Automatic Evaluation System of English Prosody Based on Word Importance Factor

Motoyuki Suzuki†, Tatsuki Konno‡, Akinori Ito‡ and Shozo Makino‡.

†Institute of Technology and Science, The University of Tokushima.
2–1, Minamijosanjima-cho, Tokushima, 770-0815, Japan.

‡Graduate School of Engineering, Tohoku University.
6–6–05, Aramaki-Aza-Aoba, Aoba-ku, Sendai, 980-8579, Japan.

ABSTRACT

Prosody plays an important role in speech communication between humans. Although several computer-assisted language learning (CALL) systems with utterance evaluation function have been developed, the accuracy of their prosody evaluation is still poor.

In the present paper, we develop new methods by which to evaluate the rhythm and intonation of English sentences uttered by Japanese learners. The novel features of our study are as follows: (1) new prosodic features are added to traditional features, and (2) word importance factors are introduced in the calculation of intonation score. The word importance factor is automatically estimated using the ordinary least squares method and is optimized based on word clusters generated by a decision tree.

Experiments conducted herein reveal the correlation coefficient (± 1.0 denotes the best correlation) between the rhythm score given by native speakers and the system was -0.55. In contrast, a conventional feature (pause insertion error rate) gave a correlation coefficient of only -0.11. The correlation coefficient between the intonation scores given by native speakers and the system was only -0.29. However, the word importance factor with decision tree clustering improved the correlation coefficient to 0.45.

In addition, we propose a method of integrating the rhythm score with the intonation score, which improved the correlation coefficient from 0.45 to 0.48 for evaluating intonation.

Keywords: computer-assisted language learning system, prosody evaluation, rhythm, intonation, decision tree

1. INTRODUCTION

It is important for non-native English speakers to be able to communicate in English. Communication skills can be improved by individual study through educational television and radio programs, textbooks, and educational materials such as CDs and DVDs. However, it is very difficult to study speaking skills such as pronunciation and prosody because the learner cannot evaluate his/her own speech.

In order to solve this problem, several Computer-Assisted Language Learning (CALL) systems with utterance evaluation function have been developed [1–3]. In these systems, acoustical features are extracted from a learner's speech and are compared with those of native speakers. Many of these systems can evaluate the pronunciation of the learner's speech. For instance, Kawai's system [3] can detect typical pronunciation errors of English made by Japanese learners. Many Japanese learners make insertion errors of vowels and modify various English phonemes into Japanese phonemes. Kawai's system is based on speech recognition technology and detects such pronunciation errors using both English and Japanese phoneme models.

On the other hand, prosody plays an important role in English communication between humans [4]. In other words, the CALL system should evaluate the correctness of prosody of the learner's speech in addition to the evaluation of pronunciation. Several CALL systems can evaluate the prosody of a learner's speech. Kobashikawa's system [5] and Imoto's system [6] evaluate the rhythm of stress in English using Hidden Markov Models. Ito's system [7] uses the duration of a word and the pause insertion error rate as prosodic features, and the distance between prosodic features of a learner's speech and the speech of a native speaker is used as a rhythm score. In Kato's system [8], the slope of pitch corresponding to a word boundary is used as a prosodic feature.

Regrettably, these systems have lower performance than that of a pronunciation evaluator. In the present paper, we develop a new prosody evaluator with new prosodic features and word importance factors [9]. The proposed system evaluates both the intonation and rhythm of a learner's speech. The rhythm score and intonation score are calculated using prosodic features and are independently used for evaluation. However, some prosodic features corresponding to rhythm affect the evaluation of intonation. Therefore, we also propose a method of integrating the rhythm score and the intonation score.

In the present paper, the target language is English, and the native language of the learners is Japanese.

