Video recordings of L1 and L2 jaw movement: effect of syllable onset on jaw opening during syllable nucleus

Yusuke Abe  s1170175  Supervised by  Prof. Ian Wilson

Abstract

Video is a convenient, inexpensive method of recording data for jaw movement during speech. However, when using markers attached to the chin, it is possible that the data will not represent actual mandible motion, because of the skin stretching over the mandible - especially true for labial consonants. In this study, we made video recordings of L1 and L2 speakers of English saying 5 trials of 34 sentences each, and we automatically measured the distance between paper markers attached to the chin and glasses. We compared jaw opening during syllable nucleus for syllables with and without labial onsets, for L1 and L2 English speakers of various proficiencies. Although speakers must stretch the lower lip upwards for a labial constriction, preliminary results show that there are no statistically significant differences for any speaker’s jaw opening during the nucleus of non-labial- versus labial-onset syllables. There is also very little intra-subject variation in the metrical structure (as measured by jaw opening) for a given sentence across trials. However, across-trial variability in the time between jaw movement peaks is a lot less for L1 than for L2, presumably because these L2 speakers have not yet mastered the metrical structure of English.

1 Introduction

1.1 Background

Prosody, including rhythm, may be more important than segmental pronunciation for overall intelligibility of speech [1]. Languages vary in their segmental and prosodic structures; second language learners must learn articulation of these two types of non-native structures in order to communicate in a multilingual environment. Previous research has shown that poor English rhythm by L2 speakers makes speech difficult to understand by native listeners [2].

Japanese speakers of English have trouble with English rhythm because the two languages have different metrical structure.

The amount of jaw opening is directly related to English rhythm: sentence stress increases the amount of jaw opening, no matter what the vowel quality [3]. So, (inexpensive) video of jaw movement could provide good articulatory feedback to a learner if the video accurately reflects mandible movement.

Eventually, the purpose is the making mobile phone feedback system for L2 learners to learn another language. However, it is possible that mouth opening is affected by skin stretching especially labial consonant when the video recorded is used. It is not mandible movement. We can see just skin movement in the video, so the video recording is needed to prove the value as the first step for making feedback system project.

Since stress and rhythm are salient in vowels, I test whether or not jaw position is different for vowels with and without a labial consonant in the syllable onset.

1.2 Objectives

My study is a first step to make the pronunciation feedback system for L2 English learners using mobile phone, so I have two objectives. First, determine feasibility of using video to measure jaw (i.e., mandible) movement during vowel production. Do this by comparing jaw opening during syllable nucleus for syllables with and without labial onset. Second, if the using video recording to infer jaw position across different tokens of the same vowel is acceptable, compare L1 and L2 English speaker’s jaw movements and find the difference between them.

1.3 Problem definition

When we record the jaw movement using the video, the problem is that video data shows us skin, not the mandible, and the skin stretches over the mandible during production of some speech sounds (especially labial consonants which means those using the lip during pronunciation like “b”, “p”, “f”, “v”, “m”). Although speakers must stretch the lower lip upwards for a labial constriction, preliminary
results show that there are no statistically significant differences for any speaker’s jaw opening during the nucleus of non-labial- versus labial-onset syllables.

1.4 Present study

Video is a very useful data collection method because it is inexpensive and convenient for recording. However, if we want to measure jaw lowering, video data shows us the skin, not the mandible. Since the skin stretches over the mandible during speech, we need to test how feasible video is for measuring a given vowel in different contexts. I especially need to test vowels in labial consonant contexts. I compared jaw opening during syllable nucleus for syllables with and without labial onsets, for L1 and L2 English speakers of various proficiencies.

2 Method

Subjects were video recorded reading stimuli from a computer screen. Jaw lowering was measured by tracking markers attached to each subject’s glasses and chin. After that the mouth opening percentage was calculated, and the maximum percentage of the nucleus for labial and non-labial onset syllables was found. After that I compared each other if there are significant differences using t-tests.

