Timing differences in articulation between voiced and voiceless stop consonants: An analysis of cine-MRI data

Masako Fujimoto\textsuperscript{1}, Tatsuya Kitamura\textsuperscript{2}, Hiroaki Hatano\textsuperscript{3} and Ichiro Fujimoto\textsuperscript{4}

\textsuperscript{1}National Institute for Japanese Language and Linguistics, Tokyo, Japan
\textsuperscript{2,3}Faculty of Intelligence and Informatics, Konan University, Kobe, Japan
\textsuperscript{4}Department of Mechanical Systems Engineering, Takushoku University, Hachioji, Japan

\texttt{mfuj@ninja1.ac.jp, t-kitamu@konan-u.ac.jp, hatano.hiroaki@gmail.com, ifujimoto@ms.takushoku-u.ac.jp}

Abstract

Laryngeal and supralaryngeal articulators coordinate work to produce speech sounds. In order to study differences in supralaryngeal manifestations of voiced and voiceless consonants, we compared the tongue movement during a minimal pair /agise/ and /akise/ using the fast scanning techniques of MRI movies. The result showed that the tongue displacement starts earlier in /k/ than in /g/ for many of the speakers of Tokyo Japanese. This agrees with our previous findings using other dialect speakers. These results suggest that many Japanese actively differentiate supralaryngeal articulation according to the voicing of the consonants, raising the tongue earlier in voiceless ones. This movement is presumably to ensure the voicelessness of the consonant. The present study also supplies evidence for the usefulness of a constructive approach for physical modeling.

\textbf{Index Terms}: Magnetic resonance imaging, cine-MRI, stop consonant, voiced, voiceless, tongue movement

1. Introduction

Laryngeal and supralaryngeal articulators coordinate work to produce speech sounds. Although voicing contrasts of phonemes is primarily manifested in the larynx with alternating abduction and adduction of the vocal folds, supralaryngeal articulators may work with this laryngeal articulation in order to accomplish the voicing distinctions of consonants. Previous study using electropalatography (EPG) showed that the tongue contact area on the hard palate is longer and larger in voiceless consonants than in voiced ones \cite{1}, which would help voiceless consonant production. In a similar way, timing of the tongue-palate contact may be earlier during voiceless consonants, because faster contact would help to terminate the vocal folds’ vibration.

The fast scanning techniques from MRI (magnetic resonance imaging) provide opportunities to observe the vocal tract shape at each moment during speech production \cite{2}, and thus have allowed us to reproduce speech sounds by precise, physical acoustic modeling of the articulatory organs \cite{3}. The constructive approach using physical modeling is expected to complement conventional, descriptive studies in acoustic phonetics and reveal causal relationships between articulatory dynamics and actual speech sounds.

Our previous study using fast MRI showed that the tongue for Japanese speakers reaches the target point earlier during voiceless consonants than during voiced consonants \cite{4}. The subjects of that study are the speakers of Kansai, i.e. Kyoto-Osaka, dialects. The durational and articulatory characteristics in Kansai dialects possibly differ from those of Tokyo (standard) Japanese \cite{4-6}. Hence, cross dialectal evaluation is needed. In the present paper we examined the supralaryngeal articulation of the speakers of the Tokyo standard Japanese by comparing the tongue movement during the stop consonants /k/ and /g/ using fast MRI techniques.

2. Materials and methods

2.1. Subjects and speech task

Subjects are seven (six males and one female) adult speakers of Tokyo standard Japanese, between 23 and 43 years old. Test words are two nonsense words, /agise/ and /akise/. These stimuli are a minimal pair since they differ only by the voicing of the consonants /g/ and /k/ in the second mora. A cine-MRI movie per test word is made using 128 repetitions. We recorded other test words for some subjects, which are not included in this paper.

2.2. Cine-MRI data acquisition

Cine-MRI data of the mid-sagittal plane were acquired by a synchronized sampling method with external trigger \cite{7}. The speakers lay supine in the MRI gantry, wearing headphones to listen to a guide sound. The guide sound was a triplet-beat sequence consisting of one tone and two noise bursts. These sound bursts had durations of 100 ms, and the onset-to-onset interval was 400 ms. Speakers were asked to pronounce each nonsense word in synchrony with the guide sound matching the first and third morae to the second and third beats, respectively. Each speaker uttered each word a total of 128 times during scanning. The scanning started 65 ms after the first beat of the guide sounds. Prior to the MRI acquisition, the speakers practiced uttering the words to minimize the variation between repeated utterances to ensure sufficient quality of the cine-MRI data. (See \cite{8} for the detail of the cine-MRI)

The MRI acquisition was carried out using a 3T MRI scanner (MAGNETOM Vero, Siemens) at the Brain Activity Imaging Center of ATR-Promotions, Kyoto, Japan. The scanning parameters were as follows. Scan sequence: FLASH, frame rate: 100 frames/s, repetition time (TR): 10.0 ms, echo time: 1.62 ms, flip angle (FA): 15°, field of view: 256 times 256 mm, matrix size: 256 times 256, slice thickness: 4 mm, without averaging.

