Effects of Mora-timing in English Rhythm Control by Japanese Learners

Shizuka Nakamura 1, Hiroaki Kato 2 and Yoshinori Sagisaka 1

1 Graduate School of Global Information and Telecommunication Studies, Waseda University
2 National Institute of Information and Communications Technology / Advanced Telecommunications Research Institute International

shizuka@akane.waseda.jp, kato.hiroaki@nict.go.jp, sagisaka@gti.waseda.ac.jp

Abstract

In this paper, we analyzed the durational differences between learners and native speakers in various speech units from the perspective of that the contrast between the stressed and the unstressed is one of the most important features to characterize stress-timing of English by comparison with mora-timing of Japanese. The results showed that the lengthening and shortening of learner speech were not enough to convey the difference between the stressed and the unstressed. Finally, it was confirmed that these durational differences strongly affected the subjective evaluation scores given by English language teachers.

Index Terms: rhythm control, mora-timing, stress-timing, syllable, Japanese, English

1. Introduction

We have studied the objective evaluation of English rhythm control by learners whose native language is Japanese (called “Japanese learners” or “learners” in this paper), and constructed an objective evaluation model which could simulate the process of evaluating rhythm control of learners by English language teachers (called “evaluators” in this paper) [1-3].

Durational information, which, among the acoustical features, can be calculated automatically in the relatively simple way, has been used for objective evaluation. Durational differences between native speakers and learners are correlated with subjective evaluation scores. Speech units, in which durational differences show strong correlations, have been selected from the phonetic viewpoint as parameters for the objective evaluation model.

Such speech units included the phoneme, the vowel (the stressed/ unstressed vowel, the vowel in the content/function word), the consonant (the voiced/unvoiced consonant, the consonant in the content/function word), the syllable (the strong/weak syllable, the closed/open syllable, the syllable in the content/function word), the word (the content/function word), the sentence, and the pause.

Figure 1 shows the process of the objective evaluation of English learners’ timing control by using our model. First, durational differences as an acoustical feature are calculated by using durational information (measured by labelling) on learner speech and native speech, which represents the subconscious evaluation standard of an evaluator, in the corresponding speech units. At this point, an evaluation measure and target need to be analyzed to precisely simulate the process of evaluating the rhythm control of learners by evaluators. Next, an objective evaluation model is constructed by using a multiple linear regression model. In order to judge the accuracy of a prediction, correlation coefficients between the objective evaluation scores predicted by our model and the seven-scale subjective evaluation scores (-3: Awful - +3: Excellent), as defined by evaluators, are used.

Speech samples were selected from the speech database of second language learners [4]. We used 480 samples uttered by a large number of Japanese learners (127 subjects), and these were compared to 215 samples uttered by English native speakers (21 subjects). Five text groups of four English sentences, 20 in total, were used to construct the speech materials. Each of the text groups consisted of four text lengths, namely, VS (Very Short), S (Short), L (Long), and VL (Very Long); these were edited by simply adding new

Figure 1. The procedure of the objective evaluation of English learners’ timing control by using a model that simulates the process of subjective evaluation by English language teachers.

Table 1. Sample texts of VS (Very Short), S (Short), L (Long) and VL (Very Long) varieties in text group A, and VL in group B, C and D. (: primary stress, , : secondary stress, : syllable boundary, /: phrase boundary)

<table>
<thead>
<tr>
<th>Group, Text length</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, VS (Very Short)</td>
<td>'Thank you very much.</td>
</tr>
<tr>
<td>A, S (Short)</td>
<td>'Thank you very much / for everything.</td>
</tr>
<tr>
<td>A, L (Long)</td>
<td>'Thank you very much / for everything / that you did for us.</td>
</tr>
<tr>
<td>A, VL (Very Long)</td>
<td>'Thank you very much / for everything / that you did for us.</td>
</tr>
<tr>
<td>B, VL</td>
<td>I'm a 'mused / by the man / and his ,very, fun,ny jokes.</td>
</tr>
<tr>
<td>C, VL</td>
<td>Why won't you 'wait / un,til 'Fri.day / when he's 'back ?</td>
</tr>
<tr>
<td>D, VL</td>
<td>I was ,ter.ri.ly an,'noyed / with the 'man / for ,beat.ing the 'dog.</td>
</tr>
</tbody>
</table>

the objective evaluation scores predicted by our model and the seven-scale subjective evaluation scores (-3: Awful - +3: Excellent), as defined by evaluators, are used.
phrases/clauses/words. Examples of the text are shown in Table 1.

