MRI-based Study of Morphological and Acoustical Properties of Mandarin Sustained Steady Vowel

Gaowu WANG12, Tatsuya KITAMURA3, Xugang LU2, Jianwu DANG2, Jiangping KONG1

(1) Phonetics Lab, Peking University, Beijing, China, 100871

（2）School of Information Science, Japan Advanced Institute of Science and Technology

1-1, Asahidai, Nomi, Ishikawa 923-1292, Japan

(3) Dept. of Information Science and Systems Engineering, Faculty of Science and Engineering, Konan University

8-9-1, Okamoto, Higashinada, Kobe, 658-8501, Japan

Abstract: To establish an elaborate vocal tract model, it is necessary to obtain more detailed 3D vocal tract shapes. Such work has been done on some languages, e.g., English, French, Japanese and Swedish, using Magnetic Resonance Imaging (MRI), while Mandarin has not been investigated in detail yet. In this study, we investigated the morphological and acoustic properties of Mandarin based on MRI measurements. MRI experiments were conducted to obtain 3D static morphologies of Mandarin vowels for one male and one female Chinese speaker. The teeth shape has been superimposed onto the volumetric vocal tract data to complement the loss of the bony tissue in MRI imaging, and then vocal tract shapes are extracted based on the 3D MRI data with the implanted teeth. A set of grid planes is designed to be perpendicular to the central line of the vocal tract, and the cross-sectional area functions are obtained by calculating the areas of the airway on the sliced grid planes. To evaluate the morphological measurements, a comparison is carried out between the formants estimated from the vocal tract area functions and the ones obtained from the real speech sound. The results showed that the MRI based formants were consistent with those from real speech sounds within 10% mismatch mostly.

1. Introduction

Magnetic resonance imaging (MRI) allows a tomographic view of body tissues in any plane of the human body, and gets the 3D shape of vocal tract, without known risks for the subject. MRI has been increasingly applied in speech research over the past 20 years (Baer et al. 1991; Lakshminarayan et al. 1991; Moore 1992; Dang et al. 1994; Yang and Kasuya 1994; Dang and Honda 1996; Demolin et al. 1996; Story et al. 1996; Tiede 1996; Alwan et al. 1997; Dang and Honda 1997; Badin et al. 1998; Honda and Tiede 1998; Engwall 1999; Fitch and Giedd 1999; Jackson and Shadle 2000; Stone et al. 2000). The articulatory data collected using MRI is valuable in understanding and modeling the vocal tract in three dimensions, particularly the pharynx area, the behavior of which during speech is traditionally hard to capture. And the volumetric production data is very important in articulatory synthesis, which the scientists have been involved for more than several decades.

MRI studies have been performed for several languages, e.g. English, French, Japanese and Swedish, and so on. However, few MRI studies work on Mandarin Chinese. In this study, we investigated the morphological properties of Mandarin vowels. The MRI experiments were conducted to obtain 3D static morphologies of Mandarin vowels for one male and female Chinese speaker. Accordingly the 3D acquisition is the main focus of this study.

2 Data Acquisition

2.1 Speech materials

Mandarin, also called Putonghua, is the official national standard spoken language of China, which is based on the principal dialect spoken in and around Beijing. This study covers the nine single vowels in Mandarin, /a o e i u ü (i)e (s)i (sh)i/, in Chinese Phoneticisation Scheme, whose IPA are / a o ɤ i u y e ɿ ʅ /, respectively. While another vowel er (/ər/), is still in dispute whether it is a single vowel, which is not covered in our present study.

The vowels are uttered in Chinese words, “啊喔屙衣乌淤噎思诗”, while in “噎思诗” we selected the stable sustained end part of the sound to do speech analysis. And all the words are in their first tone (high flat tone) to ensure the stability of the

sustained vowels.

2.2 Choice and training of subjects

There are several considerations about choosing qualified subjects:

1). The subjects should have no dyslogia or dysphonia, and are good at Mandarin, which means, they should be a native speaker of North Chinese dialects, especially those in and around Beijing, and has no dialect accent. Or he/she has passed the Putonghua Proficiency Test (PPT) and achieved Grade One, Level B (G1L2, the second highest grade, which is required for Mandarin teacher and television announcer) .

