ANALYSING TONGUE SHAPE AND MOVEMENT IN VOWEL PRODUCTION USING SS ANOVA IN ULTRASOUND IMAGING

Yu Chen & Hua Lin

University of Victoria, Canada
chenyu@uvic.ca; hualin@uvic.ca

ABSTRACT

Using ultrasound technique, this study revealed certain previously undocumented tongue behaviours of Mandarin vowels /a/, /i/ and /u/. The results were achieved by using a modified SS ANOVA method that was capable of analyzing and comparing simultaneously three or more objects (e.g., vowels) in ultrasound imaging. Also developed and used was a correction method to compensate for jaw rotation.

Keywords: SS ANOVA, ultrasound, vowel, tongue shape, jaw movement

1. INTRODUCTION

Ultrasound imaging offers a safe and low-cost way of observing real-time tongue movement. However, how to accurately analyze the recorded data is still a question whose answers are mostly tentative.

Among the methods advanced by researchers [3, 10, 11], the Smoothing Spline ANOVA (SS ANOVA) used by Davidson [4] was most promising for the analysis of tongue shape and movement. Davidson's study argued that the SS ANOVA was a useful technique for providing a statistical analysis of the differences among tongue shapes acquired by ultrasound imaging, and that change in shape, rotation, or transition was taken into account in the statistical analysis. Shortly after Davidson's study, Baker [1] released R Code for SS ANOVA which has enabled researchers to compare two objects in ultrasound imaging. Davidson's and Baker's methods have their limitations. For one thing, the *ssr::assist* package they adopted for SS ANOVA is limited [5] in that it can only compare two items at a time.

Based on the *gss* package [6], we developed a new method of SS ANOVA and used it to compare multiple items simultaneously; specifically, we used the method to process and analyze a set of ultrasound images of Mandarin vowels, /a/, /i/ and /u/, produced by native Mandarin speakers. We attempted to answer the following questions:

- How does the tongue move during the production of a specific vowel such as Mandarin's /a/, /i/ or /u/?
- What are the differences in tongue shape among the three vowels?
- Is the articulatory working space defined by /a/, /i/ or /u/ parallel to the vowel formant chart?

2. EXPERIMENT

2.1. Subjects

The subjects were a male and a female native speaker of Mandarin Chinese. Although coming from non-standard Mandarin areas, they have native fluency of standard Chinese.

2.2. Materials

Three vowels /a/, /i/ and /u/ with high level tone in standard Chinese were recorded. The choice of the vowels was based on studies such as Lindblom's [8] in which /a/, /i/ and /u/ were found to have the most distinctive tongue shapes in a given language, while the rest were found to have tongue shapes similar to one of the three. In the case of Mandarin Chinese, Bao [2] and Zhou [14] confirmed this. The reason to choose the high level tone was that we wanted to restrict the impact of larynx movement. (See [11] for larynx movement in tone production.)

2.3. Recording

The equipment used includes a portable GE Logiq-e ultrasound machine, an Allied camera, two Windows XP system computers, a Mackie 1402-VLZ3 Mixer, a M-Audio Luna Microphone and a M-Audio FireWire 410.The software for data recording and processing are UltraCap 1.3 (a custom software developed at UVic), Sony Vegas 8.0, ImageJ 1.41o, EdgTrak 1.0.0.2, Praat 5.2.16 and R 2.12.0. (See Magnuson and Coey in this volume for a description of these.)

The audio signal was collected by the Microphone to the mixer, shared by the two computers through the Fire Wire. On one computer, the ultrasound signal and the audio signal were simultaneously recorded by UltraCap 1.3 and saved as uncompressed *avi* files. On another computer, the jaw movement video was collected by the camera. This video and the audio signal recorded by UltraCap 1.3 were saved as *avi* files on the hard drive.

One participant at a time was recorded uttering each vowel five times. Before the recording of each token, the participant was asked to relax his/her tongue in its natural position. After the token was produced, the participant was instructed to wait for his/her tongue to return to its natural state before uttering the next vowel.

