/ obtained from four male Japanese subjects referred to as KI, MO, HN, and KH The outline for /a/ of subject KI was excluded because the MR image for that vowel showed indistinct laryngeal structures

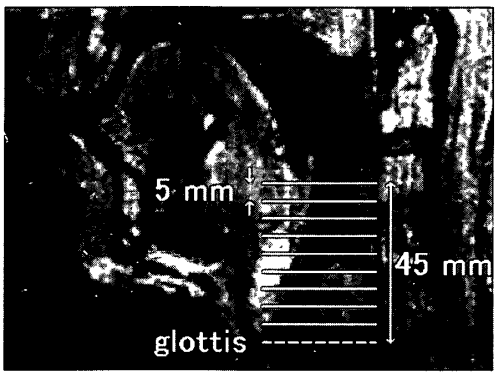


Fig. 3 The glottal position (dashed line) and positions of transverse slices extracted for morphological analysis (solid line)

slice is 512 × 512 × 201 pixels. The position of the glottis was determined visually for each vowel because it varied across vowels as shown in Fig. 2. The cross-sectional configuration of the vocal tract was then traced manually on each transverse slice.

Figure 4 illustrates tracings of the lower part of the vocal tract. Vocal tract configurations for a subject at equal distance from the glottis show similar shapes across vowels in an approximately 30-mm-long region extending from the glottis, which corresponds to the hypopharynx. The configurations of the bilateral cavities of the piriform fossa also show similar shapes regardless of vowel type, as reported in Dang and Honda [17]. These observations imply that the shape of the lower part of the vocal tract is relatively stable during vowel phonation. However, subject

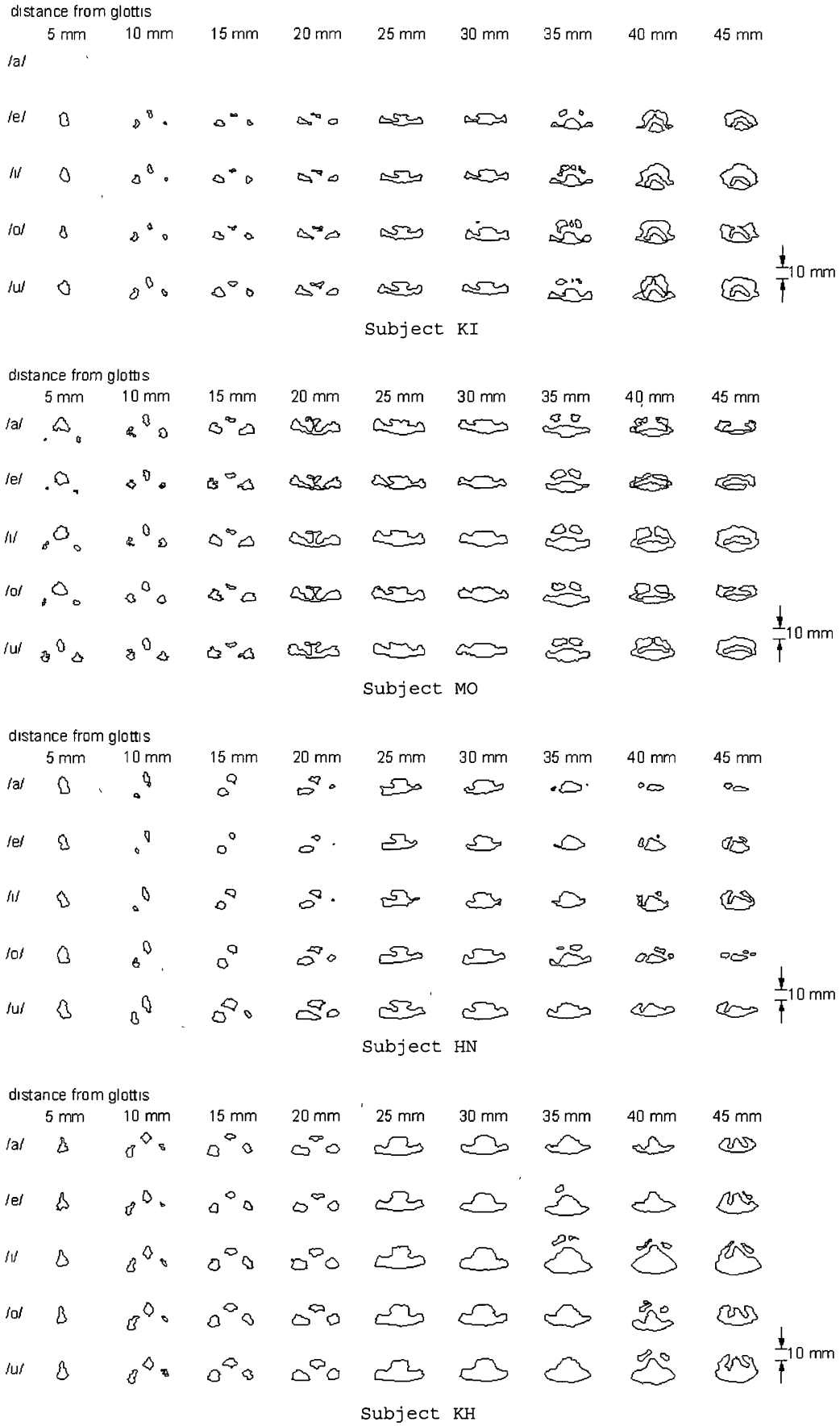


Fig. 4 Tracings on transverse slices of the vocal tract for four male Japanese subjects phonating the five Japanese vowels The tracing for /a/ of subject KI was excluded because the MR image showed indistinct laryngeal structures The pair of bilateral cavities of the vocal tract near the glottis is the piriform fossa

