INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Individual Difference of Relative Tongue Size and its Acoustic Effects

Xiaohan Zhang, Chongke Bi, Kiyoshi Honda, Wenhuan Lu, Jianguo Wei*

College of Intelligence and Computing, Tianjin University, Tianjin, China

* khonda@sannet.ne.jp

Abstract

This study examines how the speaker's tongue size contributes to generating dynamic characteristics of speaker individuality. The relative tongue size (RTS) has been proposed as an index for the tongue area within the oropharyngeal cavity on the midsagittal magnetic resonance imaging (MRI). Our earlier studies have shown that the smaller the RTS, the faster the tongue movement. In this study, acoustic consequences of individual RTS values were analyzed by comparing tongue movement velocity and formant transition rate. The materials used were cine-MRI data and acoustic signals during production of a sentence and two words produced by two female speakers with contrasting RTS values. The results indicate that the speaker with the small RTS value exhibited the faster changes of tongue positions and formant transitions than the speakers with the large RTS values. Since the tongue size is uncontrollable by a speaker's intention, the RTS can be regarded as one of the causal factors of dynamic individual characteristics in the lower frequency region of speech signals.

Index Terms: relative tongue size, individual characteristics, cine-MRI, tongue movement velocity, rate of formant frequency transition

1. Introduction

The structure of the speech organs varies across speakers and thus emanates individual characteristics of speech sounds. Such characteristics are clearly observed in static acoustic measures, such as the fundamental or formant frequencies. In addition to the static characteristics, the speech organs may also derive subtle differences that are found in the dynamic patterns of speech signals, as seen in formant transitions for example. However, since the changes of formants in speech carry essential linguistic information, how to extract dynamic speaker characteristics is a challenging task. A hint to answer the question may be found in the anatomical structure of the speech organs that cannot be altered by a speaker's intention. One obvious case is the size of the tongue. The tongue volume is constant for each speaker, while it varies from speaker to speaker, thus eliciting certain individual differences in the changing speech signals as the tongue articulates. Based on this assumption, our earlier studies examined the relationship between the relative tongue size and tongue movement velocity [1, 2]. The relative tongue size (RTS) was proposed as an index for the individual difference in tongue size relative to the dimension of the surrounding structure. The earlier studies based on the combined cine- and tagged-MRI indicated that the smaller the RTS, the faster the tongue movement. Therefore, the RTS is a potential factor for generating dynamic individual characteristics.

A number of studies have investigated articulatory variability of the tongue in normal and pathological speech, yet not much is known about the relationship between the tongue size and acoustic characteristics. To analyze the relationship, the definition of tongue size is often critical. Tomaschek et al. [3] computed the area in the mid-sagittal plane covered by the tongue during articulation to reveal substantial between-speaker variation in speech rhythm. Iida-Kondo et al. [4] calculated the ratio of the tongue volume to the oral cavity volume as one of the measurable indices in a study on the obstructive sleep apnea syndrome.

Feng et al. [2] proposed our method to measure the tongue and vocal tract areas in the mid-sagittal plane above the level of the superior genial tubercle. The relative tongue size (RTS) was defined as the ratio of the tongue area to the total area of the tongue and oropharyngeal cavity. The RTS values thus obtained revealed the aforementioned relationship between the tongue and its articulatory dynamics, while acoustic analysis was not conducted.

In this study, we use cine-MRI and acoustic data to investigate the relationship between tongue movement velocity and formant transition. Magnetic resonance imaging (MRI) excels other classical imaging methods for observing the entire tongue tissue during speech production and providing a way for quantifying the structure of the articulators, including morphological features of speakers in conjunction with their acoustics characteristics. Real-time MRI [5] and synchronized sampling MRI [6] were the choices for the present analysis of tongue movement velocity. An LPC analysis was employed for formant tracking to compute the rate of formant frequency transition from the clean speech signal, and the results were compared between two speakers having small and large RTS values.

2. Materials and Methods

2.1. MRI data and speech samples

To investigate the relationship between tongue movement velocity and formant transition, we chose two female speakers, SC and ZC, as subjects, who indicated the smallest and largest RTS values, respectively, in the former study [2]. Both subjects are native Mandarin speakers from China, and none of them reported any history of speech or language disorders.

