Pan-Pacific Association of Applied Linguistics 14(1), 1-14

Analysis of Relationship between Duration Characteristics and Subjective Evaluation of English Speech by Japanese learners with regard to Contrast of the Stressed to the Unstressed

Shizuka Nakamura*
Waseda University / Japan Society for the Promotion of Science

Nakamura, S. (2010). Analysis of relationship between duration characteristics and subjective evaluation of English speech by Japanese learners with regard to contrast of the stressed to the unstressed. *Journal of Pan-Pacific Association of Applied Linguistics, 14*(1), 1-14.

To improve more effectively the learners' proficiency to control contrast of the stressed to the unstressed in English teaching, it is necessary to analyze how the acoustical characteristics of learners' speech are related to the perceptual evaluation by teachers. This paper analyzes A) learner characteristics of durations measured in speech units related to stress, which are stressed and unstressed syllables, and strong and weak vowels, and B) the relationship between the duration characteristics and subjective evaluation of English rhythm control proficiency. As a result of analyses, the following were revealed. Duration characteristics of learner speech were as follows: 1) Lengthened duration of speech units and inserted pauses make learners' sentence duration longer compared to that of native speakers. 2) Learner speech does not provide sufficient contrast between the stressed and the unstressed as native speech does, because it is difficult for learners to shorten the durations of unstressed syllables and weak vowels. The relationship between the duration characteristics and subjective evaluation were as follows: 1) Duration of the syllable, especially the unstressed syllable, and that of the vowel, especially the weak vowel, have a stronger correlation with subjective evaluation score. 2) Shorter duration of the weak vowel rather than that of the unstressed syllable tend to be evaluated as more native-like speech, that is, shortening the duration of the weak vowel strongly affects subjective evaluation.

Key Words: English speech by Japanese learners, rhythm control, duration characteristics, subjective evaluation, contrast of the stressed to the unstressed

* This study was supported in part by the Grant-in-Aid for Research Fellow No. 21・06162, Japan Society for the Promotion of Science. The research project "Objective evaluation of English speech uttered by learners based on mathematical model reflecting prosodic control and perceptual characteristics" including this work is supervised by Prof. Yoshinori Sagisaka (Waseda University). The author also wishes to thank Prof. Michiko Nakano (Waseda University) for her valuable advices on editing this article.

1 Introduction

As internationalization has advanced, the demand to acquire the ability to speak English has increased. When we consider ways of evaluating English learners' speech, it is desirable to develop the strategy of using subjective evaluations by English language teachers to a more precise and reliable stage. To this end, it is necessary to analyze multi-dimensionally the strategy of subjective evaluation by teachers and to find how they utilize the acoustical characteristics of learners' speech in their hearing capacity (Kondo et al., 2007). The analyzed strategy of subjective evaluation can be replaced by a more effective objective evaluation system by using a computer (Ito et al., 2006; Nakano et al., 2008; Yamashita et al., 2005). The present author has studied the analysis of the strategy of the evaluation (Nakamura et al., 2007; Nakamura et al., 2009). In this process, intrinsic and significant knowledge about the relationship between the acoustical features of learner speech and subjective evaluation are obtained. This paper reports these results.

Stress contributes greatly to the characterization of English rhythm control. The physical quantities of acoustical features that relate to stress are duration, fundamental frequency, and intensity (Campbel, 2000; Ladefoged, 1993; Lehiste, 1970). They correspond to the psychological quantities of phone length, pitch, and loudness, respectively. Among the acoustical features related to the subjective evaluation of rhythm control, durations in various speech units, which are the basis of the temporal structure, are treated in this paper for the following reasons: 1) Duration can be thought to include most of the information of fundamental frequency and intensity, and 2) The information of duration, which is based at the start and end points of each phone unit, can be measured with relatively high reliability.