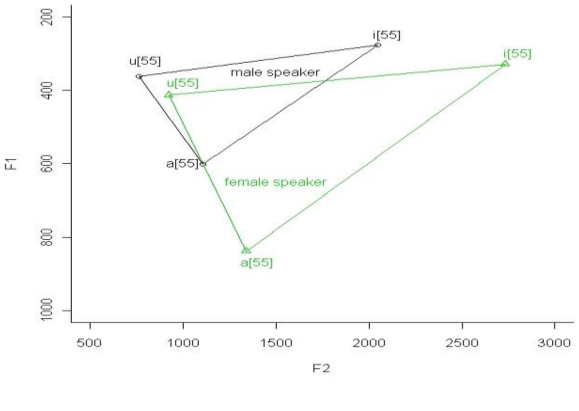
3. DATA PROCESSING

The recorded *avi* files were processed by Sony Vegas 8.0. In Vegas, we first aligned the ultrasound data with the jaw data and then segmented and saved each token of the three vowels into a separate *avi* file. For the five repetitions of each vowel, the middle three were used for further analysis. Next, we transformed each *avi* file's video frames into a sequence of *jpeg* files and extracted and saved its audio signals in a *wav* file. Thereafter, the audio data, the ultrasound data and lip-jaw data were processed separately.

3.1. Audio data

The Audio data were processed by Praat 5.2.16 and R 2.12.0. Out of each participant’s data, we first extracted the F1 and F2 values of the three vowels by Praat; then, we calculated the mean formant values by R and plotted them on a chart in Figure 1.

Figure 1: The participants’ formant charts of /a/, /i/ and /u/.



3.2. Ultrasound data

Having acquired the sequences of *jpeg* files of the tongue movement, we applied EdgeTrak to detect the tongue shapes and saved the results as *con* files and processed them in R.

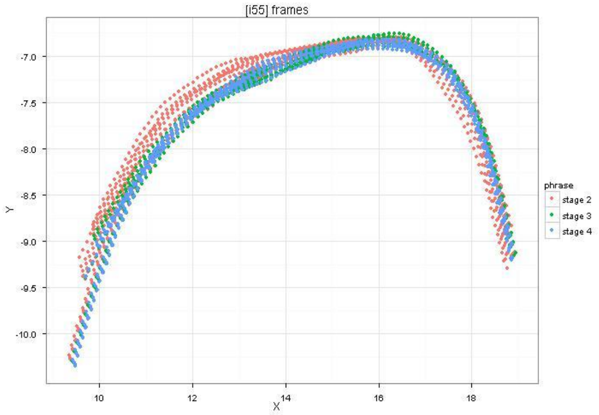
With ultrasound, one way to observe real-time tongue movement in the production of a vowel is to detect the variation among video frames. However, one problem with this method is that the different repetitions of the same vowel may not have the same number of frames; some repetition may have, say, 17 frames, whereas others may have 19. It is therefore hard to compare tongue movement among different items by frames. To address this problem, we divided the frames of each item into three stages (Stages 2, 3, and 4), and then compared these stages separately.

The data recording also collected transitional frames that occurred before and after each item. These transitional frames (Stages 1 and 5) can show information such as the neutral position of the tongue, movement prior to sound production and movement to a relaxed state following production. While this information is important in understanding the complete picture of tongue movement during the production of a vowel, we reference it here only to extract the tongue’s neutral shape to use as a reference for later comparison.

Frames between the beginning and end of phonation were categorized into Stages 2, 3, and 4. The division of frames into these stages was done as follows. In our data, there were roughly 18-20 frames for the male subject and 21-23 frames for the female subject for each token. We divided the number of frames for each token by three, and if the reminder was zero we evenly distributed the frames into the three stages sequentially. If the reminder was one or two, we allocated the remaining frames to Stage 3. Even though the one or two frames could potentially increase the variation from a statistical perspective, it would not influence the result significantly because the middle part of a vowel production was assumed to be more stable and experience less variation than other parts.

