A REQUIREMENT OF TEXTS FOR EVALUATION OF RHYTHM IN ENGLISH SPEECH BY LEARNERS

Shizuka Nakamura & Yoshinori Sagisaka

Graduate School of Global Information and Telecommunication Studies, Waseda University, Japan
shizuka@akane.waseda.jp; sagisaka@giti.waseda.ac.jp

ABSTRACT

In objective evaluation of rhythm in English speech uttered by learners, the prediction accuracy is greatly affected by the properties of texts given to the subject to speak. In order to construct an objective evaluation model for more efficient prediction of the subjective evaluation scores given by evaluators, we examined a requirement of texts for evaluation of rhythm in English speech. A model was constructed based on acoustic differences in speech between learners and native speakers. Repetition of stressed syllables alternating with unstressed syllables helps listeners perceive rhythm in English. Focusing on the property of salient stressed syllables in the repetition, the number of stressed syllables included in a text was treated as an element of a requirement. As a result of prediction experiments, it became clear that a text including three stressed syllables is an effective requirement of texts for evaluation of rhythm in English.

Keywords: evaluation of learners' speech, rhythm in English, stressed and unstressed syllables, text

1. INTRODUCTION

In previous studies aimed at objective evaluation of the prosodic aspects of the English speech of learners, the evaluation method was analyzed by comparing the speech of learners and native speakers in terms of acoustical features reflecting prosodic aspects such as intensity, fundamental frequency, and duration e. g. [1, 10]. In this study, we focus on the rhythm as an important factor for Japanese learners of English speech.

Among acoustical features concerning rhythm in English speech, we concentrate on duration for the following reasons: duration greatly contributes to stress, which builds the structure of rhythm in English speech [4], and duration also plays a role in the temporal framework of a change of the other acoustical features.

A subjective evaluation measure of a 7-point scale is used to evaluate the proficiency level of each learner's rhythm in English. An objective evaluation model to predict the subjective evaluation scores is constructed with a linear regression model. Any difference in the syllable duration of a learner compared to a native speaker is treated as an independent variable, since syllable duration greatly concerns rhythm in English.

In this type of objective evaluation of rhythm in English speech, the prediction accuracy is greatly affected by the properties of the texts given to the subject to speak. The reason is that a possibility of demonstrating proficiency levels by learners and an accuracy of judgments by human evaluators depends on the properties of the texts used in it. In previous studies, different prediction accuracies were obtained by using different lengths of texts [7, 8]. Therefore, in order to construct an objective evaluation model for more efficient prediction of the subjective evaluation scores, we examine a requirement of texts for evaluation of rhythm in English speech in this study.

In the next chapter, a requirement of texts for evaluation of rhythm in English is set. In Chapter 3, an objective evaluation model is constructed by using speech samples satisfying all requirements. In Chapter 4, appropriateness of a requirement is investigated by prediction experiments.

2. SETTING A REQUIRMENT OF TEXTS FOR EVALUATION OF RHYTHM

2.1. The number of stressed syllables included in a text

Stress characterizes rhythm in English speech. The reason that a stressed syllable is recognized as a syllable with a stress is the following: it is heard to stand out more prominently than nearby unstressed syllables by longer duration, greater intensity, and higher pitch [9]. Repetition of these stressed syllables alternating with unstressed syllables helps listeners perceive rhythm in English [4]. In this study, focusing on the property of salient stressed syllables in the repetition, the number of stressed syllables included in a text is treated as an element of a requirement of texts.

More than one stressed syllable is included in a text of an English sentence. In the case of a text including one stressed syllable, no interval between stressed syllables is formed. In the case of a text including two stressed syllables, one interval between these stressed syllables is formed, and its absolute duration is perceived. In the case of a text including at least three stressed syllables, at least two intervals between adjacent stressed syllables are formed, and the rhythm by its periodic repetition is also perceived. For these reasons, a text including at least three stressed syllables is set as a requirement of texts for evaluation of rhythm in English.

The location and degree of stress can be changed in some cases according to a general English rule to avoid having stresses too close, but to maintain regular intervals [3]. This habit is not necessarily done the same by all native speakers [9]. Furthermore, the degree of stress is not paid attention to even by native listeners in an ordinary situation [2].

Considering these facts, just a syllable with a primary stress, which stands out prominently in contrast to an unstressed syllable, is treated as a stressed syllable in this study. This can be useful for weakening the effect of a difference in the way evaluators recognize stress and having a clearer result. Hereafter, a syllable with a primary stress is called a stressed syllable, and the other syllable is called an unstressed syllable.

2.2. The difficulty of words

The target of evaluation in this study is a proficiency level of not a phonological aspect of a specific word, but a prosodic one in a whole sentence, especially rhythm in English. For this reason, it is desirable that there is no difference between learners in their knowledge of the words included in texts. Therefore, texts mainly consisting of the simple words required in English classes in Japanese junior high schools and containing no proper nouns are selected.