The minimum speech unit that can be derived from measuring duration is the phone unit. Phone units are divided into vowel and consonant units. They are combined into syllable units, which are combined into word units, which, in turn, are combined into sentence units. The speech units related to stress, in these four levels (phone, syllable, word, and sentence) are the strong vowel, weak vowel, stressed syllable and unstressed syllable. This paper analyzes the following two aspects: 1) Characteristics of learner duration measured in these speech units 2) The relationship between the characteristics and a subjective evaluation. Obtained results, which can be useful in teaching learners to speak English, are reported in detail.

2 Analysis Data

In this chapter, speech data and subjective evaluation scores used for analyses are presented.

2.1 Speech data

Speech data used for analyses were selected from the "English Speech Database Read by Japanese Students (ERJ) (Minematsu, 2004)." This database is designed to meet the multiple demands related to creating a learning-support environment for non-native speech, such as the technical demand of speech information processing and the educational demand of learning to speak English. The English speech, which is spoken by learners of a wide range of English proficiency levels, was recorded in a standardized recording environment.

The ERJ database is consisted of several groups made up of sentences/words taking into consideration learning phonological/prosodic control; moreover, each group is broken into finer ones. This paper focuses on an evaluation of rhythm control in prosodic aspects. Therefore, the speech data used for analyses was selected from the group "sentences aiming at rhythm."

2.1.1 Text

In the following chapters, the effect of text length on an objective evaluation of rhythm control is analyzed. In the analyses, factors other than text length should be equalized as much as possible. The ERJ database includes various lengths of texts edited by simply adding new phrases/clauses/words. Texts were selected from them in this paper.

Furthermore, the texts were classified into four length levels, VS (Very Short), S (Short), L (Long), and VL (Very Long), on the basis of the number of syllables in a sentence, as shown in Table 1. Five text groups of four sentences showing different text lengths, 20 in total, were used to construct the speech data.

Table 1. Sample Texts of VS, S, L, and VL Varieties in Text Group A, and VL in Groups B, C, D, and E. (': Primary Stress, ,: Secondary Stress, .: Syllable Boundary, and /: Phrase Boundary)

Group, Text length	Text
A, VS (Very Short)	'Thank you.
A, S (Short)	,Thank you ,ver.y 'much.
A, L (Long)	,Thank you ,ver.y 'much / for 'eve.ry.thing.
A, VL (Very Long)	,Thank you ,ver.y 'much / for 'eve.ry.thing / that you 'did for us.
B, VL	I'm a.'mused / by the 'man / and his ,ver.y ,fun.ny 'jokes.
C, VL	I was ,ter.ri.bly an.'noyed / with the 'man / for ,beat.ing the 'dog.
D, VL	Why won't you 'wait / un.,til 'Fri.day / when he's 'back ?
E, VL	The ,boys have ,sold some of the 'flow.ers.

2.1.2 Speech data of learners

Four hundred and eighty samples were selected for the speech data. The number of samples was 120 in each of the datasets VS, S, L, and VL. Speakers were university students whose native language was Japanese (63 males and 64 females; 127 subjects in total). Each speaker uttered four or three sentences. (Four or three sentences were uttered by each speaker.)

Every sample was approved to be spoken with proper prosodic control by the speakers themselves. During practice and recordings, speech samples uttered by English native speakers were not presented as references. Additionally, learners were given prosodic symbols indicating phrase boundaries in the texts and were requested to practice speaking them prior to the recordings. Therefore, deviation of the phrasing among speech samples is reduced.

2.1.3 Speech data of native speakers

The ERJ database also includes speech data of English native speakers for the same set of sentences uttered by learners. The speech data of native speakers corresponding to those of learners are referred to as the target of comparison. There were 200 samples (10 samples for each sentence) uttered by 20 native speakers (8 males and 12 females) who speak General American.

2.2 Subjective evaluation scores

English language teachers were asked to give subjective evaluation scores to every selected speech sample spoken by learners. The evaluators were 5 English language teachers (2 males and 3 females) who had knowledge of English phonetics and careers in teaching English to Japanese learners. Evaluators did not include the native speaker who uttered the selected speech described in the previous section.