MRI and acoustic data for the analysis were chosen from the existing MRI dataset. The MRI dataset was collected using a Siemens Verio 3T MRI scanner at the ATR Brain Activity Imaging Center (ATR-BAIC) in Kyoto, Japan. In the scan sessions, each subject was instructed to lie supine on the platform of the MRI unit. The speech sounds were recorded for acoustic analysis during the scan and at separate post hoc sessions in a soundproof room. The clean speech corresponding to both cine-MRI data was down-sampled from 44.1 kHz to 10 kHz for analysis with linear prediction coding (LPC).

http://dx.doi.org/10.21437/Interspeech.2019-2452

2.1.1. *Real-time MRI for sentence production*

Real-time MRI (rt-MRI) uses ultrafast imaging protocols, which favors speech production research for characterizing the dynamic shaping of the vocal tract during speech production for any scan plane(s) of interest with no need for repeated scans [6]. In this study, each rt-MRI data contains 512 frames of images with a rate of 10 frames/second for a sentence production at a slow speaking rate. We selected the word /tàiyáng/ from the existing rt-MRI data to perform image and acoustic analyses.

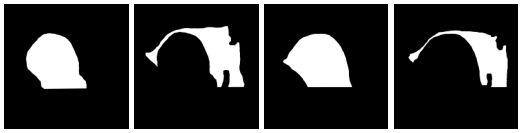
2.1.2. *Synchronized-sampling MRI for word repetition*

The synchronized sampling MRI (ss-MRI) was applied to record articulatory movements while the subject repeated a short utterance synchronized with the scan sequence [5]. In this study, each ss-MRI data contains 17 frames of images at a playback rate of 30 frames/second. Each subject lying supine in the MRI scanner repeated a short word. The utterances chosen for the analysis were two-syllable words /mìdù/ and /mùdì/. While hearing the trigger sounds, the subject repeated the utterances rhythmically to reproduce the same speaking rate as in the scanning experiment.

2.2. Calculation of the RTS

Before the image analysis, all the cine-MRI data were standardized to have the common coordinate system with the palatal plane as the abscissa. The palatal plane is defined by the line passing the anterior nasal spine (ANS) and posterior nasal spine (PNS) [1].

Feng et al. [2] proposed a method to measure the RTS based on the tongue and vocal tract areas on the mid-sagittal plane above the level of the superior genial tubercle. The RTS was defined on the static MRI data during vowel /i/ as the ratio of the tongue area to the total area of the tongue and oropharyngeal cavity. After defining the lower boundary of the tongue and the vocal tract, the binary images of the tongue and the vocal tract obtained by the manual extraction can be seen in Figure 1.



(a) (b) (c) (d)

Figure 1: *Tongue and vocal tract areas of each subject for vowel /i/. (a) SC tongue; (b) SC vocal tract; (c) ZC tongue; (d) ZC vocal tract.*

The equation (1) is used for computing the RTS in [2].

$$rel_{tongue} = \frac{A_{tongue}}{A_{tongue} + A_{VT}} \tag{1}$$

where A_{tongue} represents the tongue areas, A_{VT} is the vocal tract areas, and rel_{tongue} is the relative tongue size. The RTS values for the two subjects are shown in Table 1.

2.3. Image analysis method

To investigate the relationship between the RTS and velocity of the tongue movement (morphological changes of the tongue in articulation), it is necessary to choose features that represent the changes in tongue shape during utterances. For this purpose, the tongue movement velocity is presented by calculating the mean over a frame of the absolute difference in each pixel value (mean pixel rate of change, hereafter) between two adjacent frames [7]. The mean pixel rate of change is defined as:

Table 1: *Relative tongue size of two female subjects*

Subject	Relative tongue size
SC	54.17%
ZC	65.57%

$$\overline{P(t)} = \frac{\sum_{i,j=1}^{M} \left| p_{ij}(t+1) - p_{ij}(t) \right|}{M^2} \tag{2}$$

where $\overline{P(\text{t})}$ is the mean pixel rate of change of the current t frame, $P_{\text{ij}}(t)$ is the pixel value of the grayscale image of the t frame, M is the size of the image. In the real cases, and P is a binary image corresponding to a grayscale image.

Takemoto et al. [7] have shown that the mean pixel rate of change is a measure of the magnitude of articulatory movements. That is, when large articulatory movement occurs between two adjacent frames, the mean pixel rate of change increases. In contrast, the mean pixel rate of change decreases between the frames, when articulatory movement is small.

