Articulating rhythm in L1 and L2 English: Focus on jaw and F0

Ian Wilson[1], *Donna Erickson*[2], *Naoya Horiguchi*[3]

[1,3]Center for Language Research, University of Aizu, Aizuwakamatsu, Fukushima, Japan
[2]Showa Music University, Kawasaki, Kanagawa, Japan
wilson@u-aizu.ac.jp, ericksondonna2000@gmail.com, m5141147@gmail.com

Articulatory/acoustic recordings were made of two native speakers of American English and five Japanese speakers of English. Jaw displacement measurements and average F0 were made for each syllable in the utterances. Patterns of jaw displacement and F0 were systematic for the native speakers, but those for the Japanese speakers varied. Evaluations by American university students as to how native-like the Japanese-English utterances sounded showed groupings of skill levels, which corresponded well to the observed patterns of jaw movements and F0. Future work along these lines will explore how to apply these findings to improved teaching of spoken English to Japanese learners of English. The findings of this study could also be applied to creation of more realistic avatars/talking heads.

Index Terms: rhythm, jaw, F0, L2, English, Japanese

1. Introduction

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. With regard to prosody, rhythm and intonation are two key ingredients. Intonation is patterns of F0 changes, while rhythm is more difficult to define. According to Kohler [1], rhythm is the patterning of syllable prominences and the grouping (or chunking) thereof over time. There are said to be three types of rhythm systems of languages: stress-timed languages, such as English and other Germanic languages; syllable-timed languages, such as Spanish and many Romance languages; and mora-timed languages, such as standard Japanese [2]. According to studies by Ramus, Nespor and Mehler [3] and Low, Grabe and Nolan [4], the rhythm of a language can be described in terms of interval-based rhythm metrics; yet, this is not universally agreed upon [5].

Articulatory (EMA, Xray Microbeam) studies of rhythm suggest that an articulatory component of rhythm may be alternating strong and weak jaw movement patterns with consequent changes in F1, and that these alternations may reflect the metrical structure of spoken English [6, 7]. However, these studies were based on data from only a few speakers, due to the difficulty in using EMA or X-Ray Microbeam.

In this paper we examine the hypothesis that articulation of prosody is related to jaw movement. We use video recordings to measure jaw opening for each syllable in the utterance for a larger number of speakers. We also measure average F0 of each syllable, since language prosody involves both F0 changes and rhythm (stress) changes.

We compare native speaker patterns of F0 and jaw movement with those of non-native speakers. A secondary hypothesis is that the jaw-F0 patterns for native speakers is relatively consistent, and that these patterns vary for non-native speakers according to their perceived skill level in spoken English.

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. A selection of the audio data was played for American college students to judge native-likeness of the subjects. Pitch was measured using Praat software.

2.1. Subjects

We collected video 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 freshmen students). A1, A2, and J1 were all faculty members of a prefectural university in Fukushima, Japan, 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 marker tracking. Video files were first converted from MTS format to uncompressed AVI using *ffmpeg*, an open-source command line tool. The two blue face markers were then automatically tracked using a previously tested program written in C with OpenCV by the third author [8] (see figure 1 and section 2.3.4 for details).

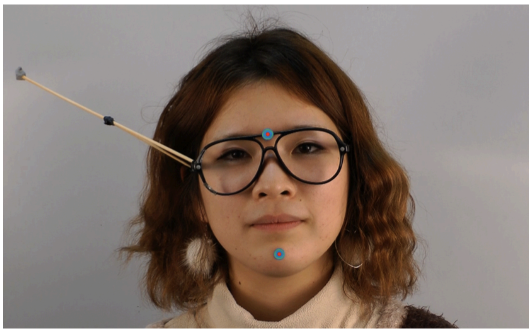

Figure 1: *Position of blue markers (red dots indicate automatically calculated marker centroids).*

2.3. Procedure

2.3.1. 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, we chose two to focus on in this analysis: (1) *Yes, I saw five bright highlights in the sky tonight*, and (2) *He sees 3 lean teepees 'neath the seaweed*. Both these utterances have a similar rhythm pattern with three phrases (or possibly four in (1) if *Yes* is counted separately). In choosing stimuli, one constraint was that each syllable in a sentence must have the same vowel, since jaw displacement changes greatly depending on vowel height, i.e., for a low vowel, the jaw opens much more than for a high vowel. Note that in sentence (1), other than the word *Yes*, every content word has a diphthong or low vowel, and in (2), every content word has a high front vowel.