2). He/she should have no mental objects in body, and have no other diseases unsuitable for MRI experiment.

After the subjects are chosen, they should be trained to ensure the articulatory stability so as to obtain clear MRI images. The training consists of producing the speech materials in a supine position with MRI noise in the earphones. Enough practices make better imaging results.

At present, we have two subjects, both of them are North Chinese around Beijing, and the female one has passed the PPT G1L2.

2.3 MRI equipment and scan specifications

MRI data were acquired by the Shimadzu-Marconi ECLIPSE 1.5T PowerDrive 250 installed at the ATR Brain Activity Imaging Center (ATR-BAIC).

Traditionally, a long standing drawback of MRI is the image acquisition time, which was several minutes per speech taken in the earliest studies.

Recently, acquisition time was shortened to around 30s, which depends mainly on the number of image planes used, the development of the MR machines and the exploration of technical possibilities done by physicists controlling the MR machines. However, this still requires that subjects hold articulatory configurations steady for durations which are not only extra-linguistic but are also motorically difficult to produce. This might result in articulatory instability and subject motion during image acquisition, which in turn might create image artifacts.

At present, a synchronized sampling method (SSM) with external trigger pulses has been developed by (Masaki et al. 1996) to record movements of the speech organs as a set of sequential images. And this method can also be used in acquiring static 3D shape of vowels. The subject repeats the vowel about 30~36times, each time sustained 3s, which can allow subject to articulate stably and naturally. Figure 1 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.

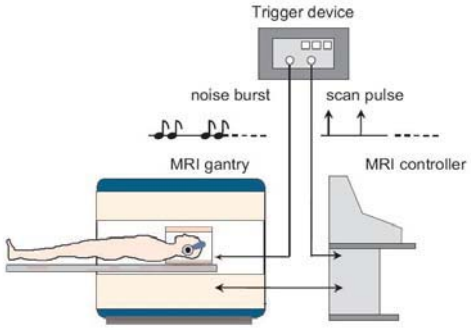

FIG. 1. Experimental setup for the synchronized sampling method. The trigger device, PC-based circuitry, controls the timing of a subject's utterance and that of MRI scanning. The device represents noise burst trains to a subject through the headset, and it sends scan pulses to the MRI controller. The noise burst trains and scan pulses synchronize with each other. after (Takemoto et al. 2006).

The parameters used in the SSM MRI scans were as follows: 3.4 ms echo time (TE), 2200 ms relaxation time (TR), 44~51 sagittal slice planes, 1.5mm slice thickness, 1.5 mm slice interval (which means the gap and overlap between slices are 0 mm), 256*256 mm field of view (FOV), and 512*512 pixel image size. The data thus obtained consist of 44~51 sagittal slices, stored as the DICOM file format.

3. MRI data processing

3.1 Image preprocessing

The images were converted from DICOM to TIFF and denoised using ImageJ software, which is released by NIH(National Institutes of Health, USA). Figure 2 shows an example of image denoising effect.

FIG. 2. An example of image denoising effect. The noise spots have been eliminated in the right panel.

3.2 Tooth superimposition

MRI has a disadvantage in imaging the bony structure because the calcified structures lacking mobile hydrogen give no resonance signals. Accordingly, the region of the teeth is found as dark as the air space. It is therefore necessary to obtain the teeth-air boundary to measure the vocal tract area function accurately from MRI data.

We referred to a method for teeth imaging to solve this problem (Takemoto, Honda et al. 2006). The same subject lay prone in the MRI gantry holding multi mineral juice in the mouth as a contrast medium, while static MRI scan was performed with the following parameters: 11 ms echo time (TE), 3000 ms relaxation time (TR), sagittal slice plane, 51 images, 1.5 mm slice thickness, a 1.5 mm slice interval, no averaging, 256*256 mm field of view (FOV), and 512*512 pixel image size. The images thus obtained demonstrated the oral cavity with high pixel values (bright), while the teeth and jaws appeared with low pixel values (dark). This contrast makes it easy to segment the teeth with the supporting rigid structures 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 MRI volumes. Figure 3 shows the result of tooth imposition.