2.3. Image processing method

The trace of the tongue surface, the highest point of the tongue surface and the trajectory of its displacement was measured from the cine-MRI data by the method of Tachibana et al. \cite{4}.
The tongue surface was first extracted from each frame of the cine-MRI data by Canny’s edge detection method [9] and was traced using a fifth-order polynomial approximation as illustrated in Figure 1. The velocity of the highest point of the tongue (illustrated by a small circle in the Figure 1) was calculated from the obtained surface shape, and the vertical movement of the point was tracked. The trajectory of the highest point of the tongue was smoothed by a sixth-order moving average and resampled at a rate of 1000 frames/s by spline interpolation. The vertical velocity was computed from the up-sampled data. Since the tongue rises toward the palate most rapidly during /g/ and /k/, the velocity increases rapidly and ceases when it reached the target position of the consonants. Hence, we defined the timing of the target position of /g/ and /k/ at the point when the velocity decreases to (nearly) zero after hitting the peak speed as shown in Figure 2. Hereafter, these timings are referred to as the target point for /g/ and /k/, and denoted by \( t_g \) and \( t_k \), respectively.

![Figure 1: Example of tongue surface tracking. The tongue surfaces in /akise/ and /agise/ production are shown as green (gray) and blue (black) lines, respectively. The black circle indicates the highest position of the tongue surface for /akise/.

![Figure 2: Example of the trajectory of the tongue movement velocity. Arrows show the target point for /g/ (\( t_g \)) and /k/ (\( t_k \)).

3. Results

3.1. Displacement of tongue surface shape

Figure 3 shows selected frames of cine-MRI data. During /a/ at the 33rd frame in the top row, the tongue position as well as its shape, is similar in /agise/ and /akise/. However, at the 48th frame in the bottom row, the tongue is more raised during /k/ in /akise/ than during /g/ in /agise/. The result suggests that the tongue is raised earlier during voiceless consonant.

Figure 4 shows selected frames of tongue traces of /a/, /k,g/, /i/, /s/ and /e/ production, respectively. Tongue surfaces of /agise/, in green (gray), and /akise/, in blue (black), are superimposed. At the 45th, 48th and 53rd frames, the traces of /k/ locate higher than those of /g/, whereas both traces stay at a similar position in other frames. This result indicates that the tongue displacement in the two tokens differs only by the voicing contrast during /g/ and /k/. Duration of the tongue being raised for /k/ is longer than that for /g/, since the tongue positions for /agise/ and /akise/ are the same in /i/ at the 63rd frame.

![Figure 3: Selected frames of cine-MRI (speaker M1). Frames during /agise/ (left) and /akise/ (right) are shown.

![Figure 4: Selected frames of tongue surface tracking (speaker M1). The tongue surfaces in /akise/ and /agise/ are superimposed in blue and green lines, respectively. The black circle indicates the highest position of the tongue surface.

956
3.2. Timing of tongue peak

Figure 5 shows the trajectory of vertical tongue-peak movement for seven speakers. The origin of the time axis was set to the beginning of the first beat of the guide sound and the origin of the height axis was set to the neutral rest point of the tongue. As seen in Figure 5, the tongue is first lowered towards /a/, raised toward /k/ or /g/, stays at a relatively higher position during /is/, is lowered again towards /e/ and returns to the rest position. This pattern is quite similar for both /agise/ and /akise/ for all speakers. However, critically, the timing to start raising the tongue and to reach the target point for /k/ is earlier than for /g/ in five out of seven speakers. In the case of speaker M1 in the top most Figure 5, the tongue reached the target point at 585 ms in /agise/ and 535 ms for /akise/ as shown. Also, the trajectories of tongue-peak displacement merge after the target point for /g/, suggesting that the duration of the tongue being raised for /k/ is longer than that for /g/. However, this tendency shows more interpersonal variation.

Figure 5: Displacement of the vertical tongue-peak movement of the speakers M1 to M6 and F1 (from top to bottom). Green line shows /agise/ and blue, /akise/. Numerals in the upper left denote the time when the tongue reached the target point of /g/ and /k/.

