Estimation of Vocal Tract Area Function for Mandarin Vowel Sequences Using MRI

Gaowu Wang [12], *Jianwu Dang* [1], *Jiangping Kong* [2]

[1] School of Information Science, Japan Advanced Institute of Science and Technology
[2] Phonetics Lab, Peking University, Beijing, China
wanggaowu@pku.edu.cn, jdang@jaist.ac.jp, jpkong@pku.edu.cn

Abstract

To fully explore the dynamic properties of speech production and investigate the relation between vocal tract geometry and speech acoustics, estimation of vocal tract area functions from measurements of the sagittal plane is an important step. In this study, we investigated the relation between the measurements on two dimensional (2D) and three dimensional (3D) MRI data and used an alpha-beta model to describe this relation. As a result, a set of parameters were derived from 3D static MRI data, and applied to time-varying vocal tract widths derived from 2D MRI movies, to synthesize Mandarin vowel sequences. An acoustic evaluation comparing the natural and calculated formants shows that the alpha-beta model can represent dynamic states of articulatory movements of vowel sequences, as well as those of the sustained vowels.

Index Terms: MRI, speech production, Mandarin, vowel, alpha-beta model

1. Introduction

Magnetic resonance imaging (MRI) allows a tomographic view of body tissues in any plane of the human body, and yields the 3D shape of the vocal tract, without any known risk for the subjects. MRI has been increasingly applied in speech research over the past 20 years [1-6]. The articulatory data collected using MRI is valuable in understanding and modeling the vocal tract accurately, particularly the pharynx area, since the behavior of this area is hard to capture during speech by traditional approaches. The volumetric morphological data is very important for articulatory synthesis, in which scientists have been involved for several decades.

At present, MRI studies have been carried out on several languages. However, few MRI studies have been conducted on Mandarin. MRI will be very valuable for research on Mandarin, since our final goal is to develop a dynamic 3D articulatory model and a visual aid system for Mandarin learning, which requires adequate morphological and articulatory data. However, due to the high cost of MRI experiments for obtaining 3D data of vocal tract, the 2D articulatory data in mid-sagittal plane are more attractive for speech research, where the generation of area functions from measurements of the sagittal section is an important step in the study of the relation between 2D and 3D geometry of the vocal tract in acoustic aspect. Several such generation methods have been proposed in the past [7-10] and the alpha-beta model is most famous.

In this study, we performed MRI experiments to get static and dynamic data of vocal tract, and uses alpha-beta model to estimate area functions of Mandarin vowel sequences.

2. Data Acquisition

In this section, we introduce the procedures for obtaining MRI data, including the selection of speech materials and subjects, the paradigm of MRI experiments, and the pre-processing of the MRI data. Due to the laborious processing of MRI data, at present, we only show the results of one female subject.

2.1. Speech materials

Mandarin, also called Putonghua, is the official national standard spoken language of China and is derived from the principal dialect spoken in and around the Beijing area.

As for the sustained vowels, our study covered the nine single vowels in Mandarin, /a o e i u ü (i)e (s)i (sh)i/, in the Chinese phoneticization scheme, whose IPA symbols are / a o ɤ i u y e ɿ ʅ /, respectively. There is one more vowel, “er” (/ər/), whose status as a single vowel is still in dispute. Therefore, we did not include this vowel in the present study. The vowels were uttered by saying the Chinese characters “啊喔嗣衣乌淤噎思诗”, where the vowels / e ɿ ʅ / in “噎思诗” are consonant-dependent and appear stably with preceding specific consonants or semivowels. For this reason, we asked subjects to pronounce the syllables with these three vowels instead of the isolated vowels, and we used the stable segments of the vowels in the following speech analysis. In addition, all the characters were produced in the first tone (high flat tone) to ensure the stability of the sustained vowels.

