A Study of Mandarin Chinese Using X-Ray and MRI

Gaowu WANG, Xugang LU, Jianwu DANG,
Huaiqiao BAO, Jiangping KONG

Abstract: This paper describes a primary study to establish a dynamic articulatory model by combining MRI technique and X-ray data, where the former is used to refine the detailed real shape of the vocal tract and the latter provides the dynamic information of articulation. In this study, MRI experiments were conducted to obtain 3D static morphologies of 9 single vowels of Mandarin, and the vocal tract shapes were investigated. A set of coefficients of the alpha-beta model has been derived from MRI data. The articulatory movement was obtained from a Mandarin X-ray video (cineradiography) database, which is the only available corpus for Mandarin, and the cross-sectional areas were calculated using the MRI based alpha-beta coefficients. For an evaluation, the formants estimated from the vocal tract area functions of both MRI and X-ray were compared with those obtained from real speech sound. The estimation was consistent with the real speech sound with a mismatch of about 10% and 15%, respectively.

Keywords: Mandarin X-ray MRI vocal tract

1. INTRODUCTION

To better understand speech production, from the phonological inputs to acoustic signals, going through motor control and physiological processes, it is important to establish an elaborate articulatory and vocal tract model. To do so, it is necessary to get a fine morphology, especially the detailed real shape of the vocal tract during speech.

So far, the technologies adopted in speech production to get vocal tract shape include X-ray photos and dynamic video (cineradiography), Electropalatography (EPG), computed tomography (CT), Ultrasonics, Magnetic Resonance Imaging (MRI), and Electromagnetic articulography (EMA). Each of them has its own advantages and disadvantages. For example, MRI can provide 3D static vocal tract shape data with high spatial resolution but poor temporal resolution, while cineradiography has higher temporal resolution but can only show 2D sagittal dynamic movement. It is required a mapping from widths in 2D plane to the area function if the transmission line model is employed in acoustic modeling.

In this research, we utilize the advantages of X-ray cineradiography and MRI observation. The latter can provide a mapping for the former by using the 2D and 3D information obtained from MRI observation. To do so, we draw a set of alpha-beta coefficients from 2D and 3D static morphologies in MRI data. These coefficients are used to estimate the cross-sectional area function from the midsagittal width in X-ray movie. So, we can establish a mapping from 2D widths to 3D areas for articulatory movements in X-ray movie. The estimation result is evaluated by using the transmission line model in acoustic modeling.

1.1 Introduction of X-ray studies in speech research

Research on speech using X-ray has a long history. According to Dart [1], as early as in 1907, Barth and Grunmach used still X-ray to study speech.

In 1942, Chiba [2] combined X-ray photography, palatography, and laryngoscopy to measure the vocal tract shape in their pioneering research.

Thereafter, Fant completed the theory of speech production based on the extensive analysis of X-ray data [3]. Also, some important studies [3-8] have been carried out using X-ray data, successively.

Although other new technologies such as MRI, CT, and EMA were developed, nowadays, the full sagittal view of vocal tract articulators during running speech provided by cineradiography remains unsurpassed by more modern techniques. Currently, the collection techniques are of two types: one permits real time tracking, but is limited to a few coplanar points (microbeam, magnetometer); the other gives full volume vocal tract images, but is limited to a static, sustainable configuration (MRI, CT). The data obtained from the cineradiography are still used in many recent research [9, 10].

In China, Zhou and Wu [11] recorded X-ray photos for Mandarin vowels and consonants uttered by a female native speaker. Bao and Yang [12] had measured the articulatory movements of five consecutive Mandarin vowels using the cineradiography. Bao [13] used X-ray still images to give a physiological explanation for the classification of Mandarin vowels. Also, some preliminary studies were performed on the relationship between the cross-sectional area function of the vocal tract and formant frequencies of vowels [14], and the synthesis of Mandarin single vowels by articulatory parameters [15].

1.2 Introduction of MRI studies

MRI allows a tomographic view of body tissues in any plane within the human body, and gets the 3D shape of the vocal tract without known risks for the subject. So, it

has been increasingly applied in speech research over the past 20 years [16-31].

The articulatory data collected using MRI are valuable in understanding and modeling the vocal tract accurately, particularly the pharynx, the behavior of which during speech is traditionally hard to capture. Also, the volumetric production data is very important in articulatory synthesis, in which scientists have been involved for several decades.