KI and HN showed an exceptionally larger area of the laryngeal tube for the vowel /u/ than that for the other vowels. In contrast to the small intra-speaker variation of those regions, the configurations and areas of the vocal tract show large inter-speaker variation. Piriform fossa configurations also display inter-speaker variation in shape, area, and depth of the cavities.

4.2. Intra- and Inter-Speaker Similarity Examination by Simple Similarity Method

Tracings of the lower part of the vocal tract are illustrated in Fig. 4. Below, we quantitatively examine the intra- and inter-speaker similarity among those tracings by employing a simple similarity method [20]

4.2.1 Simple similarity method

First, each tracing in Fig. 4 is converted into a binary image in which the pixel of the vocal tract region is set to 1 while that of the other area is set to 0. Next, the binary image is gradated and converted into a gray-scale image by a second-order Gaussian filtering. The standard deviation of the Gaussian filter is set to 10 pixels (5 mm). The simple similarity value S_s between two gray-scale images g_1 and g_2 is defined in the following formula [20]:

$$S_s = \frac{(g_1, g_2)^2}{||g_1||^2 ||g_2||^2}, \quad (1)$$

where (g_1, g_2) is the inner product of g_1 and g_2, and $||g||$ is the norm of g $(= \sqrt{(g, g)})$. The simple similarity value ranges from 0 to 1, and a larger simple similarity value between two images means the images are more similar. The above procedure to obtain the simple similarity value is schematized in Fig. 5.

The following two cases show examples of the simple similarity values computed between the cross-sectional shapes of the vocal tracts shown in Fig. 4. The simple similarity value between vowels /a/ and /e/ for subject KH on the transverse plane 10 mm from the glottis is 0.83. The simple similarity value between subject KH and subject HN for the vowel /a/ on the same transverse plane is 0.25.

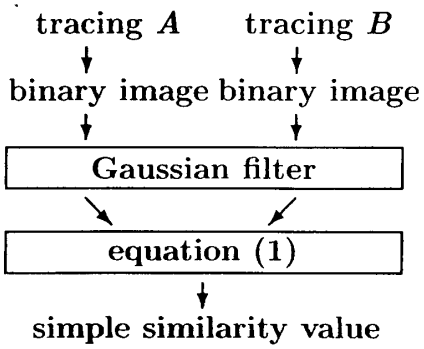


Fig. 5 Procedure to obtain simple similarity values between two tracings

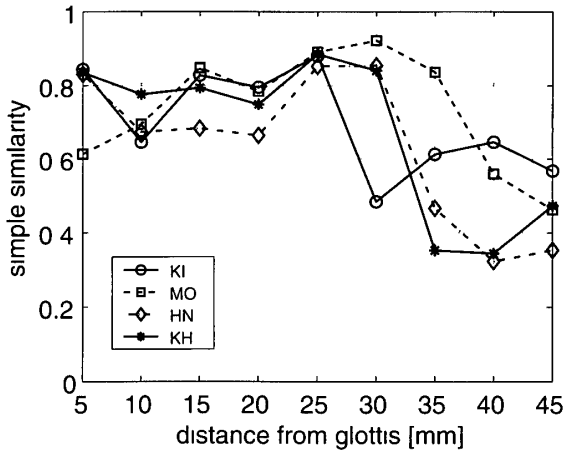


Fig. 6 Simple similarity values between the cross-sectional shapes of the vocal tract These are averaged with respect to the vowels

4.2.2. Intra- and inter-speaker simple similarity

Figure 6 shows the simple similarity values averaged with respect to the vowels, indicating intra-speaker similarity of the vocal tract shape The sections having large simple similarity values are regarded as having small variation in the shape of the vocal tract and vice versa. The intra-speaker simple similarity values in an approximately 30-mm-long region extending from the glottis, corresponding to the lower part of the vocal tract, are larger than those in the superior sections. Figure 7 shows the simple similarity values averaged with respect to the subjects. The inter-speaker simple similarity values are relatively smaller than the intra-speaker simple similarities over the same distance from the glottis, which indicates that the inter-speaker variation is larger than the intra-speaker variation. In addition, the inter-speaker simple similarity values on the transverse planes between 10 mm and 20 mm away from the glottis are small, suggesting that the shapes of the laryngeal