In the process of these previous studies, we noticed that the ingrained mora-timing of Japanese learners might unfavourably affect the stress-timing of English speech. In this paper, we analyzed the durational differences between Japanese learners and native speakers in various speech units from the perspective of that the contrast between the stressed and the unstressed is one of the most important features to characterize stress-timing by comparison with mora-timing.

The purpose of this study is to analyze the properties in detail of the speech units used for the objective evaluation model from the phonetic viewpoint. An increase of correlation coefficients between durational differences and subjective evaluation scores attributed this study can also increase an accuracy of a prediction. Furthermore, the results of this study are expected to be useful for the study to specify the effects of mora-timing in the future.

2. Durational characteristics of learner speech

In this chapter, in order to analyze the actual conditions of learner speech, the durational characteristics in speech units—sentences, pauses and syllables—are examined by comparing them with the corresponding durations of native speech.

2.1. Sentence duration

First, sentence durations which consist of all speech units are analyzed. For English texts being used as evaluation material, only the VL texts of each text group in Table 1 were used, since it was statistically confirmed that a similar or higher accuracy of analysis could be obtained by this reduced set of material texts.

The results are shown in Figure 2. Although variations by text are observed, the average sentence duration of learners is about 1.5 times that of native speakers. In addition, the individual variety of learners extends to three times as wide as that of native speakers. In the following sections, in such a long sentence duration uttered by learners, the cause of different rhythm control by learners from that by native speakers is analyzed by dividing the sentence duration into the various speech units.

2.2. Pause duration

In this chapter, the sentence durations are separated into pause and non-pause durations in order to analyze the ratio of each to the overall sentence duration. While pauses are exceptionally inserted into native speech, the sum of the pause durations in the learner speech was found to increase in proportion with the increase in the text length.

On one hand, the percentage of pause durations in the VS (Very Short) texts is low; on the other hand, that in VL (Very Long) texts is around 40%. On the contrary, the percentage of non-pause durations is not variable unlike that in the case of the pause durations.

2.3. Syllable duration

The syllable duration of English native speakers, as required by the place and the manner of articulation of the preceding consonant, can be lengthened and shortened depending on factors such as whether the syllable is stressed or not, whether it is part of a content or function word, its position within a word or its context in a sentence. However, the innate mora-timing of Japanese learners causes them to have difficulty in lengthening or shortening the syllable durations.

First, the syllable durations of learners and that of native speakers are compared to analyze the influences on the durational difference. The syllable durations of learners in general are longer than those of native speakers.

Next, the phrase durations of all speakers are normalized by that of the typical native speaker to compare the ratios of the syllable duration to that of the phrase duration. An example of normalized syllable duration is shown in Figure 3—for the phrase “for everything” in a VL (Very Long) text in text group A, “Thank you very much for everything that you did for us.” The learner characteristics of lengthening the first syllable “for” can be observed more clearly after normalization in the areas encircled by dotted line.

This observation may be affected by the contrast between the stressed and unstressed syllables, and by the content and function words. Such learner characteristics of rhythm control, which differ from the native styles, are analyzed in the next chapter.

Figure 2: The comparison of the distribution of sentence durations by learners and native speakers on the box and whisker plot for four kinds of VL (Very Long) example texts.

Figure 3: Syllable durations are plotted alongside time for the phrase “for everything,” uttered by native speakers (top figure) and learners (bottom figure). The phrase durations are normalized by that of the typical native speaker (bold line).
3. The contrast between the stressed and the unstressed

Compared to the mora-timing of Japanese, one of the most important features to characterize stress-timing of English is the contrast between the stressed and the unstressed. In the following section, the characteristics of learner speech are analyzed by comparing them to the native style from the perspective of the contrast between the stressed and the unstressed.

3.1. Stressed/unstressed syllable

Considering the fact that a mora tends to be uttered for just as long as other morae in Japanese, it can be thought that shortened unstressed syllables help to make equal intervals between stresses in English. It is expected therefore that the difficulty in adjusting these durations can be shown in the stressed/unstressed syllable durations and the relationship.

The contrast between the average durations of the stressed and unstressed syllables is shown in Figure 4. The numbers of the stressed and unstressed syllables involved in the four VL texts are 19 and 34, respectively. The durations of the stressed syllables uttered by native speakers are twice as long as those of the unstressed ones, and they gather in the small solid ellipse drawn with the average as its center and an axis along the solid line; on the other hand, there is no significant tendency among those of learners, and they expand in the large dotted circle on the dotted line with the ratio of the stressed to unstressed syllable duration of about one and a half.

3.2. Strong/weak vowel

The durational differences between the stressed and the unstressed syllables in the previous section are affected, primarily, by not the consonants but the vowels, which lengthen and shorten more easily. In this section, the vowels in stressed and unstressed syllables, that is, the strong and weak vowels, are analyzed to study the results spelt out in the last section in greater detail.