Subjective evaluation scores were made using a 7-point scale (-3: Awful - +3: Excellent) representing the level of proficiency in English rhythm control. Subjective evaluation scores were given to each sentence. Evaluators were allowed to listen to each speech sample of the sentence multiple times. Each speech sample was given one subjective evaluation score by every evaluator, with the result that every speech sample had 5 scores.

Based on these row subjective evaluation scores, an average subjective evaluation score was calculated for each speech sample. First, the average and the standard deviation of the scores of each evaluator were normalized by their average of the all five evaluators, in order to eliminate a bias caused by a difference of evaluators. Then, the average of the subjective evaluation scores for each speech sample was calculated.

In addition, some speech samples were excluded from analysis because a larger standard deviation of the average score of the five evaluators for each speech sample meant lower reliability. The criterion for exclusion was a standard deviation of over 1.5. The numbers of excluded speech samples were 6, 6, 6, and 9 in datasets VS, S, L and VL, respectively, 27 in total. The numbers of representative subjective evaluation scores after exclusion were 114, 114, 114 and 111, respectively, 453 in total.

The average subjective evaluation scores calculated in this way are simply called “subjective evaluation scores” in this paper hereafter.

3 Duration Characteristics of Learner Speech

In learner speech, durations of various speech units are lengthened, and pauses are inserted in some word boundaries as the learners try to utter unfamiliar English correctly. Consequently, the total sentence duration of learners tends to be longer than that of native speakers.

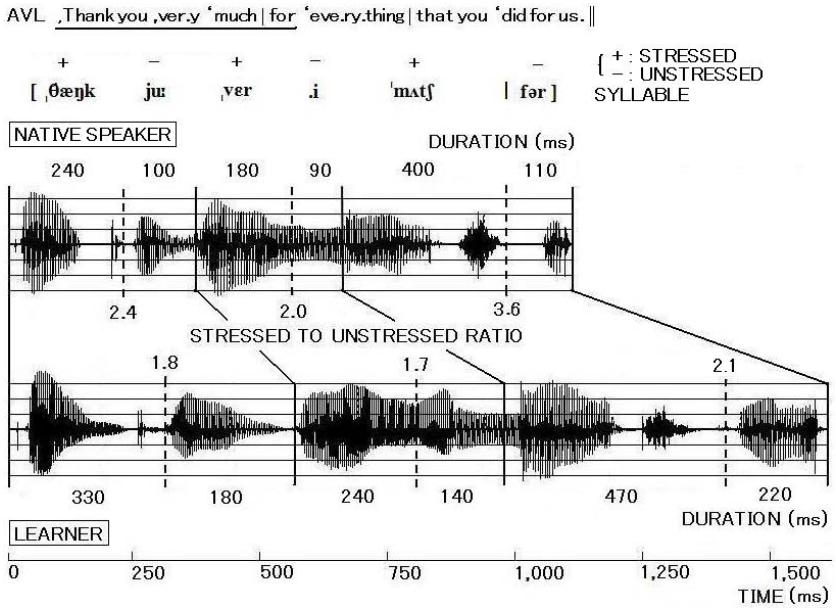
Analysis of the observed data revealed that learners’ intended rhythm control is not realized completely for these reasons, and this has the indirect effect of lowering subjective evaluation scores. However, an important role in English rhythm control is presenting the contrast of duration between stressed and unstressed speech units. This is considered to be directly connected to the subjective evaluation of learner speech.

The contrast of the stressed and the unstressed was measured by a ratio of the durations of the speech units concerned. Because this ratio is not affected by lengthening total sentence duration, it was reasonable to use raw duration without normalizing by sentence duration in the following analyses of the observed data.