At present, MRI studies have been carried out on several languages, e.g. English, French, Japanese and Swedish, and so on. However, few MRI studies have been conducted on Mandarin Chinese. In this study, we investigated the morphological properties of 9 Mandarin single vowels.

2. DATA PROCESSING

In this section, we introduce the procedures for marking X-ray movie and obtaining MRI data.

2.1 X-ray video processing

An X-ray video database provided by Bao [13, 14], is the only available cineradiographs corpus of Mandarin. Both sagittal cineradiography and frontal cine-photographs of the lips are given simultaneously. In total 786 utterances of 2 male and 2 female subjects are shown to illustrate most of the sounds of Mandarin, covering 217 syllables, besides "retroflexed" syllables and tone contrasts. This video has been segmented to syllables in avi files at 30 fps.

A platform 'VocalMarker' in Matlab is programmed to trace the midsagittal articulatory movements. Figure 1 shows an example image of the marking of articulatory movement, from which the midsagittal width of the vocal tract has been measured.

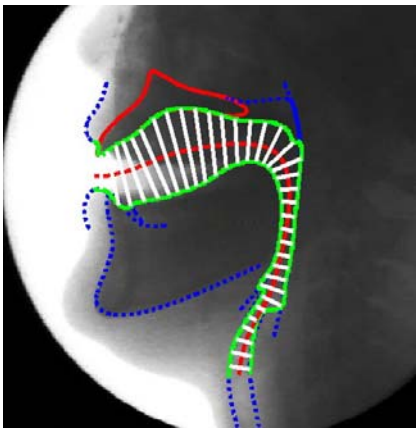

Figure 1: The midsagittal view of the marked X-ray video

2.2 MRI data acquisition and processing

2.2.1 Speech materials

This study covers the 9 single vowels in Mandarin, /a o e i u ü (i)e (s)i (sh)i/, in the Chinese Phoneticisation Scheme, whose IPA are / a o ɤ i u y e ɿ ʅ /, respectively, while another vowel /er/ (/ər/), whose status as a single vowel is still in dispute, is not included in our present study.

The 9 vowels are uttered in Chinese single syllable words, "啊喔嗣衣乌淤椰思诗", respectively, while the stable posterior parts in "椰思诗" were segmented and treated as independent vowels in the following speech analysis. In addition, all the characters are produced in the first tone (high flat tone) to ensure the stability of the sustained vowels.

2.2.2 Selection and training of subjects

The subjects have no speech or voice problems. They are native speakers of North Chinese dialects, live in or around the Beijing area, and have no other dialect accents. Alternatively, the subjects had passed the Putonghua Proficiency Test (PPT) and achieved Grade One, Level B (G1L2, the second highest grade, which is required for a Mandarin teacher and a television announcer).

After the subjects are selected, they are trained to ensure articulatory stability for obtaining clear MRI images. In the training, we require the subjects to produce the speech materials in a supine position with MRI noise in the earphones. Sufficient practice yields better imaging results.

At present, we have two subjects. Both of them are North Chinese from around Beijing area, and the female subject has passed the PPT G1L2.

2.2.3 MRI equipment and scan specifications

The MRI data was acquired with the Shimadzu-Marconi ECLIPSE 1.5T PowerDrive 250 installed at the Brain Activity Imaging Center, Advanced Telecommunications Research Institute (ATR-BAIC), in Kyoto, Japan.

Originally, a long standing drawback of MRI is the image acquisition time, which was several minutes per speech required in the earliest studies, while the acquisition time of this MRI machine was around 30 seconds for a 3D scan of the vocal tract, and the time varies with the number of image slices. To obtain a high quality image, it is still required that subjects maintain stable articulatory configurations during the image acquisition period, which might result in articulatory instability and subject motion during image acquisition, which in turn might create artifacts in the images.

Recently, a synchronized sampling method (SSM) with external trigger pulses developed by Masaki [32] was adopted in recording the movements of the speech organs as a set of sequential images. This method can also be used in acquiring the static 3D shape of vowels. The subject repeats the vowel about 30~36 times, each time sustains 3 seconds, which enables the subject to articulate in a stable manner. Figure 2 shows the experimental setup. The trigger device presents noise burst trains to the subject through a headset, and outputs the scan pulses to the MRI scanner to synchronize the data acquisition. The subject listens to the noise burst trains to pace the utterance, while the MRI scanner

initiates data acquisition synchronized with the trigger pulses.