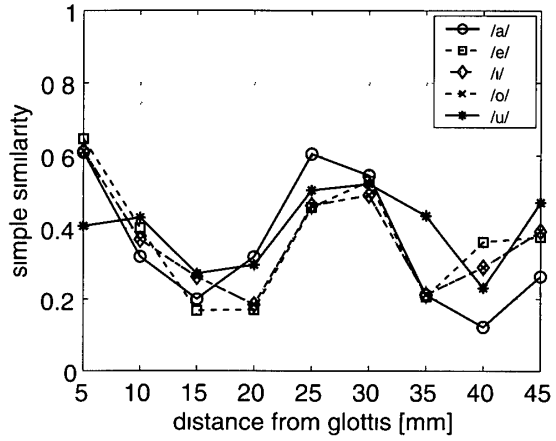


Fig. 7 Simple similarity values between the cross-sectional shapes of the vocal tract These are averaged with respect to the subjects

vestibule and the piriform fossa show particularly large inter-speaker variation Thus we can conclude that the shapes of the hypopharynx are relatively stable during vowel production and that the shapes show relatively large variation across subjects. The results are consistent with the qualitative analysis of similarity mentioned above

5. MORPHOLOGICAL ANALYSIS OF HYPOPHARYNX BY HIGH-QUALITY MRI

The quality of MR images is easily affected by a subject's body movements during scanning. The hypopharynx is prone to move when a subject breathes in and/or swallows saliva during sustained phonation. Movements of these apparatus result in a poor signal-to-noise ratio (SNR) of an MR image in that region To avoid this problem and obtain high-quality MR images of the hypopharynx, we employed the "phonation-synchronized method" and the "custom laryngeal coil" [21] We then further examined intra-speaker variation of the shape of the hypopharynx during vowel production.

5.1. High-Quality MRI Acquisition

High-quality MR images were obtained from subject KH while producing the five Japanese vowels using the same scanner In the "phonation-synchronized method," a subject is presented with a cyclic noise-burst sequence (4 beats on a 3-s cycle) in an MRI unit, and he/she breathes in and phonates repeatedly in exact timing with the noise-burst sequence. This technique allows scanning only during production and insures a high SNR in the image The "custom laryngeal coil" is a high-sensitivity coil with an elliptical antenna, and it can be placed close to the larynx to obtain high-resolution images of it [21]

The imaging sequence was a transverse RF-FAST series with 2.0-mm slice thickness, no slice gap, 2 averaging, a 128 × 128 mm field of view (FOV), a 512 × 512 pixel image size, 21 slices, 40° FA, 3,360-ms TE, and 390-ms TR. The total acquisition time was approximately 510 s — almost three times longer than that of the FSE series described in Chap. 3 The subjects were instructed to maintain a constant pitch frequency during scanning.

5.2. High-Quality MR Image on Transverse Planes

Figure 8 shows the high-quality MR images on transverse planes at distances from the glottis of $D = 6$ mm, 12 mm, 18 mm, and 24 mm during vowel production. The

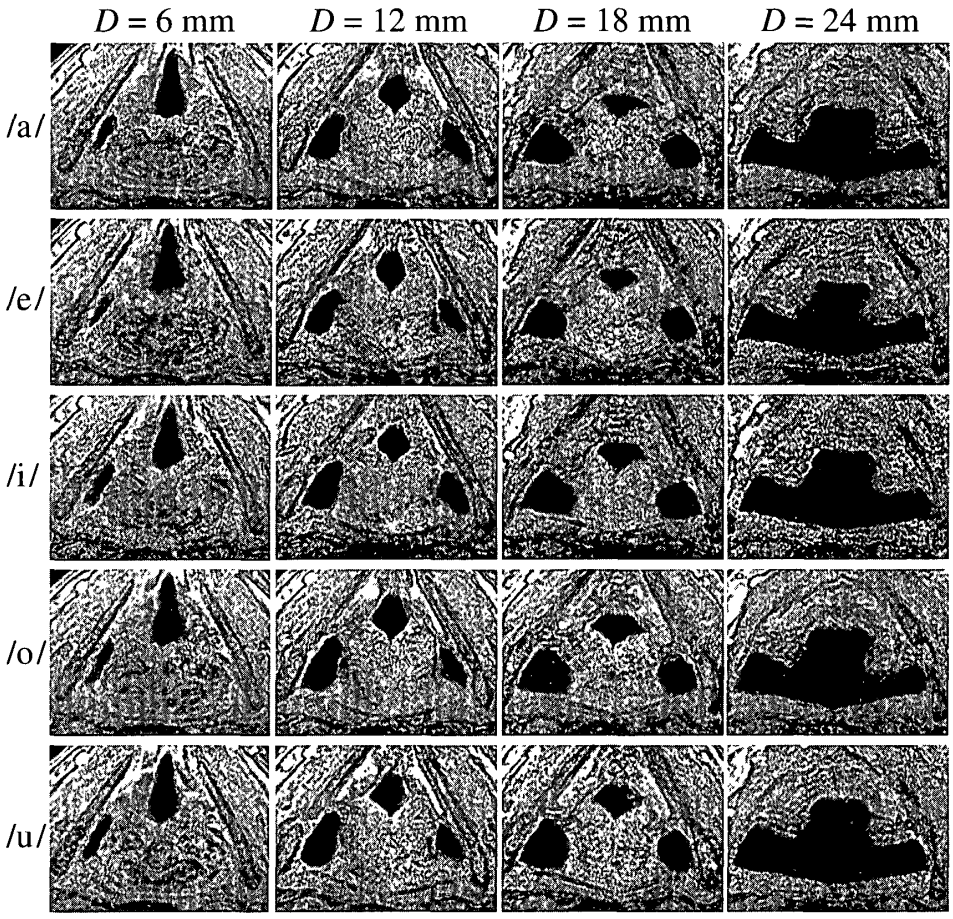


Fig. 8 High-quality MR images on transverse planes at distances from the glottis of $D = 6$ mm, 12 mm, 18 mm, and 24 mm for the five Japanese vowels These images were obtained by using the "phonation-synchronized method" and the "custom laryngeal coil" [21] The upper part of each image corresponds to the anterior part of the body