2. OVERVIEW OF THE SYSTEM

Figure 1 shows an overview of the proposed system. The basic scheme of the prosody evaluation is as follows:

0. Collect spoken sentences uttered by native speakers in advance. Prosodic features (rhythm and intona-

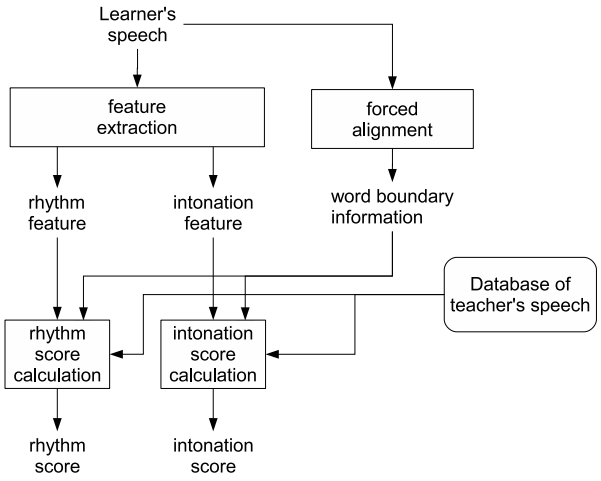

Figure 1: Block diagram of prosody evaluation.

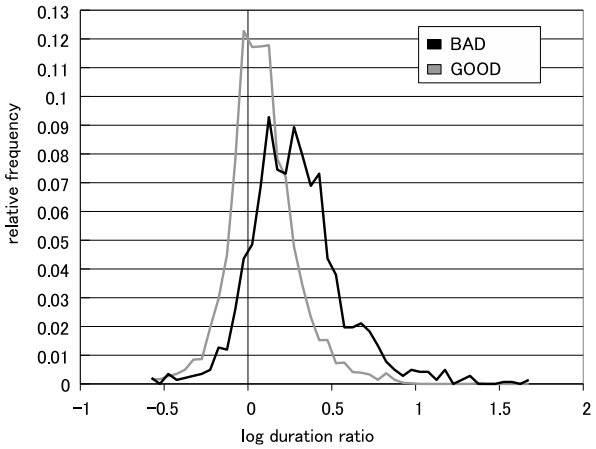

Figure 2: Histogram of the log-ratio of word duration.

tion features) are extracted from these sentences and are split word-by-word. These data will be used as reference data.

1. Extract prosodic features from the utterances of the learner.
2. Calculate the distances between words in the reference and learner data for both rhythm and intonation features. In this step, all reference data given by native speakers are used for calculation, and the smallest distance is used as the rhythm or intonation score.
3. Calculate the sentence score by the weighted sum of word scores.

In the proposed system, the rhythm and intonation scores are calculated for each word in the speech of the learner, and the total score is calculated by the sum of all scores in a sentence. In order to divide the input sentence into words, the "forced alignment" algorithm, which is a speech recognition technology, is used.

3. EVALUATION OF RHYTHM

Word duration ratio

Rhythm is made by patterns of stress and non-stress in a sentence [10]. Excellent rhythm is obtained by correct patterns of stress in words and correct durations of words.

The system proposed by Ito [7] uses two prosodic features, the relative duration of a word and the pause insertion error rate. The relative duration of the word is calculated as the duration of the word divided by the duration of the sentence. This feature indicates the correctness of the rhythm from the duration point of view. The pause insertion error rate is used as an indicator as to whether all of the prosodic phrases are uttered without pause. In general, a prosodic phrase should be uttered without pause. If a pause is inserted in a prosodic phrase, the rhythm is corrupted.

In this system, the relative duration of the word is used as a prosodic feature. This feature is not influenced by the speed of the utterance. If a learner utters a sentence having relatively the same duration as the utterance of the teacher, the score must be the highest, whether the learner speaks slowly or quickly.

However, the rhythm score given by native speakers correlates with the speed of the utterance. Figure 2 shows a histogram of the word duration ratio between the learner's speech and the teacher's speech. The X-axis represents the log-scaled duration ratio, where "0" indicates that duration of the word uttered by the learner is exactly the same as that uttered by the teacher. All of the learner's speeches were evaluated by native speakers using a five-grade scale, where 5 indicates "excellent rhythm" and 1 indicates "very poor rhythm". In this figure, a "GOOD" histogram is constructed by a learner's speech that is scored as 4 or higher, and a "BAD" histogram is constructed by a learner's speech that is scored equals to or lower than 2.

This figure indicates that there is correlation between the correctness of rhythm and the speaking speed. Many learner's speeches with higher rhythm scores were uttered at the same or a slightly lower rate than the teacher's speech. In the proposed system, the word duration ratio between the learner's speech and the teacher's speech is used as a prosodic feature.