As for the dynamic movements, we selected 39 syllables, including diphthongss (e.g. /ai ei ao/), triphthongs (e.g. /iao iou uai/), and CV syllables (e.g. /ba bi bu/). 3 syllables were uttered in each MRI section, e.g. /ai ei ao/ which were uttered by saying the Chinese characters “哀诶凹”.

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 and around the Beijing area, and have no other dialect accents. Alternatively, the subjects have 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 announcers).

The subjects were trained to ensure articulatory stability so that we could obtain clear MRI images. In the training, we required the subjects to produce the speech materials in a supine position with MRI noise in the earphones. Sufficient practice yielded better imaging results.

2.3. MRI equipment and scan specifications

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

As is well known, the major drawback of MRI is its poor time resolution for speech research. At present, a synchronized sampling method (SSM) developed by [11] is adopted in recording the movements of the speech organs as a set of sequential images. This method can also be used for acquiring the static 3D shape of vowels.

In this study, the parameters used in the SSM MRI scans for sustained vowels are shown in Table 1.

Table 1. *Parameters for sustained vowels*

Echo time (TE)	3.4 ms
Relaxation time (TR)	2200 ms
Number of slices	44-51 sagittal slice planes
Slice thickness	1.5 mm
Slice interval	1.5 mm
Field of view (FOV)	256*256 mm
Image size	512*512 pixels
Image data format	DICOM file

The parameters of SSM MRI for dynamic movement of vowel sequences were as follows: a 2200 [ms] sequence length with 128 phases, the mid-sagittal slice, an FOV of 256*256 [mm], an image size of 256*256 pixels. As the result, for a vowel sequence, e.g. /ai ei ao/, we got 128 frames in mid-sagittal plane, which formed a real time movie at 60fps.

2.4. Image preprocessing

The images were converted from DICOM to TIFF and denoised using ImageJ software, which was produced by the National Institute of Health, USA.

2.5. Teeth superimposition

MRI has a disadvantage in imaging bony structures because calcified structures that lack mobile hydrogen produce no resonance signals. Accordingly, the region of the teeth has the same darkness as the air space. To solve this problem, we measured the structure of the teeth using the teeth imaging method proposed in [12]. The maxilla and the mandible with the teeth were reconstructed to obtain “digital jaw casts”, which were then manually superimposed onto the original MRI images. Figure 1 shows the result of teeth extraction and imposition. We resliced the digital jaw casts to match the slices of the vowel data, and we located the upper and lower teeth manually in the mid-sagittal plane, to minimize superimposition error, as shown in the right panel of Figure 1.

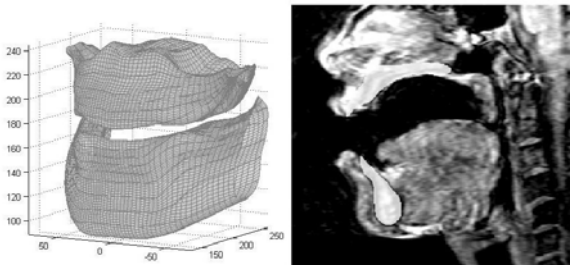

Figure 1: *left: 3D digital jaw casts; right: mid-sagittal plane after imposition.*

2.6. Data processing for sustained vowels

To study the acoustic properties of the vowels based on MRI data, we use the classical transmission line model [13-15]. This requires us to obtain the vocal tract area functions, which can be calculated from the reconstructed vocal tract shape with the teeth shown above.

The calculation was performed in three steps. First, the vocal tract midline was semi-automatically calculated in the mid-sagittal image. Then, along the midline, images perpendicular to the midline were resliced at 1~5 mm intervals. Finally, the cross-sectional area of the airway of the vocal tract in each slice was measured. The vocal tract area function is defined as a series of cross-sectional areas along the vocal tract. Figure 2 shows the midline of the vocal tract on the midsagittal plane, the grid lines for this vocal tract, and the cross-sectional images sliced according to the grid lines. The vocal tract area function is derived from these slices.