2.4. Trajectory of formant frequency

The acoustic theory of speech production states that vowel formants are resonances of the vocal tract, and tongue movement is also reflected by the changes in vocal tract resonance pattern. Thus acoustic consequence of tongue movement is observed in formant transition. What the RTS predicts is the correspondence between tongue movement velocity and the rate of formant frequency transition. Therefore, the analysis on the formant frequency transition is the most appropriate method.

In this study, we used an LPC-based formant tracking method with the Praat [8]. The parameters for speech sounds corresponding to rt-MRI data and ss-MRI data were the same, that is, a 10-kHz sampling rate, and 10th order of LPC with a Hanning window of a 25-ms length with a 10-ms shift length.

To ensure a smooth fit to the local slope of the trajectories of formant frequencies, the formant transition rate is obtained by using the delta parameter values to approximate the local slope over a finite length time window. The method for calculating the rate of formant frequency transition is inspired by the successful use of delta-based dynamic features in an application of speaker recognition systems [9].

3. Results and Discussions

3.1. Results for rt-MRI data

According to the method described in section 2, we know that subjects SC and ZC have the smallest and the largest RTS values, respectively. Figure 2 shows the mean pixel rate of changes for SC and ZC during the word utterance /tàiyáng/.

In Figure 2, the local minima in the time series correspond to the vowels' center frame, and the local maxima indicate the magnitude peaks of the articulatory movement velocities. We added markings for each vowel center. Observing the results,

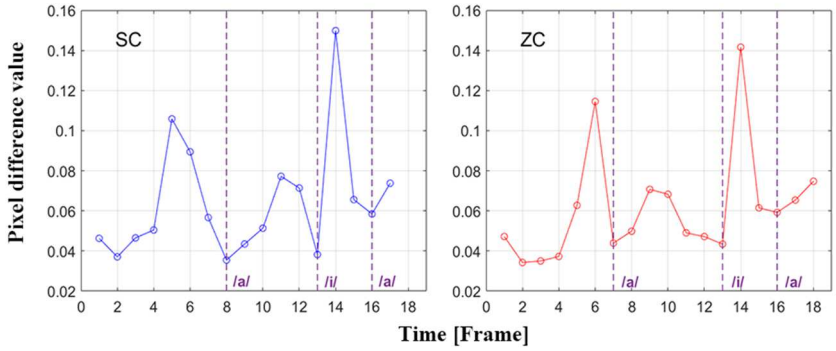
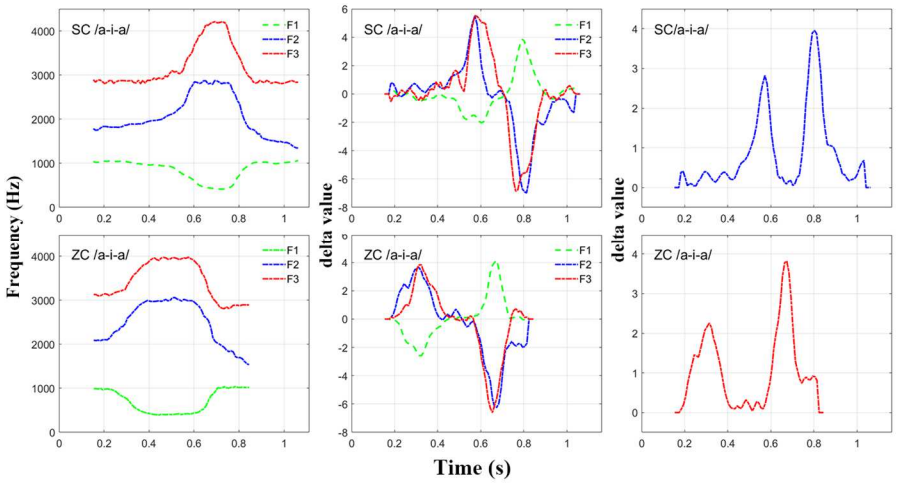


Figure 2: *Results of rt-MRI analysis showing time series of the mean pixel rate of change in /tàiyáng/. Dashed lines indicate the vowel's center frames.*



(a) (b) (c)

Figure 3: *Results of formant analysis. (a) Formant frequency transition, (b) Rate of formant frequency transition and (c) Total rate of formant transition for /aia/ in /tàiyáng/*

the peaks (between the vertical dashed lines excluding the first peak before /a/) of the mean pixel rate of change in the vowel-to-vowel transition in SC is larger than in ZC in the results in rt-MRI data. Accordingly, Figure 2 indicates that the articulatory movement velocities of SC are faster than that of ZC during both vowel transitions. In addition, although the tongue movement of ZC is earlier during the transition from vowels /a/ to /i/, the tongue movement velocity of ZC is still slower. Therefore, this result agrees with our earlier study [2], that is, the smaller the tongue, the faster the tongue movement even though the speaking rate is slower in the rt-MRI data.