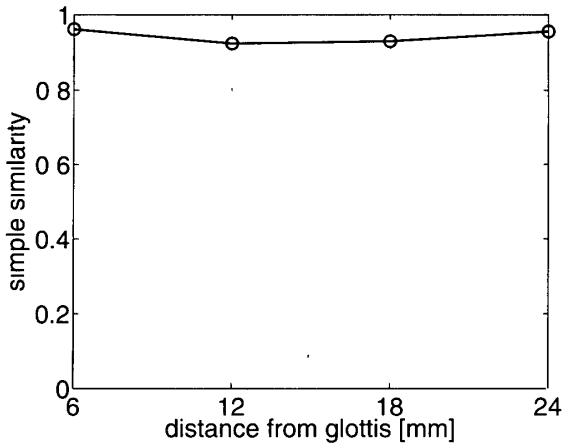


Fig. 9 Simple similarity values between the cross-sectional shapes of the vocal tract extracted from high-quality MR images These are averaged with respect to the vowels

black region in these images corresponds to air, that is, the vocal tract. Figure 9 shows the simple similarity values averaged with respect to the vowels, indicating intra-speaker similarity of the vocal tract shape. The standard deviation of the Gaussian filter was set to 20 pixels (5 mm) in the procedure to obtain simple similarity values. These figures show that cross-sectional configurations of the hypopharynx across the vowels are highly similar at equal distances from the glottis. Consequently, we can confirm that the shapes of the hypopharynx are relatively stable during production of the vowels.

6. SIMULATION USING TRANSMISSION LINE MODEL

In the previous sections, we indicated that the shapes of the hypopharynx are relatively stable across vowels, while they show a large inter-speaker variation. In this section, we estimate the acoustical effects of the individual variation in the hypopharynx by using a transmission line model To perform the simulation, a vocal tract area function of subject KH above the hypopharynx was combined with the hypopharyngeal cavities of other subjects, and their transfer functions were calculated.

6.1. Extraction of Vocal Tract Area Functions

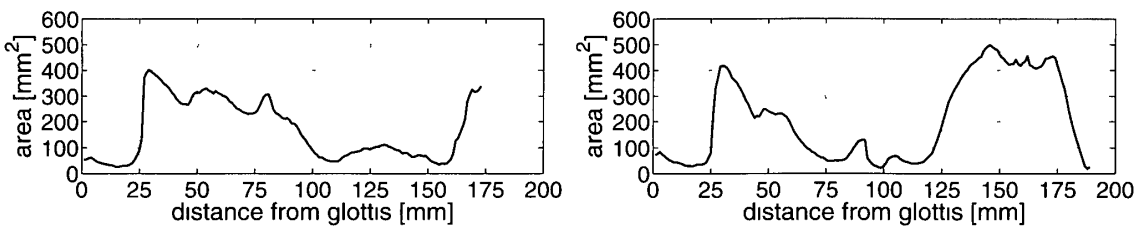
Area functions were extracted from the reconstructed volume data sets described in Chap. 4. One front vowel /e/ and one back vowel /o/ were selected for the simulation. First, volume data of the upper and lower jaws were superimposed onto the volume data using the method proposed by Takemoto *et al.* [22,23]. This procedure for dental visualization is indispensable in extracting precise vocal tract area functions. Next, vocal tract area functions were extracted from the volume data. The area functions of the main vocal tract and those of the piriform fossa were extracted separately: First, the mid-line of the vocal tract from the glottis to the lips was calculated on the mid-sagittal slice. Then, cross-sectional areas of the main vocal tract along the mid-line were measured at 1-mm intervals. The area functions of the piriform fossa were measured on transverse planes from the opening of the fossa into the pharynx to the bottom of the fossa, also at 1-mm intervals.

Figure 10 depicts the area functions of the vocal tract of subject KH. Figures 11 and 12 depict the area functions of the laryngeal tube and the piriform fossa for the same vowels. Figures 11 and 12 show that there are individual variations in the area functions of those cavities as well as differences between the two vowels in area functions of those cavities, indicating that the shape of the hypopharynx changes slightly among the vowels.