As seen in Figure 2, close to the lip end, the upper and lower lips are separated and a complete circumferential outline of the vocal tract section cannot be determined. It was necessary to answer the following two questions. Where is the end of the vocal tract in terms of acoustic meaning? How should the area of the incomplete shape be measured? In this study we adopted a method proposed by Takemoto [12]. As shown in Figure 2, 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”).

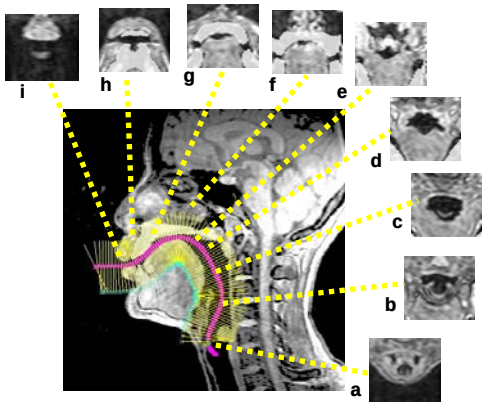

Figure 2: *Results of extraction of the cross-sectional slices.*

Subsequently, the vocal tract transfer functions were calculated based on the cross-sectional area function of the vocal tract obtained from MRI using a transmission line model, which is detailed in [15]. The calculated formant frequencies are listed in Table 2.

The speech sounds of the subjects were recorded in a soundproof room to remove the noise of MRI machine. In order to reproduce the situation of MRI acquisition, the subjects lay supine on the floor with a headset that fed the noise burst trains to the subjects. The subjects were then asked to repeat the vowels as similarly as possible to the way they spoke in the MRI machine.

We selected the stable segment of the recorded vowels, and used the Praat software to extract the four lower formants. Table 2 shows the first four formants calculated from MRI data and the formants of natural speech sounds for comparison. The absolute percentage error between all the

formant data from the transfer functions and from natural speech sounds is about 10%. The mean absolute percent error is 7.4%. This result is not as good as [12], in which the mean absolute percent error was 4.5% for Japanese vowels, but is better than a recent MRI study [16], in which the mean absolute percent error was 12.2%.

Table 2. *Comparison of the natural and calculated speech formants, "n" denotes natural speech, "c" the calculation, and "d" the percentage error.*

	a	o	e	i	u	ü	(i)e	(s)i	(sh)i
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

2.7. Data Processing for vowel sequences

A program 'VocalTractMarker' in Matlab was designed for semi-automatic tracing of the articulatory movements in the mid-sagittal plane. Figure 3 shows an example image of this marking. From the glottis to the lips along the midline of the air way, we measured the widths of the vocal tract at certain intervals.

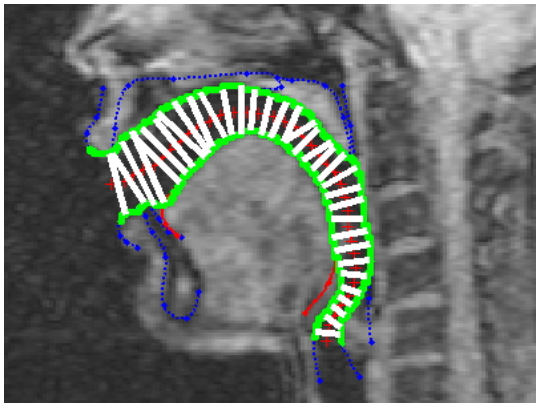

Figure 3: *Measuring the widths in the mid-sagittal plane in MRI movie.*

3. Applying alpha-beta model

3.1. For sustained vowels

Because the 2D dynamic MRI movie can only show 2D information of the vocal tract in the sagittal view, it is necessary to find a method to estimate cross sectional areas. At present, many transformations going from the mid-sagittal width to the cross-sectional area are based on the original transformation defined by [17], which is called the "α β" (alpha-beta) Model.

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

where 'd' is the width of the vocal tract in the mid-sagittal plane, 'A' is the cross-sectional area, 'x' is the distance from the glottis along the vocal tract midline, and 'α' and 'β' are the two parameters of the transformation, which are also the functions of variable 'x'.