3.1 Calculating contrast between stressed and unstressed syllables

An example of the speech waveform for a VL text of Group A ‘Thank you very much for everything that you did for us’ is shown in Figure 1. The top figure shows an utterance by a native speaker and the bottom by a learner. The duration of the phonemic segments derived by the computer program are plotted along the time axis of the speech waveform. In order to measure the durations of the speech segments, the recorded speech samples of both the native speakers and learners were stored digitally on a computer with 16 kHz and ±16 bits precision. By automatically analyzing their various acoustical characteristics, phoneme boundaries were derived within an accuracy of 10 ms. In addition, ratios of stressed to unstressed syllable duration are plotted by each dotted line of a boundary between stressed and unstressed syllables.

Figure 1. Comparison of syllable durations of a native speaker (top figure) and a learner (bottom figure). The waveform is drawn alongside time for "thank you very much for" which is the beginning portion of the Text AVL. Integers show syllable durations, and decimal numbers show the ratio of the stressed syllable duration to the following unstressed one.



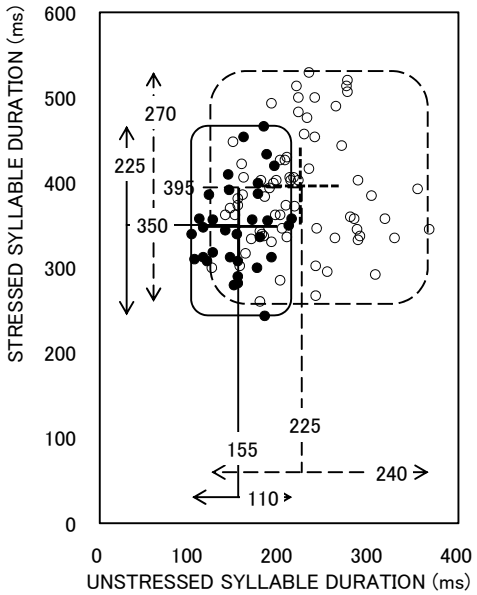
In the speech of a native speaker, the duration of stressed and unstressed syllables shows their contrastive difference. In the speech of a learner, compared to that of a native speaker, the difference is smaller and the contrast is unclear. These characteristics are shown in the result that the ratios of the duration of stressed to unstressed syllables in learner speech are lower than those in native speech.

Learner characteristics of syllable duration control as expressed by a distinction of stressed and unstressed syllables are different from native ones in this way. In Section 3.2, the duration of stressed and unstressed syllables is analyzed.

3.2 Duration of stressed and unstressed syllables

Durations of stressed syllables are mostly longer than those of unstressed syllable, as some examples were seen in Figure 1. The relationship between the durations of stressed and unstressed syllables averaged for each of the native speakers and learners is shown in Figure 2. Stressed syllable duration

Figure 2. Relationship between the intra-speaker average durations of stressed (vertical axis) and unstressed (horizontal axis) syllables uttered by native speakers (closed circle) and learners (open circle) in the three VL (Very Long) texts. + indicates the average of native speakers and learners. Native speakers are shown with solid lines and learners with solid lines.



is shown on the vertical axis, and unstressed syllable duration is shown on the horizontal axis. Three among the five texts were selected for this analysis in order to keep their text length close each other.

The analysis focuses on the distribution of syllable durations, shown with two-headed arrows. Stressed syllable durations in native speech range over 225 ms around their average of 350 ms and those in learner speech range over 270 ms (average 395 ms). Stressed syllable durations in learner speech distribute about 1.2 times as widely as those of native speech. These differences between native and learner speech is more evident in the unstressed syllable duration.

Unstressed syllable durations in native speech concentrate within 110 ms (average 155 ms) and those in learner speech spread over 240 ms (average 225 ms). Unstressed syllable durations in learner speech distribute about 2.2 times as widely as those of native speech. These results indicate it is difficult for learners to shorten syllable durations, especially unstressed syllable durations, as native speakers do.

3.3 Duration of strong and weak vowels

The duration differences between the stressed and the unstressed syllables are primarily affected by not the consonant portion in the syllable, but the vowel portion, which lengthen and shorten more easily. In this section, vowels in stressed and unstressed syllables, that is, strong and weak vowels, are analyzed to study the results described in Section 3.2 in greater detail.