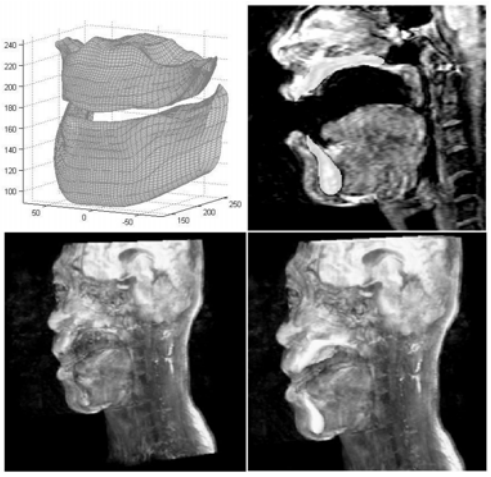

FIG. 3.The results of tooth superimposition. Upper-left is the 3D digital jaw casts; upper-right is the mid-sagittal plane after imposition; lower-left and right are the 3D show of tooth superimposition.

3.3 Extracting vocal tract area function

Vocal tract area functions were extracted from the reconstructed volumes with the teeth. The extraction was performed in three steps. First, the vocal tract midline was semi-automatically calculated in the mid-sagittal 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. We use the method proposed by Takemoto et al. 2006. Figure 4 shows the results of calculating the vocal tract midline on an actual image: the locations of cross sections from which the vocal tract area function is measured.

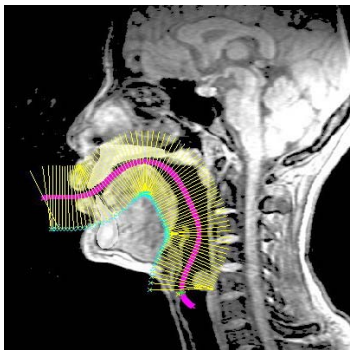

FIG. 4. The locations of grid planes

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. The solution used in this study followed Takemoto et al. 2006, to apply the measurable area nearest to the problematic sections.

In the furthest section from the glottis where the circumferential area could be measured, the length of the section was extended to half the distance from the end of that section to the last section where the upper and lower lips could still be observed.

3.4 Calculating transfer function

The vocal tract transfer functions were calculated for all the volumes obtained by MRI using a transmission line model, which is detailed in (Dang and Honda 1996).

3.5 Acoustic recording and analyses

Speech sounds were recorded from the subject in a soundproof room. 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 repeating the vowels in the same way did in MRI experiment as possible. The speech signals and noise bursts were recorded with a record system, which consists of:

Microphone: SONY ECM-G5M

Soundcard: CREATIVE SOUNDBLASTER AUDIGY2NX

Computer: DELL XPS M1210

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

4 Results

4.1 3D inside view of Mandarin articulation

To explore the inside view, we reconstructed cutaway views for 9 Mandarin vowels based on volumetric MRI data. Thus, we obtained the 3D shape of Mandarin vowels and show the inside vocal tract and articulators, which have never been achieved by other studies of Mandarin, as we know present.

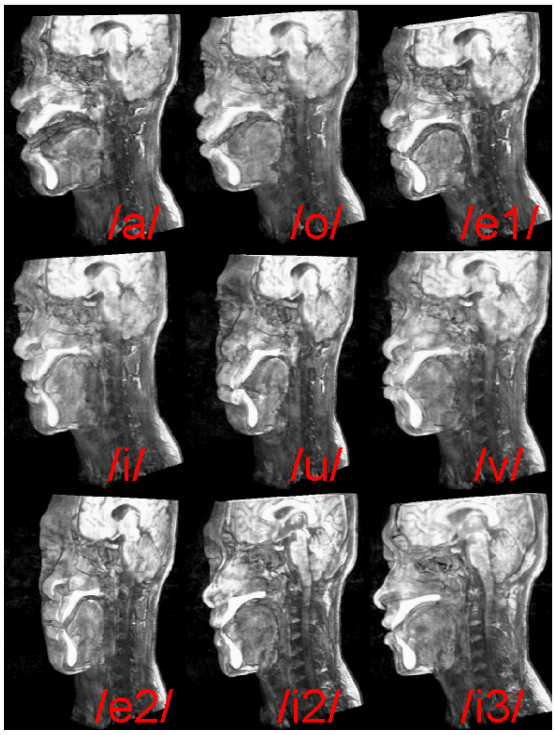

FIG. 5. The 3D shapes of 9 single vowels in Mandarin

4.2 The 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 had been extracted, and the corresponding formant frequencies had been derived and compared with those of real speech sound.