2.1 Subjects

I collected data from 2 male native speakers of American English (A1 from Indiana and A2 from Washington) and 5 Japanese speakers of L2 English (J1 – a female teacher of English, J2 – a female sophomore student, and J3-J5 – three male freshman students). A1, A2 and J1 were all faculty members of university of Aizu, and J2-J5 were all students at the same university. The 5 Japanese subjects were chosen to represent a range of English proficiency levels.

2.2 Apparatus

A tripod-mounted Panasonic HDC-TM750 digital video camera was used to collect video of the front of the face. Light from two 300W halogen bulbs (LPL-L27432) was reflected onto the face to improve automatic maker tracking. Video files were first converted from MTS format to uncompressed AVI using “ffmpeg”, an open-source command line tool.

The two blue markers were then automatically tracked using a previously tested program written in C with OpenCV by Horiguchi (see [3]).

2.3 Stimuli

Stimuli consisted of 5 blocks of 34 sentences each (for a total of 170 sentences). Each block contained a different ordering of the 34 sentences. Stimuli were presented to subjects using PowerPoint on a laptop computer positioned about 2m away slightly below eye level. From the 34 sentences, I chose four to focus on in this analysis: (1) Yes, I saw five bright highlights in the sky tonight, (2) Yes, I saw nine tight highlights in the sky tonight, (3) He sees 3 lean teepees ’neath the seaweed, (4) He sees 3 lean Genies ’neath the seaweed. The sentence (1) and (3) include labial onset (“f” of five and “br” of bright from sentence (1), “p” of teepees from sentence (3)), and the sentence (2) and (4) include non-labial onset (“n” of nine and “t” of tight from sentence (2), “n” of Genies from sentence (4)). I compared sentence (1) and (2), (3) and (4) each, especially the vowel after the labial onset and non-labial onset syllables.

2.4 Video data collection

All 7 subjects were video recorded in the same setup in the CLR Phonetics Lab at the University of Aizu. Two blue markers were attached to each subject: one between the eyes on the frame of a pair of glasses and the other attached to the front of the chin. Audio and video were collected using the setup described in 2.2. For each subject, I started by collecting images where the jaw was maximally open and maximally closed. Thus, I could express any mouth aperture as a percentage of fully open.

Figure 1: Marker Position
2.5 Data analysis

First of all, the distance between two blue markers was calculated with C program. The program extracted and imported RGB images from the collected video, converted them into HSV color space, binarized each channel with given parameters (H: 165-15, S: 180-255, V: 180-255), split and incorporated the channels, filtered the data with Gaussian, Otsu method, area and circularity, and finally calculated and displayed the centroid of each marker. Secondly, the mouth opening percentage was calculated as the maximum opening was 100% and the maximum closed was 0% using Excel. After that, the maximum percentage of the nucleus for labial and non-labial onset syllables. Finally, t-test was used to compare the vowels with and without labial onsets for sentences (1) and (2), (3) and (4).

3 Results

3.1 Result 1

The results of the t-test, shown in the table below, were calculated with Excel. There are two sentences each which I want to compare, so an unpaired t-test was used. This t-test means if these numbers are higher than 0.05, there is no significant difference.

<table>
<thead>
<tr>
<th>Subject</th>
<th>five vs nine</th>
<th>bright vs tight</th>
<th>teepeek vs Genie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 1 (L1)</td>
<td>0.30</td>
<td>0.83</td>
<td>0.82</td>
</tr>
<tr>
<td>Subject 2 (L1)</td>
<td>0.054</td>
<td>0.74</td>
<td>0.63</td>
</tr>
<tr>
<td>Subject 3 (High-level L2)</td>
<td>0.61</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Subject 4 (Mid-level L2)</td>
<td>0.09</td>
<td>0.19</td>
<td>0.98</td>
</tr>
<tr>
<td>Subject 5 (Low-level L2)</td>
<td>0.19</td>
<td>0.67</td>
<td>0.61</td>
</tr>
<tr>
<td>Subject 6 (Low-level L2)</td>
<td>0.53</td>
<td>0.63</td>
<td>0.06</td>
</tr>
<tr>
<td>Subject 7 (Low-level L2)</td>
<td>0.91</td>
<td>0.67</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Table 1: T-tests

For the native speakers and all but the low level L2 learners, there were no statistically significant differences between jaw opening for a vowel with and without a labial onset. For low-level learners, there was a significant difference for some pairs of words. This may be due to the higher degree of variation in articulation when low-level L2 learners speak.