The word duration ratio $R_L(k)$ is calculated by the following equation:

$$R_L(k) = \frac{\max\left\{L(k), L^S(k)\right\}}{\min\left\{L(k), L^S(k)\right\}} \quad (1)$$

where $L(k)$ and $L^S(k)$ denote the duration of the k-th word uttered by the learner and the teacher, respectively. $R_L(k)$ is 1 only if the durations are exactly same, and if the duration uttered by the learner increases or decreases, $R_L(k)$ becomes larger.

If a learner speaks slowly, $R_L(k)$ is large and the rhythm score may be 1 ("very poor"). If a learner speaks more slowly, $R_L(k)$ becomes very large, but the rhythm score remains 1. This means that the rhythm score shows a

plateau. In order to represent this plateau, the sigmoid function is introduced.

$$x_{ratio}(k) = \frac{e^{\gamma(R_L(k)-1)} - 1}{e^{\gamma(R_L(k)-1)} + 1}, \quad \gamma \geq 0 \tag{2}$$

γ is a pre-defined parameter and was set to 1.7 in the experiments.

Stress pattern in a word

The pattern of stress and non-stress in a word is one of the most important features. However, few studies have used this pattern as a prosodic feature. In the proposed system, the stress pattern in a word is used as a prosodic feature.

The stress pattern score is calculated for each word. First, the average log-power is calculated from all of the frames, and the log-power of each frame is normalized by subtracting the average. The stress pattern score is then calculated as the Dynamic Time Warping (DTW) distance between the learner's and teacher's log-power sequences.

The DTW is one of the most popular algorithms in the field of pattern recognition research and can be used to find the optimum correspondence between two sequences. The DTW distance can be calculated using following equations:

$$g(i,j) = \min \begin{cases} g_1(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g_2(i,j) \end{cases} \tag{3}$$

$$g_1(i,j) = \min_{2 \leq m \leq r} g(i-m, j-1) + 2d(i-m+1, j) + \sum_{t=0}^{m-2} d(i-t, j) \tag{4}$$

$$g_2(i,j) = \min_{2 \leq m \leq r} g(i-1, j-m) + 2d(i, j-m+1) + \sum_{t=0}^{m-2} d(i, j-t) \tag{5}$$

where $d(i,j)$ denotes the distance between normalized log-powers of the i-th frame in the learner's speech and the j-th frame in the teacher's speech, and r is a pre-defined parameter that can control how far-located frames can be made to correspond. Note that $g(1,1) = d(1,1)$, and $g(i,0) = g(0,j) = \infty$ for all i and j.

Finally, the score of the stress pattern for a word k is calculated using the following equation:

$$x_{DP}(k) = \frac{1}{I_k + J_k}\, g(I_k, J_k) \tag{6}$$

where I_k and J_k denote the number of frames in the k-th word of the learner's speech and the teacher's speech, respectively.

Combination of prosodic features

The rhythm score $x_{rh}(k)$ for word k is defined by the weighted sum of $x_{ratio}(k)$ and $x_{DP}(k)$.

$$x_{rh}(k) = (1-w) \cdot x_{ratio}(k) + w \cdot x_{DP}(k) \tag{7}$$

where w denotes a weighting factor and is set by hand.

4. EVALUATION OF INTONATION

Intonation is mainly represented by the flow of pitch. In the proposed system, the flow of log-power is also considered, because an utterance with a higher pitch may have a higher power. Four features, namely, the pitch, log-power, and first-order regression coefficients of both features, are used as prosodic features. Both pitch and log-power are normalized by subtracting the corresponding average values.

For each frame, the correspondence between the learner's speech and the teacher's speech is estimated using the DTW algorithm, and the weighted sum of the distance between corresponding frames is calculated. The weight $w_k(i)$ of the i-th frame of the k-th word is defined as the multiplicative inverse of the standard deviation of the frame calculated by the speech of several teachers. This means that a frame with a small weight has significant variation in the teacher's speech, and the frame is not important for the evaluation of intonation.