2.3. The length of texts

Considering the limitation of accurate evaluation by human evaluators, simple sentences or complex ones that consists of two pairs of a subject and a predicate are selected as texts. Generally, the number of words is about 7 in a simple sentence and 14 in a complex one, in view of the structure of a normal sentence.

The number of words in each text used in this study is 8-11. This is within the limit mentioned above. The number of syllables is 9-15 in these texts.

3. CONSTRUCTING AN OBJECTIVE EVALUATION MODEL TO PREDICT SUBJECTIVE EVALUATION SCORES

3.1. Materials used in experiments

3.1.1. Speech samples of learners and native speakers

Speech samples used in experiments were selected from the “English speech database read by Japanese students (hereafter ERJ database) [5],” which includes texts satisfying all requirements for evaluation of rhythm in English mentioned in the last chapter. This database consists of English speech uttered by learners of a wide range of English proficiency levels and recorded in a standardized recording environment.

Five texts satisfying the requirements were selected for experiments. As shown in the second line of each text in Table 1, every text includes three stressed syllables indicated by the symbol “@.”

Table 1: Texts including three stressed syllables. The symbol “.” indicates a syllable boundary. Stressed (@) and unstressed (-) syllables are based on the definitions described in 2.1.

Text
Why won’t you wait un.til Fri.day when he’s back? - - - @ - - @ - - - @
I’m a.mused by the man and his ver.y fun.ny jokes. - - @ - - @ - - - - - - @
Thank you ver.y much for eve.ry.thing that you did for us. - - - - @ - @ - - - - @ - -
The boys have sold some of the flow.ers. - @ - @ - - - @ -
I was ter.ri.bly an.noyed with the man for beat.ing the dog. - - - - - - @ - - @ - - - - @

One hundred and six samples were selected for speech samples of learners. Speakers were 106 university students whose native language was Japanese. The number of samples per text was approximately 21. In the process of constructing the ERJ database, speech samples uttered by native speakers were not presented as references during practices and recordings. Additionally, learners were given prosodic symbols indicating location and degree of stress in the texts and required to practice speaking them prior to the recordings.

The ERJ database also includes speech samples of native speakers for the same texts as those of learners. Fifty-eight samples uttered by 20 native speakers, corresponding to those by the learners mentioned above, were selected. The number of samples per text was approximately 11.

3.1.2. Subjective evaluation scores

Five selected English language teachers were asked to give subjective evaluation scores to selected speech samples of learners. The evaluators had knowledge of English phonetics and careers in teaching English to Japanese learners. The evaluators were different people from the native speakers who used previously to utter the selected speech described in 3.1.1.

An evaluation measure of a 7-point scale (-3: “Very low” to +3: “Very high”) representing the proficiency level of the rhythm in the English speech was used in subjective evaluation. Evaluators were allowed to listen to each speech sample multiple times.

3.2. An objective evaluation model using a linear regression model

Learners aim to control rhythm in English as native speakers do. When comparing the speech of learners to native speakers for the same texts, the closer the duration was to the native speakers, the higher the proficiency level of rhythm in English.

Duration differences were calculated for each syllable to measure the rhythm in English, since syllable duration greatly concerns rhythm in English as mentioned in 2.1. Syllable durations of learners and native speakers were calculated using the results of a forced phoneme alignment using a speech recognizer utilized HTK [11] in 10ms.

An objective evaluation model predicting subjective evaluation scores was constructed by using a linear regression model. A subjective evaluation score was treated as a dependent variable. A square error of the corresponding syllable duration of a learner from a native speaker was treated as an independent variable. A factor of an independent variable was decided to minimize the error of a subjective evaluation score with a predicted one. Prediction accuracy of an objective evaluation model was judged from a correlation coefficient between subjective evaluation scores and predicted ones. Arrangements of these data are explained in the following sections.

3.2.1. A subjective evaluation score as a dependent variable

One subjective evaluation score was given to each speech sample of a whole sentence by one evaluator in subjective evaluation mentioned in 3.1.2. As a result, five scores in total were given to each sample. It was desirable that one subjective evaluation score corresponded to one speech sample in the case of a linear model used in the objective evaluation model in this study. For this reason, a representative subjective evaluation score was calculated for each speech sample. A representative score was based on the average value of five scores [6]. A representative subjective evaluation score calculated in this way is called just a subjective evaluation score hereafter.

3.2.2. A syllable duration difference as an independent variable

Since learners speak more slowly than native speakers, learners take a longer to speak a sentence than native speakers do [6, 7]. To calculate duration differences between learners and native speakers appropriately, sentence durations of learners were normalized by the representative sentence duration of native speakers, based on the method of previous study [7].

4. INVESTIGATING THE APPROPRIATENESS OF A REQUIREMENT BY PREDICTION EXPERIMENTS

To carry out prediction experiments, we constructed an objective evaluation model after the data was arranged as previously mentioned. The data set consisted of 106 samples. Trainings and tests of the model were performed by three-fold cross-validation. In the case of a text including three stressed syllables, a correlation coefficient of 0.68 was obtained for the test set, as shown in Table 2.