2.3.2. L2 audio and 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, we started by collecting images where the jaw was maximally open and maximally closed. Thus, we could express any mouth aperture as a percentage of fully open.

2.3.3. Native listener judgment of L2 audio data

In order to test our secondary hypothesis that jaw patterns vary with the perceived proficiency of the speaker, we needed to run a native listener judgment task. Using Runtime Revolution software, a judgment task was constructed. The audio from the third repetition (for L2 speakers) and the second repetition (for native speakers) of the two sentences given in 2.3.1 was extracted and scaled to 65dB. In addition, recordings of the same sentences from 3 Japanese speakers of English from a separate (EMA) experiment were also used (9 speakers x 4 utterance types x 3 repetitions, plus, 1 speaker x 2 utterance types x 3 repetitions = 114 presentations).

The instructions to the native listeners were as follows: "When you click the start button, you will hear a sentence. Please use the slider to indicate how much like a native English speaker this person sounds. You can keep clicking the same sound as many times as you want to in order to double check your impressions. To listen to the sound again, please click the button at the top of the next screen."

Twenty American college students from a Midwestern university in the United States judged how native-like the utterances sounded. In looking at the results, one listener seemed to have moved the slider in the opposite direction, giving one of the native speakers a judgment close to 0, not close to native-like, so that listener was eliminated. Thus, there were a total of 19 listener judges.

2.3.4. Data analysis

Using the C program mentioned in 2.2, we automatically calculated the position of the markers throughout all sentences of interest. 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.

Jaw aperture measures (Euclidean distance between the markers) was calculated and expressed as a percentage of maximum jaw opening. These measures were then plotted over time and the peaks/valleys were checked against the audio to determine their corresponding syllables. The average F0 of each syllable in each utterance was calculated using Praat software, and both F0 and jaw movement were plotted together using Excel.

3. Results & discussion

The results of the judgments, shown in the table below by the university students, are in agreement with the impressionistic judgments of the first two authors, professors of phonetics in Japan.

Table 1. Judgments of native-likeness by 19 American university students. native American English speakers (A1, A2); Japanese speakers of English (J1~J5) for the *highlights* sentence and the *teepees* sentence.

Speaker	Highlights	Teepees
A1	89%	91%
A2	66%	57%
J1	53%	68%
J2	37%	33%
J3	28%	23%
J4	29%	22%
J5	21%	13%

The following figures show the acoustic and articulatory results for the sentences "Yes, I saw five bright highlights in the sky tonight" and "He sees three lean teepees 'neath the seaweed" as spoken by the American English speaker (A1), and the "best" (J1) and "worst" (J5) Japanese speakers of English, in terms of the judgments by American listeners shown in table 1. The average F0 for each of the vowels is shown by the connected squares in the upper part of the graph; the amount of jaw opening for each of the vowels, by the connected circles in the lower part of the graph. Since an acoustic indicator of stress is high F0, and an articulatory indicator of stress is a lower jaw position, if subjects use both ways to stress a syllable then we would expect to see the red line (F0) rising as the blue line (jaw position) lowers. The words associated with each of the F0/jaw measurements are shown below the jaw movement tracings. The temporal line-up point for each of the graphs is at "five" or at "three", since these are the starts of the middle phrase in each of the sentences.

Looking first at the data for A1 in figure 2, "Yes, I saw five bright highlights in the sky tonight", we notice that (1) the F0 pattern and the jaw opening pattern are different (i.e., they do not always spread away from each other as we would expect if they both functioned together as stress indicators, (2) the F0 pattern shows a peak on the word "saw" and then a gradual declination until the end of the sentence ("night" had no measureable F0), (3) the jaw pattern shows an opening for each of the key words in the sentence, (4) even though all the vowels are the same (/aJ/) and presumably should show the same amount of jaw opening, they show different amounts of jaw opening, and (5) there seems to be a pattern of jaw opening of alternating strong-weak (or vice-versa) openings, which reportedly reflect the prosodic structure of the utterance [7]. Although not shown here, the pattern of jaw movement and F0 for the second American English speaker (A2) is quite similar to that of A1.