Let $c(i)$ be the frame number of the teacher's speech corresponding to the i-th frame of the learner's speech. Here, $c(i)$ is estimated by DTW. The weight is calculated by the following equation:

$$w_k^d(i) = \frac{1/\sigma_k^d(i)}{\sum_{j=1}^{I_k} 1/\sigma_k^d(j)} \tag{8}$$

where $\sigma_k^d(i)$ denotes the d-th dimension of the standard deviation of the i-th frame of the k-th word.

$$\sigma_k^d(i) = \sqrt{\frac{1}{M} \sum_{s=1}^{M} \left(v_s^d\left(c_s(i)\right) - \bar{v}^d\left(c_s(i)\right) \right)^2} \tag{9}$$

where $v_s^d(i)$ denotes the d-th dimension of the prosodic feature vector of the i-th frame uttered by teacher s, and M denotes the number of teachers.

The distance between the i-th frame of the learner's speech and $c(i)$-th frame of the teacher's speech is calculated by the following equation:

$$D_k(i) = \sqrt{\sum_{d=1}^{4} w_k^d(i) \left(u^d(i) - v^d\left(c(i)\right)\right)^2} \tag{10}$$

Finally, the intonation score of the k-th word is calculated by the following equation:

$$y_{int}(k) = \frac{1}{N_k} \sum_{i=1}^{N_k} D_k(i) \tag{11}$$

5. CALCULATION OF SENTENCE SCORE USING WORD IMPORTANCE FACTOR

Word importance factor

We defined rhythm and intonation scores for each word. After the calculation of these scores, the sentence score

should be calculated by summing these word scores. However, native speakers appear to evaluate a learner's prosody by focusing on several keywords. In order to emulate such an evaluation strategy, the word importance factor is introduced, and the sentence score is calculated as a weighted sum of the word scores.

Let α_{ij} be the word importance factor of the j-th word of the i-th sample uttered by a learner. This factor is estimated by the ordinary least squares method. The error Q is defined as follows:

$$Q = \sum_i \left(\frac{1}{K_i} \sum_{j=1}^{K_i} \alpha_{ij} x_i(j) + \beta - e_i \right)^2 \quad (12)$$

where $x_i(j)$ denotes the prosody score ($x_{rh}(j)$ or $y_{int}(j)$) of the i-th sample, K_i denotes the number of words in the i-th sample, and e_i denotes the prosodic score (rhythm score or intonation score) given by native speakers. The ordinary least squares method can estimate α and β with minimum Q. After estimation, the sentence score S_i can be calculated using estimated values of α and β, as follows:

$$S_i = \frac{1}{K_i} \sum_{j=1}^{K_i} \alpha_{ij} x_i(j) + \beta \quad (13)$$

The word importance factor α_{ij} should be estimated separately for each sample and word. However, it is very difficult to estimate robustly because there are few samples for estimation. In order to solve this problem, the word importance factor is clustered using a decision tree.

Clustering of the word importance factor

One reasonable way to estimate α robustly is based on α, which is commonly used for each vocabulary. For instance, α_{the} is estimated using the word "the" in all samples. In this method, many samples can be used for the estimation of α. However, α cannot represent the difference of position in a sentence or the sentence style (such as a declarative sentence or a question).

In order to estimate α more robustly and flexibly, a decision tree clustering algorithm is introduced. Figure 3 shows an example of a decision tree. In this algorithm, a number of questions regarding the nature of words are prepared in advance, and a word cluster is divided into

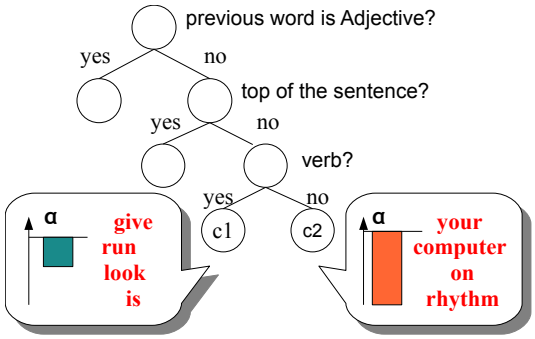

Figure 3: Example of a decision tree.

two clusters using appropriate questions. The question with highest correlation coefficient between scores given by native speakers and that given by the system is selected as the appropriate question.

The details of the algorithm are as follows:

Step 1 Make a root node L_0 in the tree. All of the words in the training samples are included in the root node.