3.2 Result 2

The following figures show the part of all results for the sentences “Yes, I saw five bright highlights in the sky tonight” and “Yes, I saw nine tight highlights in the sky tonight” as spoken by the American English speaker, and high level and low level L2 speaker. These figures indicate two things. First, we can see the percentage of jaw opening during syllable nucleus for syllable with and without labial onset used t-test in 3.1. Second, we can see the difference of jaw movement shape.

![Figure 2: Data collection Setting](image)

![Figure 3: Plot of jaw movement for native English speaking subject A1, sentence (1)](image)
Now looking at the data for A1 (fig.3 and fig.4), the percentage of jaw opening during syllable nucleus for syllable with and without labial onset is almost same, so the t-test showed no statistically significant difference. In addition, these two figures example is a part of 5 data (trials) each sentences per subjects. I could see the same result from the sentence(3) and sentence(4).

High-level L2 speaker’s jaw movement shape is almost same as native speakers when we compare the figure 3 and the figure 4. The jaw movement shape shows English rhythm, so these graphs could prove this high-level L2 speaker and native speaker are very close pronunciation.

For low-level L2 speaker (fig.5), almost all words are emphasized to pronounce because he is not familiar with English rhythm. In addition, Japanese and English have different metrical structure. Also, he has higher degree of variation in articulation. It is easy to see the difference and to compare between low-level L2 speaker and native speaker, high-level L2 speaker. For low-level L2 speaker, “tonight” is more open than “sky”. However, for native speaker and high-level L2 speaker, “sky” is more open than “tonight”. Then, for high-level L2 speaker and low-level L2 speaker pronounce “I” emphatically.

5 Discussion

The skin stretching does not effect on mouth opening during vowel production with and without labial onset. However, it’s not to say the speech is always same articulation. We can see that mouth opening percentage of some words is little difference like “sky” in figures of result 2. In figure 3, mouth opening percentage is about 32%, but figure 4 shows about 21% because this may be due to the stress patterns. In addition, the mouth opening percentage should be 0% during labial onset because lips are closed. This may be due to the head was moving and head angle was changed, but I could not the head keep in place because the important thing is natural speech. Thus, I have to think about the method to solve this.

6 Conclusions

Video recording is good to look the shape of articulation. It can compare between low-level L2 speaker and high-level L2 speaker, native speaker because the mouth opening difference is different clearly.
I have determined that the type of syllable onset (labial versus non-labial) does not affect chin marker position across different occurrences of the same vowel.

Since there is no correlation between the skin-stretching for labial onset consonants and the jaw position for following vowels, I conclude that using video recording to infer jaw position across different tokens of the same vowel is acceptable.

7 Future Research

Collect electromagnetic articulography (EMA) data and use the lower tooth marker and chin marker to determine exactly how much the skin stretches over the jaw for many different vowels and consonant. This will indicate exactly how much can be inferred about the jaw movement in many contexts.

Increase the number of video subjects for both L1 and L2 English. Collect video data in the same setting using L2 speakers of English from a different L1 language background (e.g., Chinese speakers of English).

Eventually create a feedback system on a mobile phone, for example, where L2 learners can get instant real-time feedback on their jaw movements during L2 English production.

8 Acknowledgements

Thanks to all subjects at the University of Aizu. Thanks also to Dr. Donna Erickson (Showa Music University) for her advice and Mr. Naoya Horiguchi for his advice and programming skills.

References