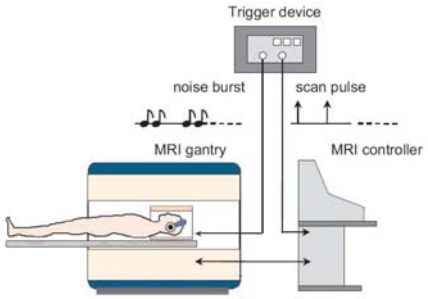

Figure 2: Experimental setup for the synchronized sampling method. (after Takemoto [33]).

The parameters were as following: a 3.4 [ms] echo time (TE), a 2200 [ms] relaxation time (TR), 44~51 sagittal slice planes, a 1.5 [mm] slice thickness, a 1.5 [mm] slice interval, a 256*256 [mm] field of view (FOV), and a 512*512 pixel image size. The data are stored in the DICOM file format.

2.2.4 MRI image preprocessing

The images were converted from DICOM into TIFF and denoised using ImageJ software, which is released by the NIH (National Institutes of Health, USA). Figure 3 shows an example of the image denoising effect, in which the articulatory structures are maintained with carefully checking.

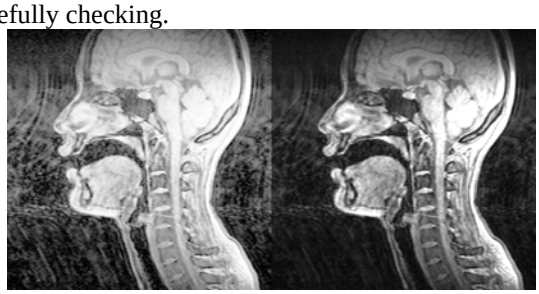

Figure 3: An example of the image denoising effect. Some noise spots (left panel) have been eliminated (right panel).

2.2.5 Teeth superimposition

MRI has a disadvantage in imaging the bony structure because the calcified structures lacking mobile hydrogen produce no resonance signals. Accordingly, the region of the teeth shows the same brightness as that of the air space. To build up an elaborate vocal tract model, however, it is necessary to obtain the teeth-air boundary to accurately reconstruct the vocal tract shape from the MRI data.

To solve this problem, we measured the structure of the teeth before or after obtaining the articulation data [33]. The subjects were asked to fill their mouth with a multi-mineral juice as a contrast medium and to lay prone in the MRI machine. An MRI scan is performed with the following parameters: an 11 [ms] echo time (TE), a 3000 [ms] relaxation time (TR), 51 sagittal slice planes, a 1.5 [mm] slice thickness, a 1.5 [mm] slice interval, a 256*256 [mm] field of view (FOV), and a 512*512 pixel image size. The data are stored in the DICOM file format.

In the teeth scan, the images showed the oral cavity with high brightness due to the contrast medium, while the teeth and jaws appeared with low brightness. This contrast makes it easy to extract the teeth and their supporting rigid structures (maxilla, mandible) from the oral cavity. The maxilla and the mandible with the teeth were reconstructed to obtain the “digital jaw casts”, which were then manually superimposed onto the original MRI volumes. Figure 4 shows the result of teeth extraction and imposition. We resliced the “digital jaw casts” to the same slices as the slices of the vowels data, and located the upper and lower teeth manually in the midsagittal plane, respectively, to minimize the superimposition error, as shown in the right panel of Figure 4.

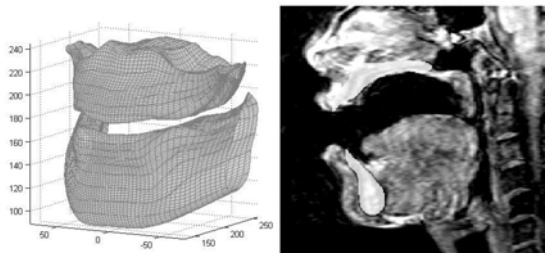

Figure 4: Left is the 3D digital jaw casts; Right is the midsagittal view after superimposition

2.2.6 Extracting the vocal tract area function

Vocal tract area functions were extracted from the reconstructed volumes with the teeth. Similar to [33], the extraction was performed in three steps. First, the vocal tract midline was semi-automatically calculated in the midsagittal image. Along the midline, then, images perpendicular to the midline were resliced at 1~5 mm intervals. Finally, the area of the vocal tract region in each section was measured to obtain the area function. Figure 5 shows the midline on an actual image of the vowel /i/, and the cross-sections from which the vocal tract area function is measured.