After allocating these frames into different stages, we used the *ggplot2* package in R to plot the contours from the frames for each item. Figure 2 displays the male speaker’s /i/ frames. (In all the figures below, the front of the tongue is to the right of the figure.)

Figure 2: The male participant's /i/ frames.



If we want to compare the three vowels in one plot, we still need to validate and calibrate the original data because there were many factors that could have influenced ultrasound data [13]. Among these factors, jaw movement may be the most prominent factor that should be corrected for. Ultrasonic tongue tracking needs to put a probe tightly under the chin. During the data recording, when the jaw goes down, the relationship between the probe and the jaw changes. In view of this, we video-recorded the jaw movement and used that information to calibrate the original ultrasound data as described in the next section.

3.3. Jaw movement data

Since ultrasound cannot record the jaw movement as X-ray can, we used an external camera to document jaw movement, and did a linear, two-dimensional transformation of the ultrasound data to correct for jaw rotation. During jaw data recording, we marked two points at the mandible (one around the angle of the mandibulae and the other around the mental foramen) and then calculated their displacement. By comparing the distances of these two points during a sound production to their neutral setting (when a sound is not being produced), we could calculate the rotation angle of jaw movement as well as the displacement distance. We assumed that the tongue aligned with the mandible and moved in a rotating way. With this basic assumption, we used the rotation angle to calibrate the original ultrasound data of /a/. We ignored the high vowels /i/ and /u/ because they involve trivial jaw movement.

The *jpeg* files of the jaw movement for /a/ were processed using ImageJ. To simplify data processing, we did not measure the distances of jaw movement in all the frames for each token, but only the middle five frames. We then calculated the mean distance of jaw movement and the rotation angle of the jaw. Finally, with this rotation angle, we calibrated both speakers' ultrasound data for /a/ using R.

4. RESULTS AND DISCUSSION

In order to answer the questions advanced at the beginning of this paper, we applied SS ANOVA among the vowels and the stages of each vowel, based on the calibrated ultrasound data.

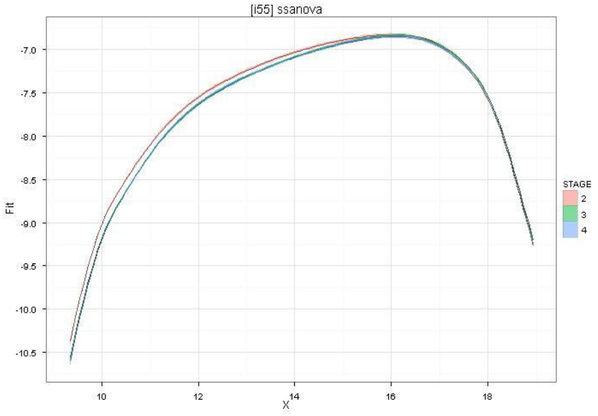
4.1. Tongue movement in producing /a/, /i/ and /u/

Iskarous [7] found that there were only two basic patterns of tongue movement involved in speech production: the pivot and the arch: the pivot pattern occurred when the tasks were at two regions in the vocal tract, while the arch pattern occurred when the tasks were at one region.

The SS ANOVA results for the tongue movement for the two speakers here were compatible with Iskarous' findings in that they exhibited a unified arch pattern. However, the extent of tongue movement in our data was dramatically smaller than that observed by Iskarous. Especially for the male speaker, the SS ANOVA results indicated that his tongue was more stable than that of his female counterpart during the production of vowels. Figure 3 shows the male participant's SS ANOVA results for /i/.

We also noted that, during the production of the three vowels, the two speakers' data showed a tendency toward less movement in the production of /i/ than of /a/, which in turn involved less movement than for /u/.