6.2. Method

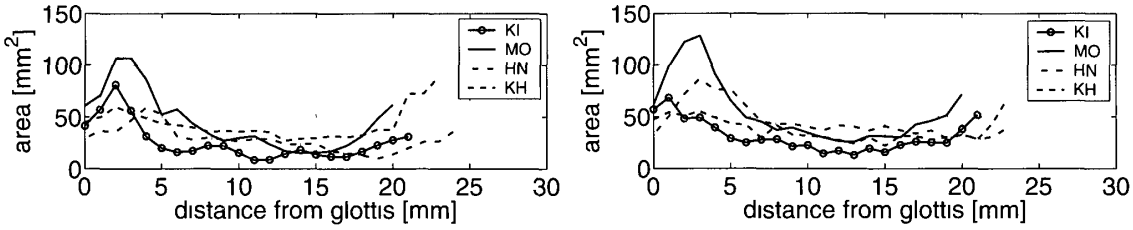
We constructed hypopharyngeal models based on the vocal tract area functions of the four subjects, and combined the areas of each hypopharyngeal model with the models of the oral and pharyngeal cavities of subject KH. Thus, the models were associated with their respective areas only for the hypopharynx, while only the models of subject KH were associated with his respective areas for the entire vocal tract. The models for the vowel /e/ produced by each subject are denoted below as `KIe`, `MOe`, `HNe`, and `KHe`, respectively. Similarly, the models for the vowel /o/ are denoted as `KIo`, `MOo`, `HNo`, and `KHo`.

Calculation of velocity-to-velocity transfer functions of the models was based on a transmission line model [24] for a frequency region up to 6 kHz. Based on the transmission



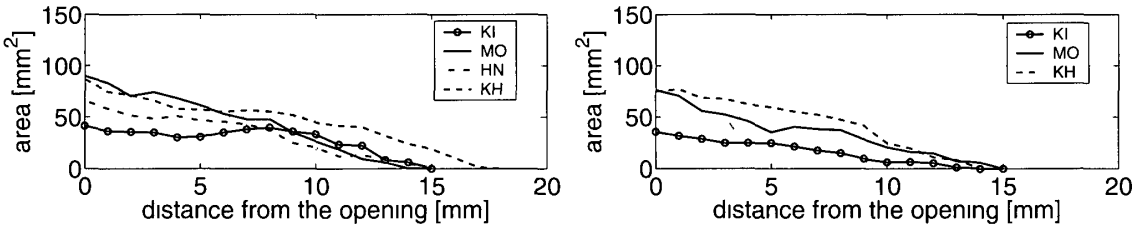
(a) Vowel /e/ (b) Vowel /o/

Fig. 10 Area functions of the vocal tract (a) the vowel /e/ and (b) the vowel /o/ of subject KH



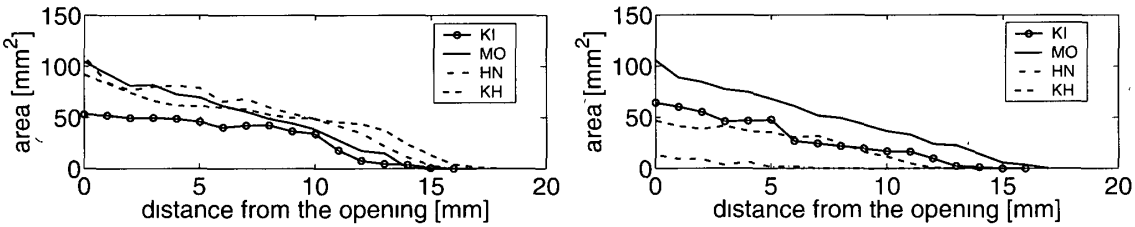
(a) Vowel /e/ (b) Vowel /o/

Fig. 11 Area functions of the laryngeal tube of (a) the vowel /e/ and (b) the vowel /o/



(a) Left cavity of piriform fossa (b) Right cavity of piriform fossa

Vowel /e/



(c) Left cavity of piriform fossa (d) Right cavity of piriform fossa

Vowel /o/

Fig. 12 Area functions of the piriform fossa (a) the left cavity and (b) the right cavity for the vowel /o/, (c) the left cavity and (d) the right cavity for the vowel /o/ The right cavity of the piriform fossa was not observed in the vowel /e/ of subject HN as shown in Fig 4

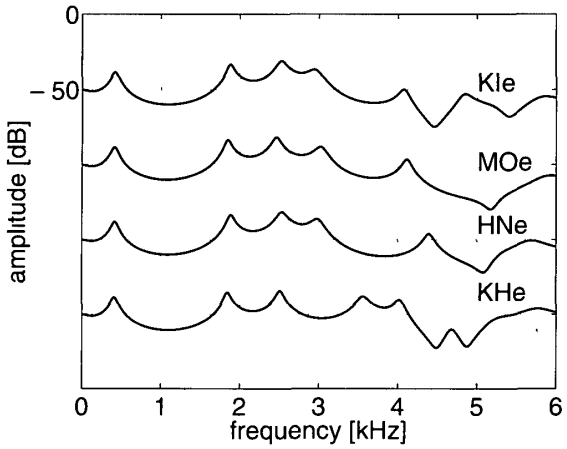
line model, a vocal tract is modeled as a cylindrical tube where plane wave propagation is assumed. This assumption is valid for a frequency region below an upper-limit frequency. This upper-limit frequency is determined by the largest diameter of the cylindrical tube, which is equal to the half-wavelength of the highest allowed frequency When we set the speed of sound $c = 353.0\,\mathrm{m/s}$, the largest area for which plane wave propagation can be assumed at 6 kHz is $(353000/2\ \ 2 \cdot 6000)^2 \cdot \pi = 680\,\mathrm{mm}^2$. The vocal tract area function of the vowel /e/ and the vowel /o/ shown in Fig. 10 are smaller than $680\,\mathrm{mm}^2$ along their entire lengths. It is therefore reasonable to calculate transfer functions of the models for a frequency region up to 6 kHz.