Step 2 Select the node L_i that has greatest number of words.

Step 3 **Step 4** and **Step 5** are carried out for all of the questions $Q_1 \cdots Q_M$.

Step 4 Divide the words in node L_i into two new nodes L_{yes} and L_{no} using question Q_j. If the number of words in node L_{yes} or node L_{no} is less than a predefined threshold θ, cancel the division using Q_j.

Step 5 Estimate α using the ordinary least squares method. All of the words in the same node use the same α. After estimation, the correlation coefficient $r(Q_j)$ between scores given by native speakers and the system is calculated.

Step 6 Select the question $\hat{Q}$ with the highest $r(\hat{Q})$, and divide the node L_i into two new nodes using the question $\hat{Q}$. If none of the questions can be used because the number of words in the new node is smaller than θ, exit this algorithm. Otherwise, go to **Step 2**.

Appropriate clusters can be acquired using this algorithm, and the number of nodes can be controlled by θ.

6. EXPERIMENTS

Experimental conditions

Several experiments were carried out in order to confirm the effectiveness of the proposed system. An English speech database read by Japanese students [11] was used as the learners' speech. All of the data were evaluated with respect to both rhythm and intonation by four native speakers. A total of 68 questions (examples are shown in Table 2) were prepared for decision tree clustering, and a 4-fold cross validation technique was used for an open test. Shirokaze's method [12] was used for extracting pitch. The other experimental conditions are shown in Table 1.

Evaluation of rhythm

First, the correlation between the scores given by four evaluators is checked. Table 3 shows the correlations between the evaluators. In this table, "mean" denotes the correlation between a score given by an evaluator and the average score calculated from three other scores. This table indicates that scores given by evaluators varied widely. The maximum correlation between evaluators was 0.57, and the correlation between an evaluator and the average score was slightly high. In the experiments, the average score was used as the scores given by native speakers.

Table 4 shows the results of rhythm evaluation with several prosodic features. We examined the proposed features, the duration ratio of the word between the learner's speech and the teacher's speech (A) and the DTW distance of the normalized log-power (B). Moreover, we also

Table 1: Experimental conditions.

Evaluation of rhythm		
	Learner's data	190 speakers (95 males, 95 females) 944 sentences 3—18 words/sentence
	teacher's data	19 speakers (8 males, 11 females)
Evaluation of intonation		
	Learner's data	190 speakers (95 males, 95 females) 938 sentences 2—18 words/sentence
	teacher's data	18 speakers (7 males, 11 females)
Evaluator		4 Americans (2 males, 2 females)
Scores		5 (Excellent) – 1 (Very poor)
Threshold θ		3

Table 2: Examples of questions used for creating a decision tree.

Is the part of speech of the current word a noun?
Is the part of speech of the previous word an adverb?
Are there less than three syllables in the word?
Is the word located at the end of the prosodic phrase?
Is the word located at the top or the second of the sentence?
Is the sentence a negative sentence?

Table 3: Correlation coefficients of rhythm evaluation between evaluators

Evaluator	B	C	D	mean
A	0.54	0.56	0.47	0.64
B	-	0.46	0.57	0.65
C	-	-	0.44	0.58
D	-	-	-	0.60

examined conventional features, the relative duration of words (C) and the pause insertion error ratio (D). In this experiment, -1.0 indicates the best correlation coefficient because the system outputs a "distance". A larger distance indicates poorer prosody, whereas a larger score given by evaluators indicates better prosody.

From these results, the conventional method gave a very low correlation. According to a previous study [7], the conventional method gave a higher correlation. The reason for this is that the prosodic phrase boundary was given.

Table 4: Comparison of features for rhythm evaluation.