Figure 6 shows the results of the straightened VT volume of vowels /a i u/, and their binarized grid planes

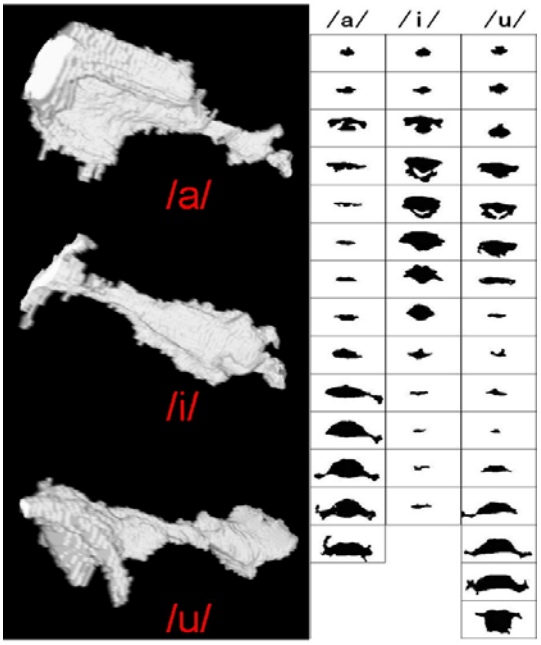

FIG. 6. The straightened 3D VT shapes of /a i u/ and their cross

sectional areas

Table 1 shows the first four formants, which are compared with those of natural speech sounds. The absolute percent error between all the formant data from the transfer functions and natural speech sounds is less than 10%. Relatively larger errors are found regionally.

Table 1: the natural and calculated speech formants of the female subject, and the percent errors for the latter relative to the former. “n” denotes the natural speech, “c” the calculation, and “d” percent 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

5 Conclusions and Discussions

In this study, the 3D static morphologies of 9 sustained Mandarin vowels had been investigated. The production data, in which the different articulators are detailed depicted, will help us to establish an articulatory vocal tract model.

The 3D inside view model will also be helpful in Visual Speech Training Aid for those disable persons who cannot hear but learn speech only by mimic the positions and movements from what they can look.

As for the formants calculated from the production data have an error no more than 10% mostly, which is advanced than early research. (Bao 1983) had calculated the formants of Mandarin of a female subject, as listed in table 2.

Table 2. The results of Bao (1983). ‘r’ denotes the revised values

	/a/	/o/	/e1/	/i/	/u/	/v/	/e2/	/i2/	/i3/
nF1	928	795	686	328	428	338	673	490	465
nF2	1300	805	1130	3000	606	2900	2575	1768	2010
cF1	621	534	487	219	304	203	467	275	297
cF1r	766	674	625	355	438	340	604	409	431
cF2	1263	1212	1542	2222	827	2070	1806	1616	1869
dF1	-33.1	-32.8	-29.0	-33.2	-29.0	-39.9	-30.6	-43.9	-36.1
dF1r	-17.5	-15.2	-8.9	8.1	2.2	0.5	-10.2	-16.6	-7.4
dF2	-2.8	50.6	36.5	-25.9	36.5	-28.6	-29.9	-8.6	-7.0

Bao (1983) obtained the area function from the X-ray photo of the 9 vowels of a female subject, and used a simplified transmission line model to calculate the formants. The percent errors of F1 are more than 30% systematically. So he adopted the experiential revised formula propose by (Lonchamp et al. 1983) to reduce the errors.