A cavity of the piriform fossa comprises a side branch of the vocal tract and is therefore modeled by two cascaded portions· the lower conical portion and the upper cylindrical portion as proposed by Dang and Honda [17] The radiation impedance of the vocal tract Z_R was approximated by the following equation suggested by Caussé *et al.* [25]:

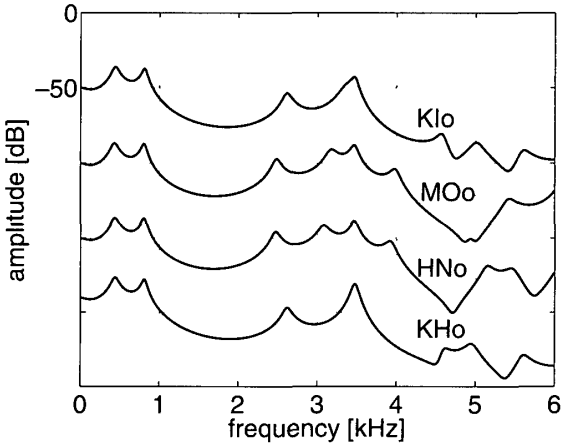
$$Z_R/\rho c = z^2/4 + 0.0127z^4 + 0.082z^4 \ln z - 0.023z^6 + j(0.6133z - 0.036z^3 + 0\ 034z^3 \ln z - 0.0187z^5), \quad (2)$$

$$z = kr. \quad (3)$$

Here, ρ is the air density, k is the wave number, and r is the radius of the radiating end. We set $\rho = 1.14\,\mathrm{kg/m}^3$ and $c = 353.0\,\mathrm{m/s}$. It should be noted that Eq. (3) is valid for a frequency region satisfying $kr < 1.5$. As the radius of the open end of the models for the vowel /e/ is 10 mm, so Eq. (3) is valid for up to 8.4 kHz In the same way, that of the models for the vowel /o/ is 3 mm, and Eq. (3) is valid for up to 28 1 kHz.



(a) Hypopharyngeal models for /e/



(b) Hypopharyngeal models for /o/

Fig. 13 Velocity-to-velocity transfer function computed using the hypopharyngeal models (a) the models for /e/ and (b) the models for /o/

6.3. Results

Velocity-to-velocity transfer functions of the hypopharyngeal models are shown in Fig. 13, and the formant frequencies from the first to the fifth are listed in Table 1. The differences across the formant frequencies of the four models for /e/ were 8 Hz (1.9%) for F_1, 39 Hz (2.1%) for F_2, 71 Hz (2.8%) for F_3, 619 Hz (19.8%) for F_4, and 321 Hz (7.7%) for F_5. The differences of the formant frequencies of the models for /o/ were 12 Hz (2.7%) for F_1, 6 Hz (0.7%) for F_2, 149 Hz (5.9%) for F_3, 386 Hz (11.7%) for F_4, and 1,102 Hz (28.8%) for F_5. These results indicate that the inter-speaker variation in the shape of the hypopharynx exerts influence over a wide frequency range above approximately 2.5 kHz. Dang and Honda [17] demonstrated that the piriform fossa cause antiresonance in speech spectra at the frequency region from 4 to 5 kHz, which results in a shift of the formant frequencies over a broad frequency region. Fant [19] and Fant and Båvegård [26] reported in their theoretical studies that the presence of the piriform fossa in their parametric model of vocal tract area functions leads to a shift in the formant frequencies over a wide frequency region. They also showed that varying the length of the laryngeal tube affects the frequencies of F_4 and F_5. Sundberg [27] demonstrated that the shapes of the pharyngeal and the hypopharyngeal cavities are different between speech and singing and that these cavities cause the "singing formant," which is a high spectrum envelope peak near 2.8 kHz in the male singing voice. Recently, Takemoto *et al.* [28] showed that varying the areas of the laryngeal tube in a vocal tract area function can have a strong effect on the frequencies of F_3 and F_4. Our results agreed with those studies.