Features	Correlation
(A)	-0.53
(B)	-0.45
Weighted sum of (A) and (B)	-0.55
(C)	0.14
(D)	-0.11
Product of (C) and (D) [7]	-0.04

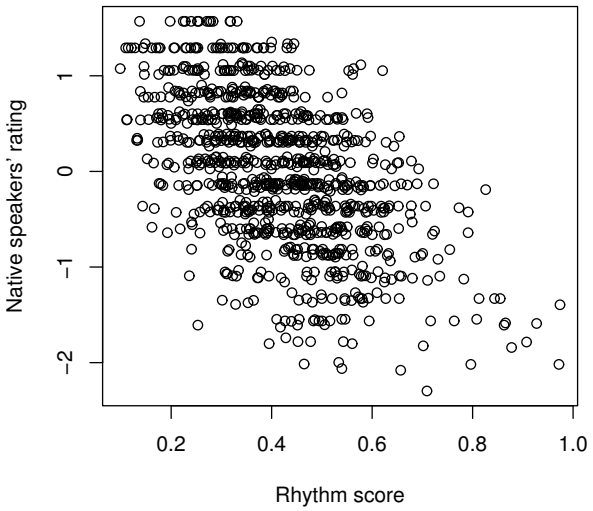

Figure 4: Scatter plot of the rhythm scores vs. scores given by human evaluators

In other words, the experiments in the previous study had easier condition than that in the present paper.

On the other hand, the proposed features gave correlations of -0.45 or better. In this experiment, the weighting factor used for the combination of (A) and (B) was set to 0.128. There is statistically significant difference between all pairs of conventional features and proposed features. There is also statistically significant difference between (A) and (B). However, there is no difference between (A) and the combination of (A) and (B). Figure 4 shows the correlation between scores given by evaluators and the distances given by the combination of (A) and (B). Note that scores given by the evaluators were normalized by subtracting the average score.

We attempted to apply the word importance factor with decision tree clustering, however, this was not effective for rhythm evaluation.

Evaluation of intonation

Table 5 shows the correlation coefficients among evalua-

Table 5: Correlation coefficients of intonation evaluation between evaluators

Evaluator	B	C	D	mean
A	0.46	0.50	0.55	0.65
B	-	0.37	0.48	0.55
C	-	-	0.39	0.51
D	-	-	-	0.60

Table 6: Results of intonation evaluation.

Features	Correlation
$(F_0, \Delta F_0)$	-0.29
$(POW, \Delta POW)$	-0.26
$(F_0, \Delta F_0, POW, \Delta POW)$	-0.27

Table 7: Results of intonation evaluation using importance factor estimation.

Word importance factor	Correlation
without	-0.27
with (closed)	0.59
with (open)	0.45

tors. Intonation scores given by evaluators also varied widely. The correlation among evaluators was approximately 0.4—0.5, and the correlation among the evaluators and the average score was 0.5—0.65.

Table 6 shows the correlation for each feature. In this table, F_0 denotes the pitch, POW denotes the normalized log-power, and Δ denotes the first-order regression coefficients.

$(F_0, \Delta F_0)$ was used in the conventional method [7] and provided low correlation. There is no statistically significant difference between any of the pairs of feature sets in the table.

However, the word importance factor improved the performance. Table 7 shows the results obtained with the word importance factor. The word importance factor improved the correlation coefficients to 0.45. The relationship between the proposed score and the scores given by the evaluators is shown in Figure 5.

7. INTEGRATION OF BOTH SCORES

Correlation between rhythm and intonation scores In previous sections, we proposed the rhythm score and the intonation score, and these scores were used independently for the evaluation of each prosodic factor. However, some prosodic features corresponding to rhythm may affect the evaluation of intonation, or vice versa. An utterance with good rhythm causes the evaluator to evaluate