Figure 3 shows the results from the acoustic analysis for the segment /aia/ in /tàiyáng/ for SC and ZC. The lineup point in the three panels corresponds to the vowel center frame of /a/. Panels (a) and (b) in Figure 3 show the formant frequency transition, the rate of formant frequency transition during the segment /aia/ in SC and ZC, where green, blue and red lines correspond to F1, F2 and F3 trajectories, respectively. Panel (c) gives the total rate of F1, F2 and F3 transitions after summing up the absolute values in Figure 3 (b). Between the two female subjects, the peak values of formant frequency transition are larger in SC than in ZC. In addition, the peaks of the tongue movement velocity in Figure 2 roughly correspond to the peaks of the formant transition rate in Figure 3 (b) or (c). Thus, the rate of formant transition is larger for the subject with the higher peak of tongue movement velocity. Overall, the results of rt-MRT data indicate that the smaller the relative tongue size, the higher the rate of formant frequency transition. The difference in the rate of F1 transition is not obvious in Figure 3 (b), which infers a certain contribution of jaw movement. The above results suggest that the RTS predicts temporal changes in F2 and F3 from one vowel to another.

3.2. Results for ss-MRI data

The synchronized sampling MRI data provides us more adequate motion images to analyze word utterances. In the data, the sequences of MRI and acoustic data are synchronized with each other because the clean speech was recorded with listening the MRI scan noise. Two Chinese words /mìdù/ and /mùdì/ were selected for the analysis, and the same set of analysis was conducted. Figure 4 shows the mean pixel rate of change for SC and ZC. To interpret the ss-MRI data, we added markings for the vocal-tract closure timing (for /d/) to the data from SC and ZC in Figure 4.

In the next, formant frequencies were calculated for each utterance. Confirming the accuracy of the formant tracking, we calculated the lower two formant frequencies for all the repetitions of the words /mìdù/ and /mùdì/. Observing all the formant frequency transitions, the token-to-token differences

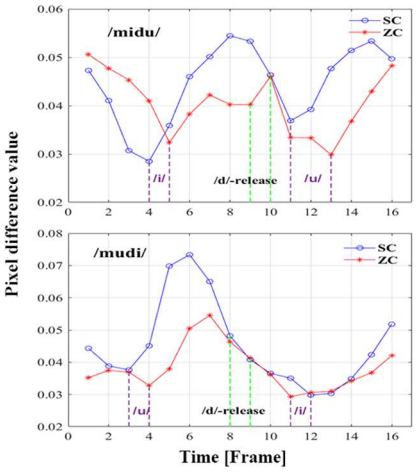


Figure 4: *Time series of the mean pixel rate of change in /mìdù/ and /mùdì/*

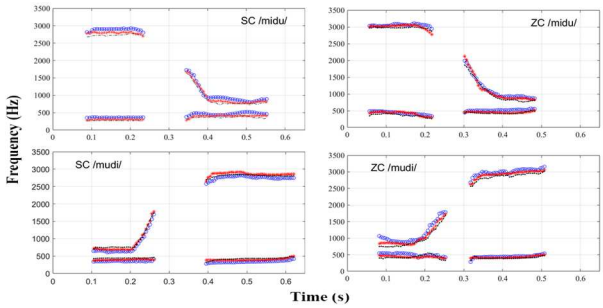


Figure 5: *The trajectories of the formant frequency in /mìdù/ and /mùdì/*

are found to be small. Therefore, the results from the three repetitions were used as the representative results as shown in Figure 5, which correspond to the green, red and black lines. In this figure, the primary differences seen in the formant frequency transition are F2 in both utterances /mìdù/ and /mùdì/. In addition, for the two female subjects, the slope of the formant frequency transition between vowels /i/ and /u/ in SC is larger than in ZC in the above results in both utterances.