An example of the speech waveform for a VL text of Group E 'The boys have sold some of the flowers' is shown in Figure 3. The top figure shows an utterance by a native speaker and the bottom by a learner. The duration of the syllables and vowels in the syllables derived by the computer program are plotted along the time axis of the speech waveform.

Durations of the strong vowels are mostly longer than those of weak vowels, as seen in the examples of Figure 3. Figure 4 shows the relationship between the durations of strong and weak vowels averaged for each of the native speakers and learners. Strong vowel duration is shown on the vertical axis, and weak vowel duration is shown on the horizontal axis.

Figure 3. Speech waveform. Comparison of durations of strong and weak vowels in stressed and unstressed syllables of a native speaker (top figure) and a learner (lower figure). The waveform is drawn alongside time for "the boys have sold" which is the beginning portion of the Text EVL. Integers show measured syllable and vowel durations.

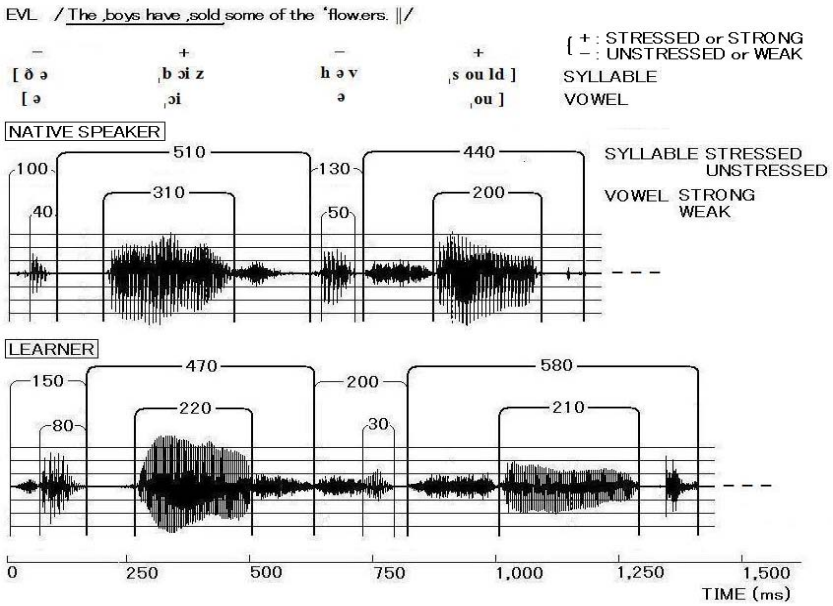
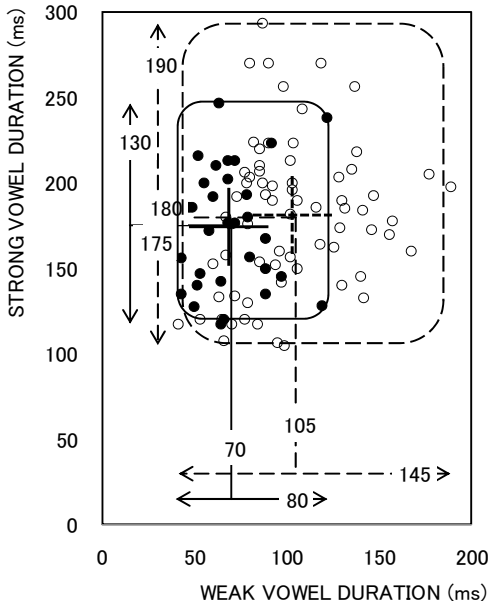


Figure 4. Relationship between the intra-speaker average durations of strong (vertical axis) and weak (horizontal axis) vowels, uttered by native speakers (closed circle) and learners (open circle) in the three VL (Very Long) texts. + indicates the averages of native speakers and learners. Native speakers are shown with solid lines and learners with dotted lines.