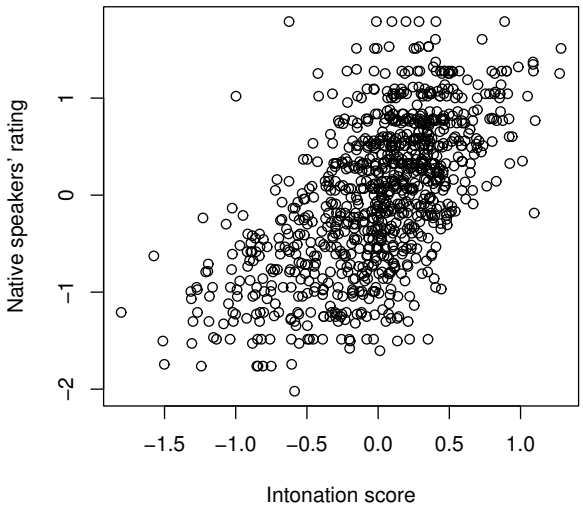

Figure 5: Scatter plot of intonation scores vs. scores given by human evaluators

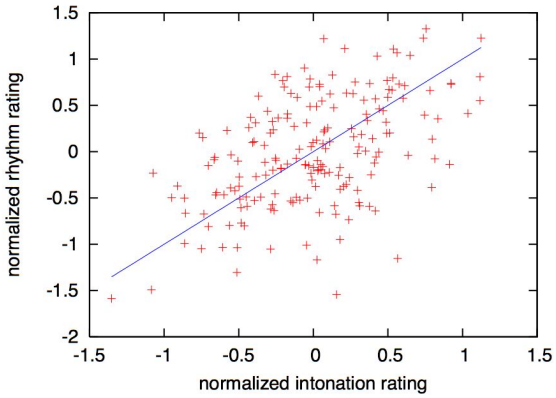

Figure 6: Correlation between rhythm and intonation scores given by evaluators.

the intonation highly, and an utterance with good intonation causes the evaluator to evaluate the rhythm highly. In other words, there may be a correlation between the rhythm score and the intonation score.

In order to confirm this hypothesis, we have investigated the correlation between the intonation score and the rhythm score. Figure 6 shows the correlation between the rhythm and intonation scores given by evaluators. In this figure, each score was normalized by subtracting the average score. The blue line indicates $y = x$.

This figure indicates that there is correlation between the rhythm score and the intonation score. The correlation coefficient is 0.50. An utterance with good rhythm has good intonation, and an utterance with poor rhythm has

Table 8: Results of intonation evaluation using integration of both scores

	Intonation only	Both scores
Closed	0.59	0.64
Open	0.45	0.48

poor intonation. Therefore, the rhythm score x_{rh} is useful for evaluating not only rhythm, but also intonation. The intonation score y_{int} is also useful for both evaluations.

In this section, we propose a new evaluation score that is calculated using both x_{rh} and y_{int}.

Integration of two scores

The new evaluation score $\tilde{S}_i$ of the i-th sample is calculated by the approach described in Section 5. The new score can be defined as follows:

$$\tilde{S}_i = \frac{1}{K_i} \sum_{j=1}^{K_i} \left(\alpha_{ij} x_{rh,i}(j) + \beta_{ij} y_{int,i}(j) + \gamma \right) \qquad (14)$$

where $x_{rh,i}(j)$ and $y_{int,i}(j)$ denote the rhythm and intonation scores of the j-th word of the i-th sample, respectively. α_{ij}, β_{ij}, and γ can be estimated by the ordinary least squares method to minimize the following error function:

$$Q = \sum_i \left\{ \frac{1}{K_i} \sum_{j=1}^{K_i} \left(\alpha_{ij} x_{rh,i}(j) + \beta_{ij} y_{int,i}(j) + \gamma - e_i \right) \right\}^2 \qquad (15)$$

The word importance factors α_{ij} and β_{ij} are also clustered using the decision tree clustering.

Note that the new evaluation score $\tilde{S}_i$ is not used for evaluating the *total* prosody, which means both rhythm and intonation. When evaluating rhythm, rhythm scores given by evaluators are used as e_i in Eq. (15), and three parameters (α, β, and γ) are estimated for rhythm evaluation. As a result, $\tilde{S}_i$ is used as the rhythm score. In the same manner, if the intonation is to be evaluated, another three parameters are estimated using the intonation scores given by the evaluator, and $\tilde{S}_i$ is used as the intonation score.

Evaluation experiments

In order to investigate the effectiveness of the new score $\tilde{S}_i$, several experiments were carried out. The experimental conditions are the same as those described in Section 6.

Table 8 shows the correlation coefficients between the score given by the evaluators and the proposed score for evaluating intonation. The integration of the rhythm score and the intonation score improves the correlation coefficient from 0.45 to 0.48 in the open condition, which means that the prosodic features corresponding to rhythm affect the evaluation of intonation.

On the other hand, the integration method was not effective for rhythm evaluation. The correlation coefficient was 0.51. However, the rhythm score gave a correlation coefficient of -0.55 (shown in Table 4). The integration method could not outperform the evaluation using only the rhythm score.