Table 2: Comparison of correlation coefficients of subjective and objective evaluation scores depending on the number of stressed syllables included in a text.

The number of stressed syllables	Training	Test
3	0.70	0.68
2	0.52	0.49
1	0.38	0.34

To investigate the appropriateness of this requirement, prediction experiments were carried out in the same way after the number of stressed syllables included in a text was decreased. Since the limitation of accurate evaluation by human evaluators was taken into account as mentioned in 2.3, the number of stressed syllables was not increased. An example of a text including two or one stressed syllable(s) is shown in Table 3. These texts were founded on the text including three stressed syllables such as a text shown in the top line of Table 3. They were formed by eliminating a few phrases from the top text.

Table 3: An example of texts including three, two, and one stressed syllable(s) included in a text. The symbol "." indicates a syllable boundary. Stressed (@) and unstressed (-) syllables are based on the definitions described in 2.1.

The number of stressed syllables	Text
3	Why won't you wait un.til Fri.day when he's back? - - - @ - - @ - - - @
2	Why won't you wait un.til Fri.day? - - - @ - - @ -
1	Why won't you wait? - - - @

Each data set, which was also selected form the ERJ database, consisted of the same number of speakers and samples in the last experiment. As shown in Table 2, correlation coefficients of 0.49 and 0.34 were obtained for the test set in the case of two and one stressed syllable(s), respectively. There was a significant difference of the results between in the case of three and one, and also three and two stressed syllable(s), respectively, at the 0.01 level of significance. It became clear that a text including three stressed syllables is an effective requirement of texts for evaluation of rhythm in English compared to a text including two or one stressed syllable(s).

5. CONCLUSIONS

In objective evaluation of rhythm in English speech spoken by learners, the prediction accuracy is greatly affected by the properties of texts given to the subject to speak. In order to construct an objective evaluation model for more efficient prediction of the subjective evaluation scores given by evaluators, we examined the requirement of texts for evaluation of rhythm in English speech. A model was constructed based on acoustic differences in speech between learners and native speakers.

Repetition of stressed syllables alternating with unstressed syllables helps listeners perceive rhythm in English. Focusing on the property of salient stressed syllables in the repetition, the number of stressed syllables included in a text was treated as an element of a requirement.

In the case of a text including one stressed syllable, no interval between stressed syllables is formed. In the case of a text including two stressed syllables, one interval between these stressed syllables is formed, and its absolute duration is perceived. In the case of a text including at least three stressed syllables, at least two intervals between adjacent stressed syllables are formed, and the rhythm by its periodic repetition is also perceived. For these reasons, a text including at least three stressed syllables was set as a requirement of texts for evaluation of rhythm in English.

Our experiments confirmed that prediction accuracy in the case of a text including three stressed syllables was significantly higher than that in the case of two or one stressed syllable(s). Consequently, it became clear that a text including three stressed syllables is an effective requirement of texts for the evaluation of rhythm in English. This requirement can be widely applied to select texts to use in subjective and objective evaluation of rhythm in English.

6. ACKNOWLEDGEMENTS

This study was supported in part by Grant-in-Aid for Research Fellow No. 21・06162 of JSPS and Waseda University Grant for Special Research Project 2010B-367.

7. REFERENCES

[1] Ito, A., et. al. 2008. Improvement of automatic English prosody evaluation based on word clustering using a decision tree. *IEICE Trans. Inf. Syst.* J91-D, 358-366.

[2] Jones, D. 1960. *An Outline of English Phonetics.* Tokyo: W. Heffer & Sons LTD.: Cambridge and Maruzen Company LTD., 245-247.

[3] Ladefoged, P. 1975. *A Course in Phonetics.* Texas: Hartcourt Brace Jovanovich, 102-103.

[4] Lehiste, I. 1970. *Suprasegmentals.* Cambridge: MIT Press, 106-153.

[5] Minematsu, N., et. al. 2003. Read speech database for foreign language learning. *J. Acoust. Soc. Jpn.* 59, 345-350.

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

[7] Nakamura, S., et. al. 2007. Tempo-normalized measurement and test set dependency in objective evaluation of English learners' timing characteristics. *Proc. ICPhS* Saarbrucken, 1733-1736.

[8] Nakamura, S., et. al. 2009. Objective evaluation of English learners' timing control based on a measure reflecting perceptual characteristics. *Proc. IEEE ICASSP* Taipei, 4834-4840.

[9] Roach, P. 2009. *English Phonetics and Phonology.* Cambridge: Cambridge University Press.

[10] Yamashita, Y., et. al. 2005. Automatic scoring for prosodic proficiency of English sentences spoken by Japanese based on utterance comparison. *IEICE Trans. Inf. Syst.* E88-D, 496-501.

[11] Young, S., et. al. 2006. *The HTK Book (for HTK Version 3.4).* Cambridge: Cambridge Uni. Engineering Dept.