A set of 'α' and 'β' parameters was calculated using 'd' and 'A' from real MRI 3D data of 9 vowels, by minimizing the estimation errors. This set of alpha-beta parameters reflects the morphology characteristics of this subject, so that we can use Eq. (1) to estimate cross-sectional areas from the mid-sagittal widths of different vowels of this subject with a mean absolute error of 0.29 [cm^2] for all 9 vowels, while the errors are 0.32, 0.57, 0.24, 0.18, 0.44, 0.11, 0.24, 0.28, 0.21 [cm^2] for / a o ɤ i u y e ɪ ɻ /, respectively.

3.2. For vowel sequences

We used this set of alpha-beta parameters to calculate the cross-sectional areas of the vowel sequences, from which the formants were calculated and compared to the formants of real speech.

Figure 4 shows the movement of articulators during the diphthongs /ai/ in vowel sequence /ai ei ao/. From the articulatory posture of /a/ to that of /i/, the varying cross-sectional areas have been calculated by the alpha-beta model. One can see that while producing /ai/ the contours of the tongue have an invariant point, which is reflected in the area functions also.

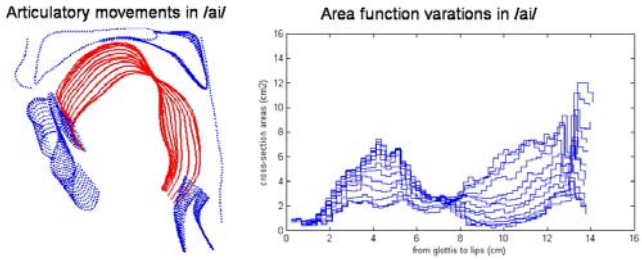

Figure 4: A*rticulatory movements and area function variations in diphthongs /ai/.*

4. Conclusions and discussion

For a clear contrast, we compared the formant trajectories in the acoustic plane and in the time domain, taking /ai ei ao/ as an example. In Figure 5, the large black dots indicate the natural formants of the 9 sustained vowels. The large blue asterisks indicate the formants calculated from the cross-section area functions estimated using the alpha-beta model on 2D data. And the small black dots demonstrate the trajectories of the formants of the recorded natural speech in

the acoustic plane. The small blue asterisks are the calculated formants from the 2D dynamic MRI movie. The upper panel shows the trajectories of F1 and F2 for /ai ei ao/ in the acoustic space, while the lower panel shows the trajectories in the time domain.

In the upper panel, two of the four open circles indicate the articulation places of /a/ and /i/ in diphthongs /ai/, respectively. One can see that the /a/ in /ai/ is more forward than single vowel /a/, and the articulation place of /i/ is not fully achieved in diphthongs /ai/.

In the lower panel, the formant trajectories calculated using alpha-beta model are consistent with those from natural speech. This shows that the alpha-beta model also performed well on dynamic vocal tracts, and the coarticulation trajectory information has been preserved.