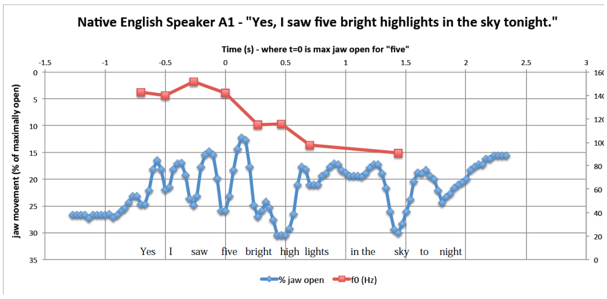

Figure 2: *Plot of F0 (in red) and jaw movement (in blue) for native English speaking subject A1, sentence (1)*

Now looking at the data for J1 (fig. 3), the relatively good speaker of English, we see (1) a gradual declination of F0 throughout the sentence. However, (2) there is an alternating rhythmic pattern of F0 and jaw opening such that for each word high F0 is associated with a large jaw opening, and low F0 with a smaller jaw opening. It is as if this good speaker of English is using both F0 and jaw opening to produce a rhythmic pattern of English.

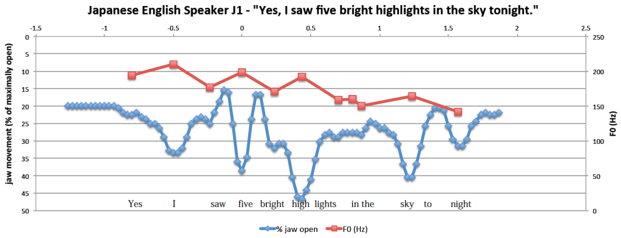

Figure 3: *Plot of F0 (in red) and jaw movement (in blue) for advanced L2 speaker J1, sentence (1)*

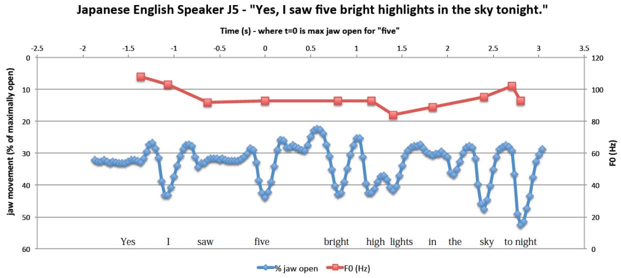

Figure 4: *Plot of F0 (in red) and jaw movement (in blue) for very low-level L2 speaker J5, sentence (1)*

For J5 (fig. 4), the relatively poor speaker of English, we see (1) neither a gradual declination of F0 (as seen for A1), nor a rhythmic alternation of F0 (as seen for J1), and (2) a pattern of jaw opening very different from that of A1 or J1. For one thing, the jaw seems to be open to almost the same degree for "I", "five", "bright", "high", "lights", "sky", and "night", with the largest jaw opening on the final word, "night." Moreover, we also see additional small jaw openings for the coda of "saw" and "five", and the word "the" has a large separate jaw opening.

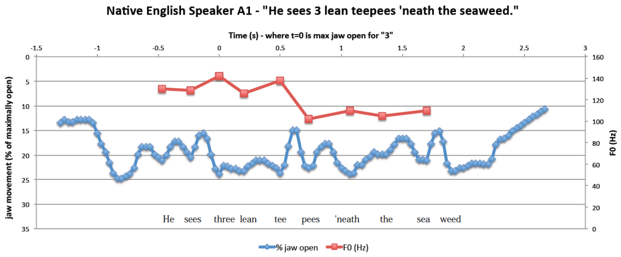

Figure 5: *Plot of F0 (in red) and jaw movement (in blue) for native English speaking subject A1, sentence (2)*

Turning now to figures 5, 6, and 7, "He sees three lean teepees 'neath the seaweed", we see patterns of jaw movement somewhat similar to the other utterance except that, of course, the amount of jaw opening is much less for /i/-vowels than for the /aJ/ vowels. For the A1 and J1 speakers, there is a similar pattern of jaw opening and F0, except that for J1, the jaw opening seems to be even more dramatic than for A1. Also, for J1, F0 is highest for the initial word “he”. This is also seen for J5, and is most likely a carry-over from Japanese prosody, where sentence initial syllables/words tends to have the highest F0. Also noteworthy is that for the poor speaker of English (J5), the largest jaw opening is on the non-content word “the”, whereas for the good speaker of Japanese and the A1 speaker, we see no such jaw opening.