The tendency found in the syllable duration is also noticed in the vowel duration. The distribution of vowel duration shown with two-headed arrows is discussed. Strong durations in native speech range over 130 ms (average 175 ms), and those in learner speech range over 190 ms (average 180 ms). Strong vowel durations in learner speech distribute about 1.5 times as wide as those of native speech.

Weak vowel durations in native speech range over 80 ms (average 70 ms), and those in learner speech range over 145 ms (average 105 ms). Weak vowel durations in learner speech distribute about 1.8 times as widely as those in native speech. These results indicate it is difficult for learners to shorten vowel durations, especially weak vowel durations, as native speakers do.

3.4 Comparing contrast between the stressed and the unstressed

Figure 5 shows the ratios of average durations of stressed to unstressed syllable and strong to weak vowel. The ratio of average duration of stressed to unstressed syllable in native speech is 2.2, as indicated by the solid line

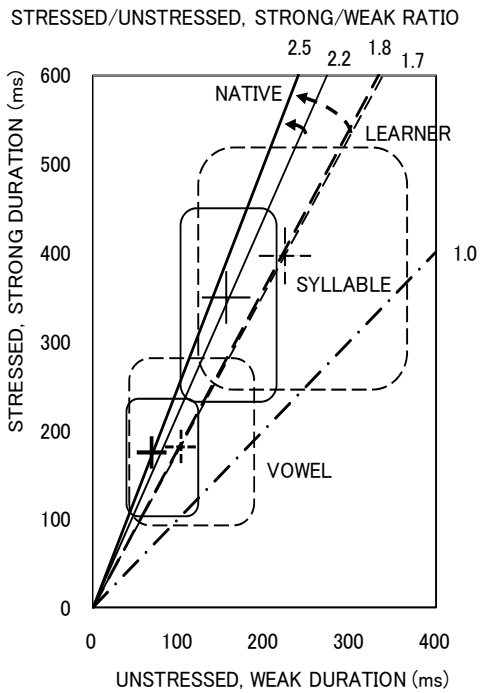
drawn from the origin through its average point in the distribution area, and that in learner speech is 1.7. The ratio in syllable duration of native speech is greater than that in learner speech (i-a).

The ratio of average duration of strong to weak vowel in native speech is 2.5, and that in learner speech is 1.8. The ratio in vowel duration of native speech is also greater than that in learner speech as indicated by dotted arrowed arc in Figure 5 (i-b).

These ratios imply that the lengthening and shortening of duration for conveying the stressed and the unstressed is not large enough in learner speech. The ratio of the stressed to unstressed in vowel duration is greater than that in syllable duration. This tendency is clearly observed especially in native speech as indicated in solid arrowed arc in Figure 5 (ii).

As stated above (i-a, i-b and ii), the characteristics of the contrast between the stressed and the unstressed is observed.

Figure 5. Ratios of stressed/unstressed syllable and strong/weak vowel durations, uttered by native speakers and learners in the three VL (Very Long) texts. + indicates the averages of native speakers and learners. The lines (solid: native speakers; dotted: learners) are drawn from the origin through their averages.