Table 1 Formant frequencies of the velocity-to-velocity transfer function (a) the models for /e/ and (b) the models for /o/ Frequencies are expressed in Hz The frequency of F_5 for KHo is excluded because the peak of that formant is not distinguished

(a) Models for /e/

model	F_1	F_2	F_3	F_4	F_5
KIe	421	1,882	2,529	2,942	4,070
MOe	414	1,848	2,461	3,018	4,110
HNe	417	1,886	2,532	2,975	4,391
KHe	413	1,847	2,511	3,561	4,019

(b) Models for /o/

model	F_1	F_2	F_3	F_4	F_5
KIo	442	810	2,615	3,465	4,564
MOo	430	804	2,484	3,180	3,462
HNo	439	807	2,470	3,079	3,467
KHo	438	808	2,619	3,472	—

Moreover, the differences in F_3 and F_4 between /e/ and /o/ were relatively small except for F_4 of subject KI. Therefore, we can speculate that the small variance in the formant frequencies originates in the hypopharynx Taking into account the result of Takemoto *et al.* [28] mentioned above, it is natural to conclude that this region is relatively stable regardless of vowel type. For the same reason, it is also reasonable to state that F_3 and F_4 are stable enough to signal individual characteristics.

Antiresonance frequencies caused by the piriform fossa at around 5 kHz in the transfer functions are different between the vowels One possible cause may be the difference in the area functions of the piriform fossa between the vowels, as shown in Fig. 12

7. DISCUSSION

The results lead to the following conclusions

(1) The shapes of the hypopharynx, i.e., the laryngeal tube and the piriform fossa, are relatively invariant during production of the vowels.

(2) On the contrary, the shapes of those cavities show a wide inter-speaker variation. The inter-speaker varia-

tion of those cavities affects spectra at the frequency range above approximately 2.5 kHz.

The results of this study are in agreement with those of psychoacoustic studies. Takahashi and Yamamoto [29] in their study of Japanese vowels reported that spectral regions containing speaker characteristics occupy up to 4 kHz for /a/, up to 4.5 kHz for /e/, /i/, and /u/, and up to 2 2 kHz for /o/. Furui and Akagi [30] showed that speaker individualities exist mainly in the frequency range from 2.5 to 3 5 kHz, within the range of telephony Kitamura and Akagi [31,32] also reported that speaker individualities exist mainly in the frequency range above the peak near 20 ERBs (1,740 Hz). The frequency range shown above is in agreement with the range where the acoustical effect of the lower part of the vocal tract is most prominent. Consequently, it can be concluded that the hypopharynx is one source of speaker individualities.

The results obtained in this study complement those of studies showing the correlation between the global shape of the vocal tract and the lower formant frequencies, as suggested by Yang and Kasuya [9–12] and Honda [15] Summarizing their results and our conclusion, we can say that the global shape of the vocal tract provides speaker individualities in the lower frequency region of speech spectra, while the hypopharynx provides those in the higher frequency region.

The nasal and paranasal cavities are also stable during speech production, and their shape and size show a large individual variation [33] Since these cavities constantly contribute nasal spectra to vowels, they can also provide sources of individual characteristics of speech [34]

8. CONCLUSIONS

In order to explore possible sources of speaker individualities in speech sounds, this study investigated inter-speaker and intra-speaker variation in the geometry of the vocal tract using means of MRI. Focusing on the hypopharynx, we carried out morphological analysis of intra- and inter-speaker variation of the vocal tract shape during sustained vowel production. In addition, we estimated the acoustical effects of the cavities by model simulations for the four male subjects. Our results indicate that the hypopharynx shows relatively small intra-speaker variation and relatively large inter-speaker variation. The steady shapes of the cavities were confirmed by a high-quality MRI technique Furthermore, the results from the acoustical simulation showed evidence that the inter-speaker variation of the hypopharynx is reflected by the individual variation of the transfer function of the vocal tract

There are some problems due to limitations of the MRI technique Because the subjects uttered vowels while lying in the supine position in the MRI scanner as the MR images were obtained, the movement of the hypopharyngeal shape may have been limited due to the body posture. In addition, MRI scanning required sustained phonation over a few minutes. The vocal tract shape during sustained phonation is somewhat different from that during natural continuous utterance [35]. These problems may have caused the small intra-speaker variation of the hypopharynx. Further efforts will be needed to clarify and overcome these problems.

ACKNOWLEDGEMENTS

This research was conducted as part of "Research on Human Communication" with funding from the National Institute of Information and Communications Technology. We thank Dr. Yasuhiro Aoki of Toshiba Social Network and Infrastructure Systems Company for his advice on the simple similarity method. We also wish to thank Dr. Jianwu Dang of JAIST/ATR Human Information Science Laboratories and Mr. Hiroyuki Hirai of SANYO Electric Co , Ltd. for their helpful comments

REFERENCES

[1] T Chiba and M Kajiyama, *The Vowel Its Nature and Structure* (Tokyo-Kaiseikan, Tokyo, 1942)

[2] T Baer, J C Gore, L C Gracco and P W Bye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging Vowels," *J Acoust Soc Am* , **90** 799–828 (1991)

[3] C A Moore, "The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images," *J Speech Hear Res* , **35**, 1009–1023 (1992)

[4] B H Story, I R Titze and E A Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J Acoust Soc Am* , **100**, 537–554 (1996)

[5] H Saito, N Suzuki, Y Fujita, K Michi and T Takahashi, "3-dimensional measurements of vocal tract shape using MRI — methodology and area function in normal subjects —," *J Jpn Stomatol Soc* , **49**, 92–101 (2000)

[6] S S Narayanan, A A Alwan and K Haker, "An articulatory study of fricative consonants using magnetic resonance imaging," *J Acoust Soc Am* , **98**, 1325–1347 (1995)