One remaining problem is how to measure the cross-sectional area of the vocal tract near the lip end, where the upper and lower lips are separated and a complete circumferential outline of the vocal tract section cannot be determined. We followed the method in [33]: as seen in Figure 5, we determined the furthest section from the glottis where the circumferential area could be measured as the last section (slice “h”), and the length of this section was extended to halfway from the end of this section to the last section where the upper and lower lips could still be observed (slice “i”).

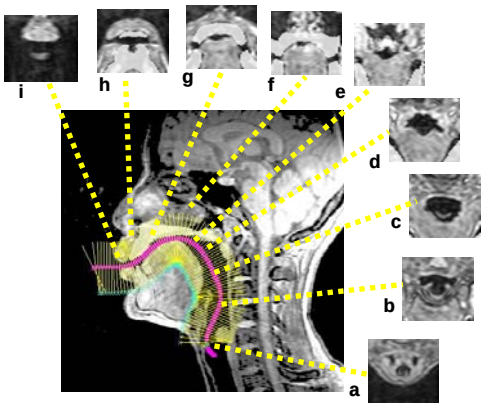

Figure 5: The results of extraction

2.2.7 Calculating transfer functions

The vocal tract transfer functions were calculated and the formants were estimated for all the volumes obtained by MRI using a transmission line model, which is detailed in [21].

2.2.8 Acoustic recording and analysis

Speech sounds were recorded from the subject in a soundproof room as the natural speech sound. The subject lay supine on the floor with a headset to listen to the noise burst trains. This is to reproduce the environment of MRI acquisition. In this situation, the subject is asked to repeat the vowels as much in the same way as in the MRI experiment. The speech signals and noise bursts were recorded with a recording system, consisting of a SONY ECM-G5M Microphone, a BEHRINGER EURORACK UB502 Mixer, a CREATIVE AUDIGY2NX Soundcard, and a DELL XPS M1210 Computer.

We selected the stable segment of the recorded vowels, and use the Praat software to extract the four lower formants.

3. RESULTS

3.1 MRI 3D inside view of Mandarin articulation

To explore the inside view, we reconstructed cutaway views based on volumetric MRI data for 9 Mandarin vowels. As a result, we obtained 3D shapes of Mandarin vowels, and show the inside vocal tract and articulators in Figure 6. The 3D images are produced by the 3DMed toolkit released by the Medical Image Processing Group, IA, CAS. To our knowledge, this is the first time to show a 3D shape for Mandarin vowels with the inner view. For this reason, we enlarge the special apical vowel of Chinese at the bottom of the figure. Such a result can be applied to the deaf or to people who have difficulties in learning Mandarin.

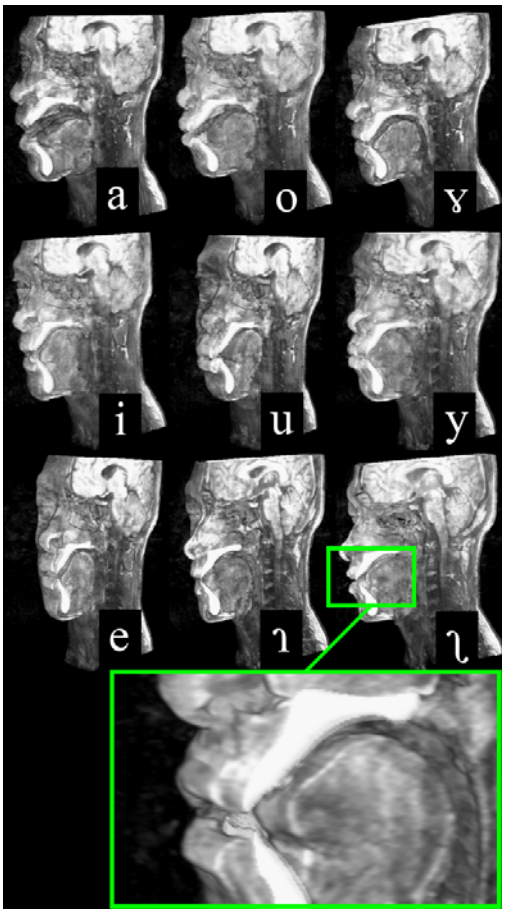

Figure 6: The 3D shapes of 9 single vowels in Mandarin.