4 Relationship between Duration Characteristics of Learner Speech and Subjective Evaluation

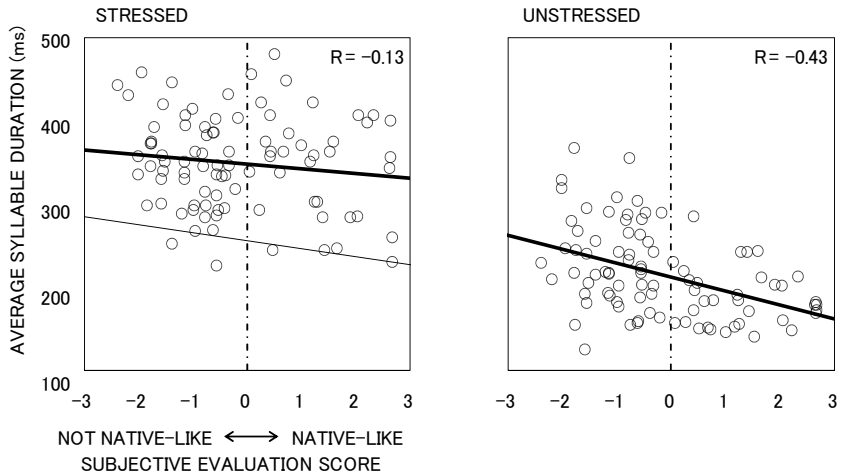
The duration characteristics of learner speech were revealed in the last chapter. In this chapter, correlations between durations calculated in speech units treated in Chapter 3 and subjective evaluation scores are studied to investigate how the duration characteristics affect subjective evaluation.

4.1 Duration of stressed and unstressed syllables

The relationship between the duration of stressed/unstressed syllables and subjective evaluation is analyzed. First, Figure 6 (left part) shows the result of stressed syllable duration and subjective evaluation. Stressed syllable durations normalized by calculating intra-speaker averages are shown on the vertical axis, and subjective evaluation scores are shown on the horizontal axis. A very weak correlation coefficient showing -0.13 of stressed syllable durations with subjective evaluation scores is obtained.
Next, Figure 6 (right part) shows the result of unstressed syllable duration and subjective evaluation. Unstressed syllable durations normalized by calculating intra-speaker averages are shown on the vertical axis, and subjective evaluation scores are shown on the horizontal axis. A weak correlation coefficient showing -0.43 of unstressed syllable durations with subjective evaluation scores is obtained.

These results indicate that shorter syllable duration and especially unstressed syllable duration tend to be evaluated as more native-like speech.

Figure 6. Relationship between subjective evaluation scores and durations of stressed (left part) / unstressed (right part) syllables



4.2 Duration of strong and weak vowels

The relationship between the duration of strong/weak vowels and subjective evaluation is analyzed. First, Figure 7 (left part) shows the result of strong vowel duration and subjective evaluation. Strong vowel durations normalized by calculating intra-speaker averages are shown on the vertical axis, and subjective evaluation scores are shown on the horizontal axis. A very weak correlation coefficient showing -0.15 of stressed syllable durations with subjective evaluation scores is obtained.

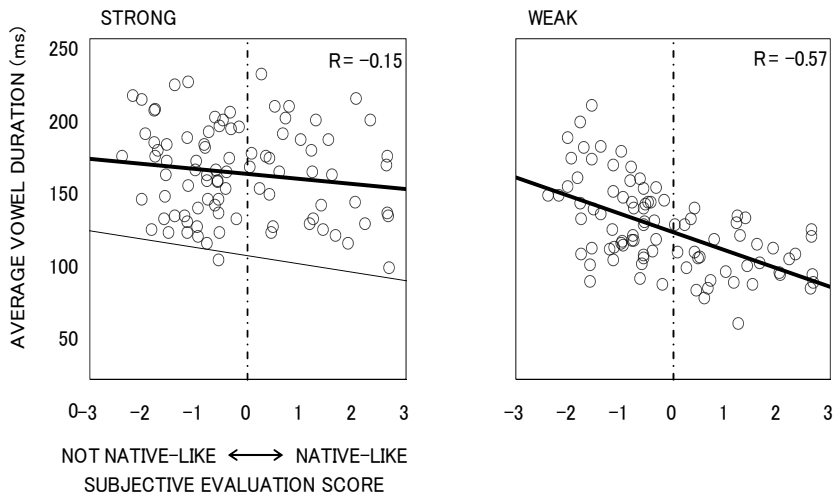
Next, Figure 7 (right part) shows the result of weak vowel duration and subjective evaluation. Weak vowel durations normalized by calculating intra-speaker averages are shown on the vertical axis, and subjective evaluation scores are shown on the horizontal axis. A slightly strong correlation coefficient showing -0.57 of stressed syllable durations with subjective evaluation scores is obtained.