8. CONCLUSION

A prosodic evaluation method for English has been developed. The proposed method evaluates the rhythm and intonation of a learner's speech. For rhythm evaluation, the word duration ratio and normalized log-power were used as prosodic features. The correlation coefficient between scores given by native evaluators and that obtained by the proposed method was -0.55.

For intonation evaluation, the normalized log-power, pitch, and first-order regression coefficients of both features were used, and the word importance factor was also introduced. A decision tree was used for clustering of the word importance factor in order to obtain a robust estimation. The proposed method gave a correlation coefficient of 0.45. Moreover, we also proposed a method by which to integrate the rhythm score with the intonation score in order to introduce the effectiveness of a prosodic feature corresponding to rhythm to the intonation evaluation. This provided a correlation coefficient of 0.48, which is a higher correlation coefficient than that given by the intonation score alone.

Both the results of rhythm and intonation evaluation are statistically significant compared with the results of the conventional method.

9. ACKNOWLEDGMENT

The present study was supported in part by a JSPS Grant-in-Aid for Scientific Research (B)16300260 and (B) 20320075.

10. REFERENCES

[1] F. Ehsani and E. Knodt, "Speech technology in computer-aided language learning: strengths and limitations of a new CALL paradigm," *Language learning and technology*, vol. 2, no. 1, pp. 45–60, 1998.

[2] M. Eskenazi, "Using automatic speech processing for foreign language pronunciation tutoring: some issues and a prototype," *Language learning and technology*, vol. 2, no. 2, pp. 62–76, 1999.

[3] G. Kawai, A. Ishida, and K. Hirose, "Detecting and correcting mispronunciation in non-native pronunciation learning using a speech recognizer incorporating bilingual phone models," *Journal of the acoustical society of Japan*, vol. 57, no. 9, pp. 569–580, 2001, (in Japanese).

[4] A. Cutler, D. Dahan, and W. Donselaar, "Prosody in comprehension of spoken language: A literature review," *Language and Speech*, vol. 40, no. 2, pp. 141–201, 1997.

[5] S. Kobashikawa, N. Minematsu, K. Hirose, and D. Erickson, "Modeling of stressed syllables for their detection in English sentences to develop an English

rhythm learning system," Technical Report of IEICE NLC2001-65, SP2001-100, The institute of Electronics, Information and Communication Engineers, 2001, (in Japanese).

[6] K. Imoto, M. Dantsuji, and T. Kawahara, "Automatic error detection of English sentence stress spoken by Japanese for CALL system," The 2001 autumn meeting of the acoustical society of Japan 3–7–2, The Acoustical Society of Japan, 2001, (in Japanese).

[7] A. Ito, T. Nagasawa, H. Ogasawara, M. Suzuki, and S. Makino, "Automatic detection of English mispronunciation using speaker adaptation and automatic assessment of English intonation and rhythm," *Educational Technology Research*, vol. 29, pp. 13–23, 2006.

[8] K. Kato, Y. Yamashita, K. Nozawa, and Y. Shimizu, "Prosodic scoring of the English learners' speech based on utterance comparison for word boundaries," The 2002 autumn meeting of the acoustical society of Japan 1–6–3, The Acoustical Society of Japan, 2002, (in Japanese).

[9] A. Ito, T. Konno, M. Suzuki, and S. Makino, "Improvement of automatic English prosody evaluation based on word clustering using a decision tree," *The IEICE Trans. Information and Systems (Japanese Edition)*, vol. J86-D-II, no. 2, pp. 195–203, Feb. 2003, (in Japanese).

[10] I. Kawagoe, *Eigo no Onsei wo Kagaku suru*, Taishukan Publishing Co., Ltd., 1999, (in Japanese).

[11] N. Munematsu, K. Nishina, and S. Nakagawa, "Read speech database for foreign language learning," *Journal of the acoustical society of Japan*, vol. 59, no. 6, pp. 345–350, 2003, (in Japanese).

[12] T. Shirokaze, S. Makino, and K. Kido, "Extraction of fundamental frequency using temporal continuity over an input speech," *Trans. IEICE*, vol. 73-A, no. 9, pp. 1537–1539, 1990, (in Japanese).