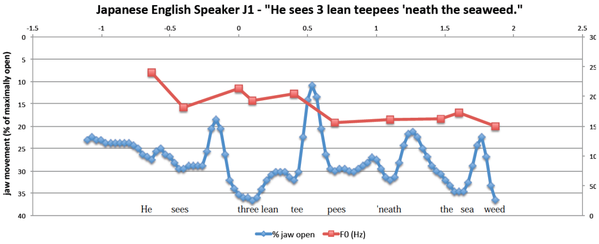

Figure 6: *Plot of F0 (in red) and jaw movement (in blue) for advanced L2 speaker J1, sentence (2)*

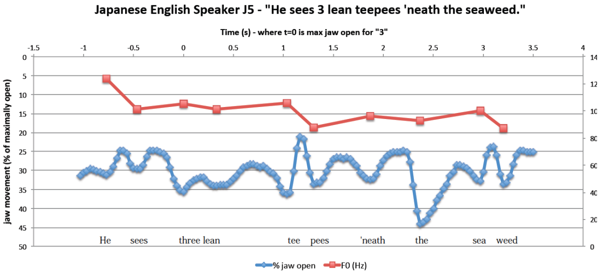

Figure 7: *Plot of F0 (in red) and jaw movement (in blue) for very low-level L2 speaker J5, sentence (2)*

4. Conclusions & future research

In summary, it would seem that there are patterns of F0 and jaw opening in American English that reflect an aspect of the rhythmical structure of English. Japanese speakers of English vary in their ability to re-produce these patterns. This study suggests that relatively good Japanese speakers of English do a better job of matching the F0 and jaw opening patterns of native speakers of English than do the relatively poor speakers of English. One interesting observation seems to be that the good L2 speakers of English maybe even “over articulate” the rhythmic patterns of English in terms of patterns of strong-weak F0 and jaw opening.

This study leads to many questions. Future work will examine the variability of jaw movement patterns for native speakers, in order to compare this with non-native speakers. We also plan to examine the stress/rhythm pattern in Japanese to see what Japanese listeners are paying attention to when they speak English. If we can make implications about what native rhythm should be, this has implications for teaching and correcting stress patterns in second language learning, and specifically for Japanese speakers of English. If it helps to produce the right stress pattern, then teaching jaw opening patterns as well as F0 patterns would be advisable.

Future research also will confirm the technique of video recording of jaw opening. Specifically, we want to examine similar sentences, but without labial consonants in order to make sure that labial consonants are not causing the skin to stretch over the jaw. Note that even though the skin may stretch over the mandible, the depth of the grooves that we see in the figures here would be unchanged, since those grooves correspond to vowel sounds for which the skin is stable. Also, we plan to do EMA experiments on the same subjects to confirm whether there is any difference in results due to different measuring techniques (i.e., video versus EMA).

5. Acknowledgements

Thanks to subjects at the University of Aizu and to student listeners at Capital University, Columbus, Ohio. Thanks also to Albert Rilliard, LIMSI, France for help creating the Runtime Revolution judgment test, and to Erin Brown, for help administering that test. This work was supported by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (C)#23520467 to the first author and (C)#22520412 to the second author.

6. References

[1] Kohler, K., “Whither speech rhythm research?” Phonetica, 66:5-14, 2009.

[2] Ladefoged, P., “A Course in Phonetics”, New York: Harcourt Brace Jovanavich, 1975.

[3] Ramus, F., Nespor, M. and Mehler, J., “Correlates of linguistic rhythm in the speech signal”, Cognition, 73(3):265-292, 1999.

[4] Low, E. L., Grabe, E. and Nolan, F., “Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English”, Language and Speech, 43(4):377-401, 2000.

[5] Arvaniti, A. and Ross, T., “Rhythm classes and speech perception”, in Proceedings of Speech Prosody (paper 887), 2010.

[6] Erickson, D., “An articulatory account of rhythm, prominence and phrasal organization”, in *Proceedings of Speech Prosody* (paper 2006), 2010.

[7] Erickson, D., Shibuya, S. and Suemitsu, A., “Rhythm and Emphasis in American English: Comparison of native and non-native speakers’ productions”, in Proceedings of International Seminar of Speech Production, pp.345-352, 2011.

[8] Horiguchi, N., “How L2 pronunciation learners interpret articulation instructions: An ultrasound study of the tongue”, MSc thesis, University of Aizu, 2012.