3.2 MRI vocal tract shape and cross-sectional areas

In order to evaluate the reliability of the area functions extracted from the 3D MRI data, the areas of the cross-section have been extracted, and the corresponding formant frequencies have been estimated and compared with those of natural speech sound. Figure 7 shows the 3D vocal tract shape for vowel /a/.

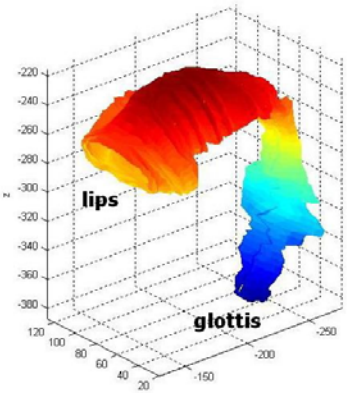

Figure 7: The 3D Vocal Tract for the vowel /a/

Table 1 shows the frequencies of first four formants, which are compared with those of natural speech sound. The percentage errors between them are less than 10%. Relatively large errors are found locally. The mean

absolute percentage error is 7.4%. This result is better than [34], whose mean absolute percentage error is 12.2% for English, but no so good as [33], whose error is 4.5%, for Japanese. While the Difference Limen (DL) for formant frequency discrimination range between 3% and 14%, which were reported in [35-37]

Therefore, in the future work, we will decrease the errors, and synthesize Mandarin single vowels using the estimated formants to perform a perceptual evaluation.

Table 1: The natural and calculated speech formants of the female subject in MRI, and the percentage errors for the latter relative to the former. "n" denotes the natural speech, "c" the calculation, and "d" percentage errors (%).

	/a/	/o/	/e1/	/i/	/u/	/v/	/e2/	/i2/	/i3/
nF1	814	590	600	325	390	320	560	420	410
nF2	1312	1000	1225	2660	770	1960	2120	1450	1810
nF3	3214	3160	3140	3460	2950	2470	2850	3170	2550
nF4	4354	4380	4380	4550	4150	3800	4430	4170	3370
cF1	737	550	554	321	382	300	504	440	442
cF2	1512	880	1382	2776	832	2015	1907	1385	1790
cF3	3307	2855	3474	3348	2984	2566	2649	3315	3162
cF4	3846	3845	4441	4368	3745	4030	3876	4308	3726
dF1	-9.5	-6.8	-7.7	-1.2	-2.1	-6.3	-10.0	4.8	7.8
dF2	15.2	-12.0	12.8	4.4	8.1	2.8	-10.0	-4.5	-1.1
dF3	2.9	-9.7	10.6	-3.2	1.2	3.9	-7.1	4.6	24.0
dF4	-11.7	-12.2	1.4	-4.0	-9.8	6.1	-12.5	3.3	10.6

3.3 From the X-ray midsagittal widths to cross-sectional areas using the alpha-beta model

Because X-ray video can only show the sagittal view of the vocal tract, it is necessary to find a method to estimate cross-sectional areas. At present, most transformations going from the midsagittal distance to the cross-sectional area are based on the original transformation defined by [5], which is the α β (alpha-beta) Model.

$$A(x) = \alpha(x) d(x)^{\beta(x)} \qquad (1)$$

where 'd' is the midsagittal distance, 'A' the cross-sectional area, 'x' the position along the vocal tract mid-line, and 'α' and 'β' are the two coefficients of the transformation, which are also functions of the variable 'x'.

In this study, a set of 'α' and 'β' coefficients has been calculated using the 'd' and 'A' from real MRI 3D data for 9 vowels, minimizing the estimation errors:

$$\arg\min_{\alpha(x),\beta(x)} \left\{ \sum_{V=/\,aoeiuv\ldots/} \left[A(x,V) - \alpha(x) d(x,V)^{\beta(x)} \right]^2 \right\} \qquad (2)$$

where 'V' represents the 9 single vowels in Mandarin. It means that 'A' and 'd' are phoneme dependent.

This set of alpha-beta coefficients reflects the morphological characteristics of this subject in MRI, so that we use Eq. (1) to estimate cross-sectional areas from midsagittal width of different vowels of this subject within limited errors.