[7] S S Narayanan, A A Alwan and K Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data Part I the laterals," *J Acoust Soc Am* , **101**, 1064–1077 (1997)

[8] A Alwan, S Narayanan and K Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data Part II the rhotics," *J Acoust Soc Am* , **101**, 1078–1089 (1997)

[9] C -S Yang and H Kasuya, "Dimensional differences in the vocal tract shapes measured from MR images across boy, female and male subjects," *J Acoust Soc Jpn (E)*, **16**, 41–44 (1995)

[10] 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)

[11] C -S Yang and H Kasuya, "Speaker individualities of vocal tract shapes of Japanese vowels measured by magnetic resonance images," *Proc ICSLP 96*, pp 949–952 (1996)

[12] C -S Yang and H Kasuya, "Invariance and individuality of the vowel Evidence from articulatory and acoustic observa-

tions," *Tech Rep IEICE*, SP96-120 (1997)
[13] L Apostol, P Perrier, M Raybaudi and C Segebarth, "3D geometry of the vocal tract and inter-speaker variability," *Proc ICPhS 99*, San Francisco, Vol 1, pp 443–446 (1999)
[14] K Honda, S Maeda, M Hashi, J S Dembowski and J R Westbury, "Human palate and related structures Their articulatory consequences," *Proc ICSLP 96*, pp 784–787 (1996)
[15] K Honda, "Individuality of orofacial form reflected in vowel spaces," *Proc Spring Meet Acoust Soc Jpn*, pp 237–238 (1997)
[16] W T Fitch and J Giedd, "Morphology and development of the human vocal tract A study using magnetic resonance imaging," *J Acoust Soc Am*, **106**, 1511–1522 (1999)
[17] J Dang and K Honda, "Acoustic characteristics of the piriform fossa in models and humans," *J Acoust Soc Am*, **101**, 456–465 (1997)
[18] H Takemoto, K Honda, S Masaki, Y Shimada and I Fujimoto, "Measurement of temporal changes in vocal tract area function during a continuous vowel sequence using a 3D cine-MRI technique," *Proc 6th Int Semin Speech Production*, Sydney, pp 284–289 (2003)
[19] G Fant, *Acoustic Theory of Speech Production* (Mouton, the Hague/Paris, 1960)
[20] T Iijima, *Theory of Pattern Recognition* (Morikita Shuppan, Tokyo, 1989)
[21] S Takano, K Honda, S Masaki, Y Shimada and I Fujimoto, "High-resolution imaging of vocal gesture using a laryngeal MRI coil and a synchronized imaging method with external triggering," *Proc Spring Meet Acoust Soc Jpn*, pp 291–292 (2003)
[22] H Takemoto, T Kitamura, H Nishimoto and K Honda, "Teeth filling method for MRI measurement of the vocal tract shape," *Proc Spring Meet Acoust Soc Jpn*, pp 293–294 (2003)
[23] H Takemoto, T Kitamura, H Nishimoto and K Honda, "A method of tooth superimposition of MRI data for accurate measurement of vocal tract shape and dimensions," *Acoust Sci & Tech*, **25**, 468–474 (2004)
[24] J L Flanagan, *Speech analysis synthesis and perception 2nd Edition* (Springer-Verlag, Berlin/Heidelberg/New York, 1972)
[25] 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)
[26] G Fant and M Båvegård, "Parametric model of VT area functions vowels and consonants," *TMH-Q Prog Status Rep*, **31**, 1–20 (1997)
[27] J. Sundberg, "Articulatory interpretation of the "singing formant"," *J Acoust Soc Am*, **55**, 838–844 (1974)
[28] H Takemoto, K Honda, S Masaki, Y Shimada and I Fujimoto, "Modeling of the inferior part of the vocal tract based on analysis of 3D cine-MRI data," *Proc Autumn Meet Acoust Soc Jpn*, pp 281–282 (2003)
[29] M Takahashi and G Yamamoto, "On the physical characteristics of Japanese vowels," *Res Electrotech Lab*, **326** (1931)
[30] S Furui and M Akagi, "Perception of voice individuality and physical correlates," *Trans Tech Com Psychol Physiol Acoust*, H85-18 (1985)
[31] T Kitamura and M Akagi, "Relationship between physical characteristics and speaker individualities in speech spectral envelopes," *J Acoust Soc Am*, **100**, 2600 (1996)
[32] T Kitamura and M Akagi, "Significant cues in spectral envelope of isolated vowels for speaker identification," *J Acoust Soc Jpn. (J)*, **53**, 185–191 (1997)
[33] J Dang and K Honda, "Morphological and acoustical analysis of the nasal and the paranasal cavities," *J Acoust Soc Am*, **96**, 2088–2100 (1994)
[34] K Honda, "Facial appearance and vocal quality in man," *J Acoust Soc Jpn (J)*, **57**, 308–313 (2001)
[35] K Honda, H Takemoto, T Kitamura, S Fujita and S Takano, "Exploring human speech production mechanisms by MRI," *IEICE Trans Inf Syst*, **E87-D**, 1050–1058 (2004)