The inter-speaker average sentence durations and the vowel durations are shown in Figure 5. The difference between the strong and the weak vowels uttered by native speakers is approximately 100 ms. Though some parts of the strong vowel by learners are incidentally close to those by native speakers since learner speech rate as a whole is slow, the difference is only about 50 ms.

However, the degree of the decrease of learner speech is smaller than that of native one. These results indicate that lengthening and shortening of learner speech are not enough to convey the difference between strong and weak vowels. Furthermore, the falling tendency of the curves indicates that the average vowel durations shorten as the text length increases.

3.3. Syllables in content/function words

These contrasts between the stressed and unstressed syllables and vowel durations are also affected by the differences between content and function words, since the vowel in content rather than in function words tends to be stressed. In this section, the vowel durations, which greatly influence syllable durations in content and function words, are analyzed. The number of vowels involved in the content and function words in the four VL (Very Long) texts are 29 and 27, respectively.

The ratios of the intra-speaker average vowel duration in content words to that in function words in native speech gathered around two, while those in learner speech expanded from two to one. In order to analyze how these ratios were consistent among the learners, both values were divided by the average vowel duration of the native speakers to normalize the differences across different texts.

As shown in Figure 6, the data on native speakers gather in a small solid ellipse with the axis along the function to content ratio of about one; in comparison, the data on learners expands over a large dotted ellipse with the axis along the ratio of about one and a half. The ratios deviate from the native data to almost two. These results show that the shortening of the vowel duration by learners are not enough to convey the contrast between content and function words, in addition, the deviation from the native data are not consistent among the learners.

---

**Figure 4:** Relationship between the intra-speaker average durations of stressed and unstressed syllables, uttered by native speakers (closed circle) and learners (open circle) in each of the four VL (Very Long) texts. + markes indicate the averages of native speakers (large) and learners (small). The lines (solid: native speakers, and dotted: learners) are drawn from the origin through their averages.

**Figure 5:** Relationship between the inter-speaker average sentence durations and the vowel durations. The lines show the approximation of the strong and weak values of learners (dotted line) and native speakers (solid lines) by using four kinds of text length (VS, S, L and VL) in four text groups (A, B, C and D).
Figure 6: Relationship between the intra-speaker average durations of vowels in content words and function words uttered by native speakers (closed circle) and learners (open circle). The values are normalized by dividing by their respective native average durations of each text. + markes and the lines are drawn in the same way in Figure 4.

4. Correlations between durational differences with subjective evaluation scores

In this chapter, we shall study the relationship between the unnatural speech durations of learners and their subjective evaluation by evaluators based on the correlation coefficients between durational differences and subjective evaluation scores.

The durational differences were the RMS scores calculated by using the corresponding typical native speech taken in each speech unit by texts. If the durational differences affect subjective evaluation, the former should have negative correlation with the subjective evaluation scores, whereby smaller scores would indicate a greater degree of unnaturalness.

Figure 7 shows the correlation coefficients of durational differences with subjective evaluation scores. As a result, it was confirmed that durational differences strongly affected subjective evaluation scores. Although variation by texts is observed, for the speech units referred to in the last chapter, the results show that it is the unstressed rather than the stressed syllable, the weak rather than strong vowel, and vowels in function rather than content words have the stronger correlation with the subjective evaluation scores.

5. Conclusions

In this paper, we analyzed durational differences between learners and native speakers in various speech units from the perspective of the contrast between stressed and unstressed syllables, devoting attention to the fact that stress-timing of English speech is unfavourably affected by the ingrained mora-timing of Japanese learners. These analyses revealed the durational characteristics of the learner speech and indicated that lengthening and shortening in learner speech was not enough to convey the difference between the stressed and the unstressed. Furthermore, it was confirmed that durational differences between learners and native speakers strongly affected the subjective evaluation scores.

Speech units, in which the calculated durational differences have strong correlations with subjective evaluation scores, are thought to be places where the characteristics of English rhythm control by Japanese learners can be observed with clarity. Since these should be effective in the evaluation of learner proficiency at rhythm control, the results of this study is expected to be taken in the objective evaluation model.

6. Acknowledgements

This study was supported in part by the Grant-in-Aid for Research fellow No. 21 · 06162 and for Scientific Research (B) No. 20300069, Japan Society for the Promotion of Science. We express our thanks to Prof. Minoru Tsuzaki (Kyoto City University of Arts) and Dr. Shigeki Matsuda (NICT) for their cooperation in this study, and Prof. Shizuo Hiki (Waseda University) for his advice.

7. References