And this set of alpha-beta coefficients is applied to the X-ray video data, using Eq. (1), to estimate 'A' from 'd' of the female subject in X-ray movie. From the estimated cross-sectional areas for the female subject in the X-ray database, we calculated the lower four formants of vowels (at present only /a i u/, due to the arduous X-ray video tracing), as shown in table 2, generally with a 15% mismatch as compared with the formants obtained from the real speech sound. Although the subjects are different, this is better than the empirical formula used in [14].

Table 2: The natural and calculated speech formants of the female subject 1 in the X-ray database

	/a/	/i/	/u/
nF1	911	390	450
nF2	1364	2862	921
nF3	3650	3699	3365
nF4	4232	4260	4165
cF1	786	398	489
cF2	1289	2476	977
cF3	3393	3423	2701
cF4	4027	4078	3386
dF1	-13.7	2.1	8.7
dF2	-5.5	-13.5	6.1
dF3	-7.0	-7.5	-19.7
dF4	-4.8	-4.3	-18.7

4. CONCULUSIONS AND DISCUSSION

This paper established a mapping from 2D widths to area functions based on MRI data, and applied to 2D articulatory movements in X-ray movie. By utilizing the advantages of X-ray movie and MRI data, we make use of morphology information in both of them.

At first, we extracted the articulatory movements from a Mandarin X-ray database, and measured the midsagittal widths of the vocal tract. Next, a set of alpha-beta coefficients were estimated from the MRI data. Finally, the area function of the vocal tract of X-ray was calculated using this set of alpha-beta coefficients. The

acoustic comparison showed that the combination of MRI and X-ray is available for further development.

However, the alpha-beta coefficients are speaker dependent, which will bring error when we apply them on the speaker in X-ray movie. In the future we will include more phonemes and syllables in Mandarin, not only some cardinal vowels /a i u/ in this paper, to evaluate the alpha-beta coefficients in both static and dynamic ways.

5. ACKNOWLEDGES

We sincerely thank the anonymous reviewers for their helpful comments and constructive suggestions.

This study is in part supported by a Grant-in-Aid for Scientific Research of Japan (No. 20300064) and SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC) of Japan.

6. REFERENCES

[1]Dart, S.N., A bibliography of X-ray studies of speech. UCLA Working Papers in Phonetics, 1987. **66**: p. 1-97.

[2]Chiba, T. and Kajiyama, M., *The Vowel: Its Nature and Structure*. 1942, Tokyo.

[3]Perkell, J.S., *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*. Research Mono. No. 53. 1969, Cambridge, MA.: MIT press.

[4]Ohman, S.E.G. and Stevens, K.N., Cineradiographic Studies of Speech: Procedures and Objectives. Journal of the Acoustical Society of America, 1963. **35**: p. 1889.

[5]Heinz, J.M. and Stevens, K.N., On the Derivation of Area Functions and Acoustic Spectra from Cineradiographic Films of Speech. Journal of the Acoustical Society of America, 1964. **36**(1): p. 37.

[6]Moll, K.L. and Daniloff, R.G., Investigation of the Timing of Velar Movements during Speech. The Journal of the Acoustical Society of America, 1971. **50**(2B): p. 678-684.

[7]Mermelstein, P., Articulatory model for the study of speech production. The Journal of the Acoustical Society of America, 1973. **53**(4): p. 1070-1082.

[8]Harshman, R., Ladefoged, P., and Goldstein, L., Factor analysis of tongue shapes. Journal of the Acoustical Society of America, 1977. **62**(3).

[9]Beautemps, D., Badin, P., and Bailly, G., Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. The Journal of the Acoustical Society of America, 2001. **109**(5): p. 2165-2180.

[10]Iskarous, K., Patterns of tongue movement. Journal of Phonetics, 2005. **33**(4): p. 363-381.

[11]Zhou, D.F. and Wu, Z.J., *The Articulation Album of Mandarin*. 1963.

[12]Bao, H.Q. and Yang, L.L., Study on cineradiography of consecutive vowels. Report of Phonetic Research, Phonetics Lab, CASS, 1982.

[13]Bao, H.Q., A physiological account of single vowels in Putonghua. ZHONGGUO YUWEN, 1984.

[14]Bao, H.Q., On the relationship between cross-sectional area function of vocal tract and formant frequencies of vowel: A preliminary report. Report of Phonetic Research, Phonetics Lab, CASS, 1983.