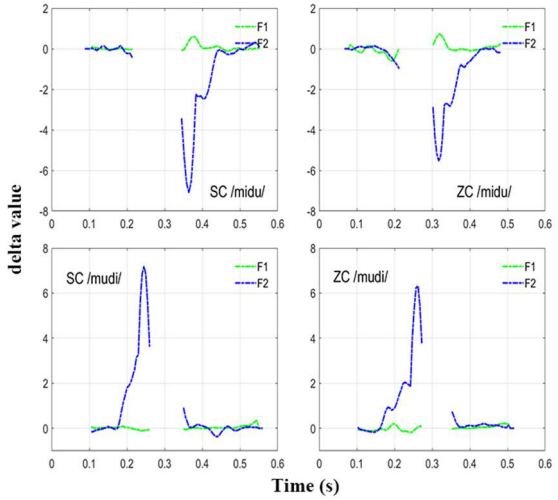


Figure 6: *The rate of formant frequency transition in /mìdù/ and /mùdì/*

Figure 6 presents the rate of formant frequency transition corresponding to the red line. Combining Figures 4 and 6 for the two female subjects, the peaks of the F2 transition rate occur after the word-medial consonant in utterance /mìdù/. In the utterance /mùdì/, the peaks of the F2 transition rate appear before the word-medial consonant. Moreover, the peaks of formant transition between /i/ and /u/ in Figure 6 indicate that the formant frequency transition of SC is faster than that of ZC before or after the word-medial consonants.

4. Conclusion

In this study, we examined two types of cine-MRI data and acoustic data to investigate the relationship between tongue movement velocity and the rate of formant transition. Calculating the mean pixel rate of change, the results showed a good agreement with our earlier studies, that is, the smaller the tongue, the faster the tongue movement. To confirm acoustic consequence of the individual difference of the RTS, speech signals were analyzed by computing the rate of formant frequency transition. The acoustic analysis showed a moderate tendency that the smaller the relative tongue size, the higher the rate of formant frequency transition. In a vocalic utterance of /aia/, the rate of formant transition in the subject with the smaller RTS is larger even though the subject's speaking rate is slower. In the word utterances with the paced speaking rate, the difference is more evident: the subject with the smaller RTS demonstrates the higher rate of formant transition before or after the word-medial consonants.

To summarize, the RTS is reflected by tongue movement velocity and thus relevant to the varied rate of the changes in vocal-tract resonance. This acoustic effect results in the differences of the rate of formant frequency transition. The result of this study agrees with what was speculated from speaker's tongue size and its range of movement limited by the space surrounding the tongue.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61573254, No. 61876131 and No. 61702360), and National Key Research and Development Plan (No. 2018YFC0806802).

6. References

[1] H. Bao, W. Lu, K. Honda, J. Wei, Q. Fang, and D. Dang, "Combined Cine- and Tagged-MRI for Tracking Landmarks on the Tongue Surface," *Proceedings of Conference of the International Speech Communication Association. Dresden, Germany: INTERSPEECH*, 2015, 359-363.

[2] X. Feng, W. Lu, J. Zhang, Y. Chi, and K. Honda, "Relative tongue size as an index to predict individual articulatory difference," In *Proc.10th Biennial Asia Pacific Conference on Speech, Language and Hearing (APCSLH 2017)*, Narita, Japan, September 17-19.

[3] F. Tomaschek, and A. Leemann, "The size of the tongue movement area affects the temporal coordination of consonants and vowels—A proof of concept on investigating speech rhythm," *The Journal of the Acoustical Society of America*, 144(5), EL410–EL416.

[4] C. Iida-Kondo, N. Yoshino, T. Kurabayashi, S. Mataki, M. Hasegawa, and N. Kurosaki, "Comparison of tongue volume/oral cavity volume ratio between obstructive sleep apnea syndrome patients and normal adults using magnetic resonance imaging," *Journal of Medical & Dental Sciences*, 2006, 53(2): 119-126.

[5] A. Toutios, and S. S. Narayanan, “Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research,” *APSIPA Transactions on Signal and Information Processing*, vol. 5, p. e6, 2016.
[6] 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*ournal of the Acoustical Society of Japan (E)*, 20(5), 375–379.
[7] 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,” *The Journal of the Acoustical Society of America*, 119(2), 1037.
[8] P. Boersma, and D. Weenink, “Praat: doing phonetics by computer (Version 5.3.76)”, http://www.praat.org/.
[9] F. K. Soong, and A. E. Rosenberg, “On the use of instantaneous and transitional spectral information in speaker recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(6), 871–879.