These results indicate that shorter vowel duration and especially weak vowel duration tend to be evaluated as more native-like speech.

5 Conclusions

Learner characteristics of durations measured in speech units related to stress, and their relationship to subjective evaluation of English rhythm control proficiency were analyzed in this paper.

Figure 7. Relationship between subjective evaluation scores and durations of strong (left part) / weak (right part) vowels



Obtained results on duration characteristics of learner speech were as follows:

1) Lengthened duration of speech units and inserted pauses make learners' sentence duration longer compared to that of native speakers.
2) Learner speech does not provide sufficient contrast between the stressed and the unstressed as native speech does, because it is difficult for learners to shorten the durations of unstressed syllables and weak vowels.

Obtained results on the relationship between the duration characteristics and subjective evaluation were as follows:

1) Duration of the syllable, especially the unstressed syllable, and that of the vowel, especially the weak vowel, has a stronger correlation with subjective evaluation score.
2) Shorter duration of the weak vowel rather than that of the unstressed syllable tend to be evaluated as more native-like speech, that is, shortening the duration of the weak vowel strongly affects subjective evaluation.

These findings on duration characteristics of learner speech and their relation to subjective evaluation, which were obtained through detailed acoustical and statistical analyses, will serve as relevant knowledge toward teaching to control native-like English rhythm.

References

Campbell, N. (2000). Timing in Speech: a Multi-level Process. In M. Horne (Ed.), *Prosody, theory and experiment: Studies presented to Gosta Bruce* (pp. 281-334). Boston: Kluwer Academic Publishers.

Ito, A., Nakagawa, T., Ogasawara, H., Suzuki, M., & Makino, S. (2006). Automatic detection of English mispronunciation using speaker adaptation and automatic assessment of English intonation and rhythm. *Educational Technology Research*, *29*, 13-23.

Kondo, Y., Tsutsui, E., Tsubaki, H., Nakamura, S., Sagisaka, Y., & Nakano, M. (2007). Examining predictors of second language speech evaluation. *Proceedings of the PAAL*. 176-179.

Ladefoged, P. (1993). *A Course in Phonetics*. Fort Worth, Texas: Harcourt Brace Jovanovich.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge: MIT Press.

Minematsu, N., Tominaga, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S.,

Dantsuji, M., & Makino, S. (2004). Development of English speech database read by Japanese to support CALL research. *Proceedings of the International Congress on Acoustics*. 557-560.

Nakamura, S., Matsuda, S., Kato, H., Tsuzaki, M., & Sagisaka, Y. (2009). Objective Evaluation of English Learners' Timing Control Based on a Measure Reflecting Perceptual Characteristics. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 4834-4840.

Nakamura, S., Tsubaki, H., Kondo, Y., Nakano, M., & Sagisaka, Y. (2007). Tempo-normalized measurement and test set dependency in objective evaluation of English learners' timing characteristics. *Proceedings of the 16th International Congress of Phonetic Sciences*. 1733-1736.

Nakano, M., Kondo, Y., & Tsutsui, E. (2008). Fundamental Research on Automatic Speech Evaluation. *Proceedings of the 9th APRU Distance Learning and the Internet Conference—New Directions for Inter-institutional Collaboration: Assessment & Evaluation in Cyber Learning*. 207-212.

Yamashita, Y., Kato, K., & Nozawa, K. (2005). Automatic scoring for prosodic proficiency of English sentences spoken by Japanese based on utterance comparison. *IEICE Transactions on Information and Systems, E88-D*, 496-501.

Shizuka Nakamura
Research Fellow,
Japan Society for the Promotion of Science
Ph.D. Student,
Graduate School of Global Information and Telecommunication Studies,
Waseda University
1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, Japan
Phone: +81 3 5286 3841
E-mail: shizuka@akane.waseda.jp

Received: January 22, 2010
Revised: June 01, 2010
Accepted: June 15, 2010