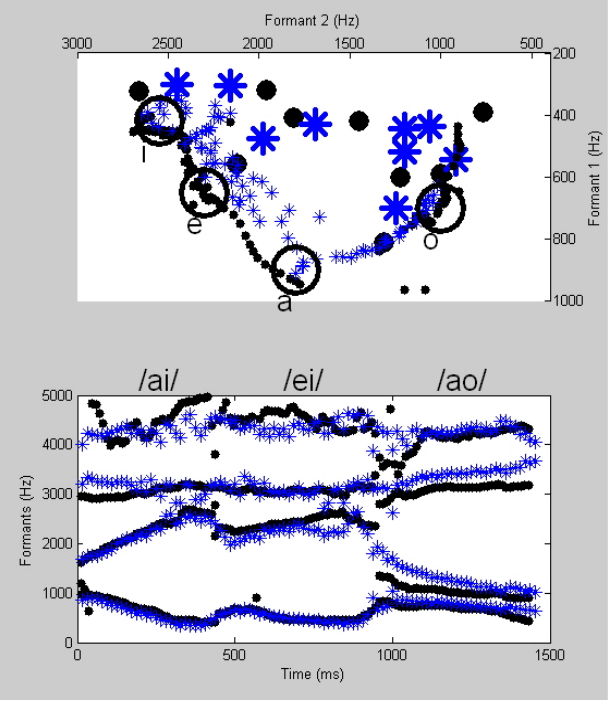

Figure 5: *Comparison of the real and calculated formant trajectories in the acoustic plane and in the time domain for vowel sequence /ai ei ao/.*

However, as shown in the lower panel of Figure 5, in the diphthongs /ao/, there is a large difference between the formant trajectories of the real speech sound and the calculation. This may be introduced from the difference between the states of the subject in the sound proof room and the MRI room. The noise in the MRI room and fatigue prevent the subjects from maintaining stable articulation. In the case of /ao/, it was confirmed that the trajectory for sounds during the MRI experiment was close to the calculation, but the trajectories were generally not clear due to the MRI noise. This will be investigated in future research.

Due to the laborious processing of MRI data, in this study, at present, we only show some examples of one female subject. In the future, for more generality, we will apply this study on more subjects.

5. Acknowledgement

This study is supported in part by SCOPE (071705001) of the Ministry of Internal Affairs and Communications (MIC) of Japan.

6. References

[1] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *Journal of the Acoustical Society of America*, vol. 90, pp. 799-828, 1991.

[2] J. Dang, K. Honda, and H. Suzuki, "Morphological and acoustical analysis of the nasal and the paranasal cavities," *Journal of the Acoustical Society of America*, vol. 96, pp. 2088-2100, 1994.

[3] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *Journal of the Acoustical Society of America*, vol. 100, pp. 537-554, 1996.

[4] M. Tiede, "An MRI-based study of pharyngeal volume contrasts in Akan and English," *Journal of Phonetics*, vol. 24, pp. 399-421, 1996.

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

[6] O. Engwall, "Vocal tract modeling in 3D," *KTH STL-QPSR*, pp. 31-38, 1999.

[7] G. Fant, "Vocal tract area functions of Swedish vowels and a new three-parameter model," in *the International Conference on Spoken Language Processing*,, Banff, 1992, pp. 807-810.

[8] J. Sundberg, "On the problem of obtaining area functions from lateral X-ray pictures of the vocal tract " *STL QPSR.*, pp. 43-45, 1969.

[9] P. Perrier, L. J. Boe, and R. Sock, "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: modeling the transition with two sets of coefficients. ," *J. Speech Hear. Res.*, vol. 35, pp. 53-67, 1992.

[10] A. Soquet, V. Lecuit, T. Metens, and D. Demolin, "Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI," *Speech Communication*, vol. 36, pp. 169-180, Mar 2002.

[11] S. Masaki, M. K. Tiede, K. Honda, Y. Shimada, I. Fujimoto, Y. Nakamura, and N. Ninomiya, "MRI-based speech production study using a synchronized sampling method," *J. Acoust. Soc. Jpn.(E)*, vol. 20, pp. 375-379, 1996.

[12] H. Takemoto, K. Honda, S. Masaki, Y. Shimada, and I. Fujimoto, "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," *Journal of the Acoustical Society of America*, vol. 119, pp. 1037-1049, Feb 2006.

[13] G. Fant, *Acoustic Theory of Speech Production: With Calculation Based on X-Ray Studies of Russian Articulation*. Mouton: The Hague, 1960.

[14] J. Flanagan, L., *Speech Analysis Synthesis and Perception*. New York: Spinger, 1972.

[15] J. Dang and K. Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation," *Journal of the Acoustical Society of America*, vol. 100, pp. 3374-3383, 1996.

[16] B. H. Story, "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*, vol. 123, pp. 327-335, 2008.

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