[15]Zu, Y.Q., The primary study of synthesizing vowels with articulatory parameters Report of Phonetic Research, Phonetics lab, CASS, 1983.

[16]Tiede, M., An MRI-based study of pharyngeal volume contrasts in Akan and English. Journal of Phonetics, 1996. **24**: p. 399-421.

[17]Demolin, D., Metens, T., and Soquet, A. Three-dimensional measurement of the vocal tract by MRI. in *Proceedings of 4th ICSLP*. 1996. Philadelphia.

[18]Baer, T., et al., Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. Journal of the Acoustical Society of America, 1991. **90**(2): p. 799-828.

[19]Lakshminarayan, A.V., Lee, S., and McCutcheon, M.J., MR imaging of the vocal tract during vowel production. Journal of Magnetic Resonance Imaging, 1991. **1**: p. 71-76.

[20]Moore, C.A., The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images. Journal of Speech and Hearing Research, 1992. **35**: p. 1009-1023.

[21]Dang, J. and Honda, K., Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation. Journal of the Acoustical Society of America, 1996. **100**(5): p. 3374-3383.

[22]Dang, J. and Honda, K., Acoustic characteristics of the piriform fossa in models and humans. Journal of the Acoustical Society of America, 1997. **101**(1): p. 456-465.

[23]Dang, J., Honda, K., and Suzuki, H., Morphological and acoustical analysis of the nasal and the paranasal cavities. Journal of the Acoustical Society of America, 1994. **96**(4): p. 2088-2100.

[24]Story, B.H., Titze, I.R., and Hoffman, E.A., Vocal tract area functions from magnetic resonance imaging. Journal of the Acoustical Society of America, 1996. **100**(1): p. 537-554.

[25]Alwan, A., Narayanan, S.S., and Haker, K., Toward articulatory-acoustic models for liquid consonants based on MRI and EPG data. Part II: The rhotics. Journal of the Acoustical Society of America, 1997. **101**: p. 1078-1089.

[26]Badin, P., et al. A three-dimensional linear articulatory model based on MRI data. in *Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*. 1998.

[27]Honda, K. and Tiede, M. An MRI study on the relationship between oral cavity shape and larynx position. in *Proceedings of 5th ICSLP*. 1998.

[28]Engwall, O., Vocal tract modeling in 3D. KTH STL-QPSR, 1999: p. 31-38.

[29]Fitch, W.T. and Giedd, J., Morphology and development of the human vocal tract: A study using magnetic resonance imaging. Journal of the Acoustical Society of America, 1999. **106**(3): p. 1511-1522.

[30]Jackson, P.J.B. and Shadle, C.H., Frication noise modulated by voicing, as revealed by pitch-scaled decomposition. Journal of the Acoustical Society of America, 2000. **108**(4): p. 1421-1434.

[31]Stone, M., et al. Modelling the Internal Tongue using Principal Strains. in *5th Seminar on Speech Production: Models and Data*. 2000. Kloster Seeon, Bavaria, Germany.

[32]Masaki, S., et al., MRI-based speech production study using a synchronized sampling method. J. Acoust. Soc. Jpn.(E), 1996. **20**: p. 375-379.

[33]Takemoto, H., et al., Measurement of temporal changes in vocal tract area function from 3D cine-MRI data. Journal of the Acoustical Society of America, 2006. **119**(2): p. 1037-1049.

[34]Story, B.H., Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002. The Journal of the Acoustical Society of America, 2008. **123**(1): p. 327-335.
[35]Flanagan, J., *A Difference Limen for Vowel Formant Frequency*. 1955, ASA. p. 613-617.
[36]Mermelstein, P., Difference limens for formant frequencies of steady-state and consonant-bound vowels. The Journal of the Acoustical Society of America, 1978. **63**(2): p. 572-580.
[37]Kewley-Port, D. and Zheng, Y., *Auditory models of formant frequency discrimination for isolated vowels.* 1998, ASA. p. 1654-1666.

Gaowu WANG, Jiangping KONG, Phonetics Lab, Peking University, Beijing, China, 100871
Xugang LU, ATR Spoken Language communication Research Labs, National Institute of Information and Communications Technology, Japan
Jianwu DANG, School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan, 923-1292
Huaiqiao BAO, Institute of Ethnology and Anthropology, CASS