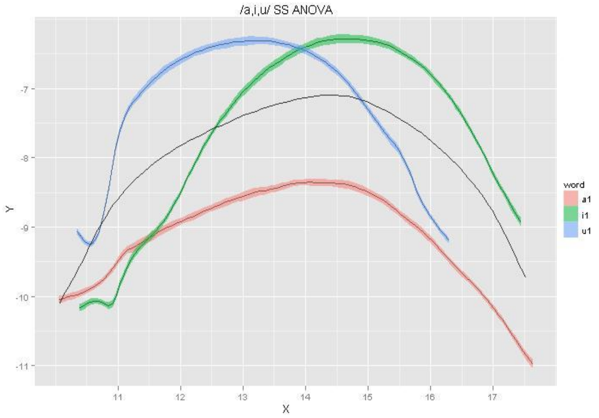
Figure 3: SS ANOVA results of /i/ of the male speaker.



4.2. Tongue configuration comparison of /a/, /i/ and /u/

Figure 4 displays the female speaker’s /a/, /i/ and /u/ tongue positions. In this figure, a reference curve (in black) indicating the tongue’s neutral shape is also included along the curves of the three vowels.

Figure 4: SS ANOVA results of /a, i, u/ of the female speaker.



For /a/, the tongue moved downwards without any dorsal prominence; for /i/, the front part of the dorsum ached while the radix went downwards to yield its configuration; to produce /u/, the back part of the dorsum humped and the apex descended relative to the tongue’s neutral position.

4.3. The articulatory working space

Based on the SS ANOVA results shown in the previous section, we standardized the articulatory working spaces of the two speakers and found that the articulatory space was roughly aligned with the formant chart. The only exception is the relationship between the male speaker’s /i/ and /u/. According to [8], the tongue height has a linear relation to F1: the higher the tongue, the higher the F1. However, the male speaker’s /i/ and /u/ in our analysis do not support this argument.

5. REFERENCES

[1] Baker, A. 2006. Smoothing Spline ANOVA. *NWAV* 35, Columbus, OH.

[2] Bao, H. 1984. Putonghua Danyuan de Shengli Jieshi. *Zhongguo Yuwen* 2, 117-125.

[3] Bressmann, T., Thind, P., Uy, C., et al. 2005. Quantitative three-dimensional ultrasound analysis of tongue protrusion, grooving and symmetry: data from 12 normal speakers and a partial glossectomee. *Clinical Linguistics & Phonetics* 19(6/7), 573-588.

[4] Davidson, L. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *J. Acoust. Soc. Am*. 120, 407-415.

[5] Fruehwald, J. 2010. SS ANOVA. *http://www.ling.upenn.edu/~joseff/*.

[6] Gu, C. 2009. gss: General Smoothing Splines. *http://www.stat.purdue.edu/~chong/software.html*

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

[8] Lindblom, B., Sundberg, J. 1971. Acoustical consequences of lip, tongue, jaw, and larynx movement. *J. Acoust. Soc. Am*. 50, 1166-1179.

[9] Magnuson, T., Coey, C. Submitted. *Synchronizing Video, Ultrasound and Audio with Water Balloon*.

[10] Mielke, J., Baker, A., Archangeli, D., Racy, S. 2005. Palatron: A technique for aligning ultrasound images of the tongue and palate. *Coyote Papers* 14, 97-108.

[11] Moisik, S., Lin, H., Esling, J.H. 2010. An investigation of laryngeal behavior during Mandarin tone production using simultaneous laryngoscopy and laryngeal ultrasound. *Proceedings of the 9th Phonetics Conference of China*, Nankai University, Tianjin, China.

[12] Parthasarathy, V., Stone, M., Prince, J.L. 2005. Spatiotemporal visualization of the tongue surface using ultrasound and Kriging (SURFACES). *Clinical Linguistics & Phonetics* 19(6/7), 529-544.

[13] Stone, M. 2005. A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics* 19(6/7), 455-501.

[14] Zhou, D., Wu. Z. 1963. *Putonghua Fayin Tupu*. Beijing: Shangwu Yinshuguan.