Table 1. Differences in timing between voiced and voiceless consonants when the tongue reaches the target point.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>( t_{k} ) (ms)</th>
<th>( t_{g} ) (ms)</th>
<th>( t_{g} - t_{k} ) (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>535</td>
<td>585</td>
<td>50</td>
</tr>
<tr>
<td>M2</td>
<td>604</td>
<td>602</td>
<td>-2</td>
</tr>
<tr>
<td>M3</td>
<td>479</td>
<td>504</td>
<td>25</td>
</tr>
<tr>
<td>M5</td>
<td>602</td>
<td>598</td>
<td>-4</td>
</tr>
<tr>
<td>M6</td>
<td>567</td>
<td>591</td>
<td>24</td>
</tr>
<tr>
<td>M7</td>
<td>534</td>
<td>560</td>
<td>26</td>
</tr>
<tr>
<td>F4</td>
<td>553</td>
<td>567</td>
<td>14</td>
</tr>
<tr>
<td>mean</td>
<td>553.4</td>
<td>572.4</td>
<td>19.0</td>
</tr>
<tr>
<td>SD</td>
<td>40.3</td>
<td>31.4</td>
<td>17.2</td>
</tr>
</tbody>
</table>

Table 1 summarizes the timing when the tongue reached the target point of /g/ and /k/. For five out of seven speakers, \( t_{k} \) precedes \( t_{g} \), i.e., the tongue reached the target point earlier in /k/ than in /g/. For the remaining two speakers, the timing between \( t_{k} \) and \( t_{g} \) are virtually identical. Hence, interspeaker variation appeared as regards the manifestation of tongue movement during voiced and voiceless consonants. A difference between male and female is not found among the
present speakers. The average time difference that \( t/k \) precedes \( t/g \) for seven speakers is 19.0 ms (SD 17.2 ms). When counted only those whose timing for \( t/k \) is earlier than \( t/g \), the difference is 27.8 ms (SD 11.9 ms). These results suggest that many speakers of Tokyo standard Japanese differentiate the timing of supralaryngeal articulation by the voicing contrast of the consonants. They do so in a way to raise the tongue earlier for the voiceless consonants than the voiced cognates.

4. Discussion

For five out of seven Tokyo speakers, the tongue began to rise and reached the target point earlier for \( t/k \) than for \( t/g \). This is the same tendency found for all four Kansai speakers in Tachibana et al. (2012). This occurred even though the speakers were instructed to pronounce the words with exactly the same timing as the guide sound. These results indicate that many Japanese tend to initiate moving the tongue towards the palate earlier in voiceless consonants than in voiced ones. This is presumably in order to cooperate with laryngeal manifestation movements in producing voiceless consonants.

It has been reported that the air pressure in the oral cavity during stop closure is higher in voiceless consonants than in voiced ones [10-12]. This difference may be primarily due to the amount of air supplied into the vocal tract, since the vocal folds abduct during voiceless consonants to stop vibration, whereas they are kept closed during voiced consonants [13]. Supralaryngeal gestures also differ depending on voicing of consonants. As noted above, the tongue contact area on the hard palate is longer and larger in voiceless consonants than in voiced ones [1], which would work to increase the air pressure in the oral cavity. The current finding of tongue movement for the preceding \( k / \) suggests this contributes to increasing the air pressure for voiceless stops.

Although the materials are limited, our study provides a step toward investigating detailed articulatory movements using MRI measurements. To evaluate the present findings, further examinations involving more subjects and other consonant pairs such as \( /t/-/d/ \) and \( /p/-/b/ \) are necessary. Cross-linguistic examinations are also desirable.

5. Conclusions

In the present study, we compared the tongue movement of the voiced consonant \( /g/ \) and voiceless consonant \( /k/ \) using fast MRI movies. The results showed that the tongue is raised earlier in \( /k/ \) than in \( /g/ \) for many of the speakers. This agrees with our previous findings. These results suggest that many Japanese speakers differentiate supralaryngeal articulation according to the voicing of the consonants, and raise their tongue earlier for voiceless ones. Faster closure helps to ensure the voicelessness of the consonant for voiceless stops.

The present study also supplies evidence for the usefulness of a constructive approach for physical modeling. It complements the conventional, descriptive studies in acoustic phonetics to reveal causal relationships between articulatory dynamics and actual speech sounds.

6. Acknowledgements

This study was supported in part by JSPS KAKENHI (21300071 and 23520539) and MEXT, Japan.

7. References