Compared with Bao (1983), our improvement may be attributed to these reasons:

1). The X-ray photo can only show the mid-sagittal plane, so Bao has to use an experiential formula to calculate the cross sectional areas from mid-sagittal dimensions. While now we can measure the real cross sectional areas using MRI.

2). Bao (1983) used the simplified transmission line model, while we adopt a model including wall impedance.

In the future, we will get more detailed vocal tract shape of Mandarin vowels and consonants, and the branched parts such as nasal cavity, piriform fossa, which will help us to establish an elaborated articulatory vocal tract model.

Acknowledgement

This study is supported by Grant-in-Aid for Scientific Research of Japan (No. 17300182) and in part by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan.

References

[1]Alwan, A., S. S. Narayanan, et al. (1997). "Toward articulatory-acoustic models for liquid consonants based on MRI and EPG data. Part II: The rhotics." Journal of the Acoustical Society of America 101: 1078-1089.

[2]Badin, P., G. Bailly, et al. (1998). A three-dimensional linear articulatory model based on MRI data. Proceedings of the Third ESCA/COCOSDA International Workshop on Speech Synthesis: 249-254.

[3]Baer, T., J. C. Gore, et al. (1991). "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels." The Journal of the Acoustical Society of America 90(2): 799-828.

[4]Bao, H. Q. (1983). "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.
[5]Dang, J. and K. Honda (1996). "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation." The Journal of the Acoustical Society of America 100(5): 3374-3383.
[6]Dang, J. and K. Honda (1997). "Acoustic characteristics of the piriform fossa in models and humans." The Journal of the Acoustical Society of America 101(1): 456-465.
[7]Dang, J., K. Honda, et al. (1994). "Morphological and acoustical analysis of the nasal and the paranasal cavities." The Journal of the Acoustical Society of America 96(4): 2088-2100.
[8]Demolin, D., T. Metens, et al. (1996). Three-dimensional measurement of the vocal tract by MRI. Proceedings of 4th ICSLP. Philadelphia: 272-275.
Engwall, O. (1999). "Vocal tract modeling in 3D." KTH STL-QPSR: 31-38.
[9]Fitch, W. T. and J. Giedd (1999). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging." The Journal of the Acoustical Society of America 106(3): 1511-1522.
[10]Honda, K. and M. Tiede (1998). An MRI study on the relationship between oral cavity shape and larynx position. ICSLP 5th.
[11]Jackson, P. J. B. and C. H. Shadle (2000). Aero-Acoustic Modelling of Voiced and Unvoiced Fricatives based on MRI Data. 5th Seminar on Speech Production: Models and Data. Kloster Seeon, Bavaria, Germany: 185-188.
[12]Lakshminarayan, A. V., S. Lee, et al. (1991). " MR imaging of the vocal tract during vowel production." Journal of Magnetic Resonance Imaging 1: 71-76.
[13]Lonchamp, F., J. P. Zerling, et al. (1983). Estimation vocal tract area functions: A progress report. Proceding of 10th ICPS: 3271.
[14]Masaki, S., M. K. Tiede, et al. (1996). "MRI-based speech production study using a synchronized sampling method." J. Acoust. Soc. Jpn.(E) 20: 375-379.
[15]Moore, C. A. (1992). "The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images." Journal of Speech and Hearing Research 35: 1009-1023.
[16]Stone, M., D. Dick, et al. (2000). Modelling the Internal Tongue using Principal Strains. 5th Seminar on Speech Production: Models and Data. Kloster Seeon, Bavaria, Germany: 133-136.
[17]Story, B. H., I. R. Titze, et al. (1996). "Vocal tract area functions from magnetic resonance imaging." The Journal of the Acoustical Society of America 100(1): 537-554.
[18]Takemoto, H., K. Honda, et al. (2006). "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data." Journal of the Acoustical Society of America 119(2): 1037-1049.
[19]Tiede, M. (1996). "An MRI-based study of pharyngeal volume contrasts in Akan and English." Journal of Phonetics 24: 399-421.
[20]Yang, C. S. and H. Kasuya (1994). Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects. ICSLP. Yokohama, Japan: 623-626.