Evaluation of English Intonation based on Combination of Multiple Evaluation Scores

Akinori Ito, Tomoaki Konno, Masashi Ito and Shozo Makino

Graduate School of Engineering, Tohoku University, Sendai, Japan
{aito,tomoaki,itojin,makino}@makino.ecei.tohoku.ac.jp

Abstract

In this paper, we proposed a novel method for evaluating intonation of an English utterance spoken by a learner for intonation learning by a CALL system. The proposed method is based on an intonation evaluation method proposed by Suzuki et al., which uses “word importance factors,” which are calculated based on word clusters given by a decision tree. We extended Suzuki’s method so that multiple decision trees are used and the resulting intonation scores are combined using multiple regression. As a result of an experiment, we obtained correlation coefficient comparable to the correlation between human raters.

Index Terms: computer-assisted language learning system, prosody evaluation, intonation, decision tree, multiple regression

1. Introduction

Nowadays, English skills are crucial for non-native English speakers. Computer-assisted language learning (CALL) systems have been used for self-learning of English. In addition to reading, writing and listening exercises, some CALL systems have a function for speaking lesson [1, 2]. 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.

Besides correct pronunciation of phonemes and words, prosody plays an important role in English communication [3]. The traditional CALL system provides a learner with multimedia teaching materials for learning English intonation [4]. However, it is difficult for a learner to assess his/her own intonation. Therefore, it is important for a CALL system to 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. The CALL system by Tsubota et al. [5] evaluates the stress of the learner’s speech. We developed a system for evaluating English intonation and rhythm as well as pronunciation of words [6]. The accuracy of automatic prosody evaluation was improved by incorporating word-by-word evaluation framework [7].

In this paper, we make further improvement of evaluation of English intonation by enhancing Suzuki’s work [7]. The essential improvement of our work is to incorporate a framework that combines multiple intonation scores, each of which is calculated using an utterance by one teacher.

2. Evaluation of Intonation based on Word Importance Factor

2.1. Overview

First, we explain the framework of intonation evaluation by Suzuki et al. [7]. Figure 1 shows the schematic diagram of the intonation evaluation. This system first extracts features for intonation evaluation from English utterance by Japanese learner. At the same time, the input utterance is split into words using forced alignment. The extracted features are then compared with those from teacher utterances word by word, and then an intonation score of the utterance is calculated. Then the score is fed back to the learner.

As there are two or more teacher utterances for one utterance of the learner, the system calculates intonation scores for all teacher utterances. Then the best score among the all scores is chosen as the final score [6].

2.2. Feature extraction

We exploited fundamental frequency (F0), log power and their first derivatives as features. The F0 is normalized using forced alignment. Then the log power is normalized using the maximum value. Using the four-dimensional feature vectors, the teacher utterances are aligned with the input speech of a learner using DP matching. Then the weighted Euclidean distances between the learner’s and teacher’s utterances are calculated frame by frame.

2.3. Calculation of intonation score

The frame-based distances are accumulated word-by-word, and then averaged for all words in a sentence. Let $D_{k}^{(i)}(i)$ be the distance between the learner’s utterance and the $i$-th teacher’s utterance.
one at the \(i\)-th frame in the \(k\)-th word in a sentence. Let \(N_k\) be the number of frames of the \(k\)-th word. Then the intonation score of the \(k\)-th word is calculated as follows:

\[
y_{int}(k, t) = \frac{1}{N_k} \sum_{t=1}^{N_k} D_k^{(i)}(t) \quad (1)
\]

Then the intonation score of the entire utterance is calculated by averaging \(y_{int}(k, t)\) for all words.

\[
s_{int}(t) = \frac{1}{K} \sum_{k=1}^{K} y_{int}(k, t) \quad (2)
\]

Here, \(K\) is the number of words in a sentence. Finally, the best score among all \(t\) is chosen as the final score of the learner’s utterance. As the score given by Eq. (2) is based on distance, the minimum score is the best one.

\[
s_{int} = \min_{t} s_{int}(t) \quad (3)
\]

### 2.4. Introduction of word importance factors

In the evaluation score using Eq. (2), all words are evaluated in equal importance. 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 \(\alpha_{ik}\) be a word importance factor of the \(k\)-th word of the \(i\)-th sample uttered by a learner. This factor is estimated by the least squares method. Let \(K_i\) be the number of words in the \(i\)-th utterance, and \(y_{int,k}(i)\) is the intonation score of the \(k\)-th word of the \(i\)-th utterance with respect to the \(t\)-th teacher utterance. Then the error \(Q\) is defined as follows:

\[
\hat{t}_i = \arg\min_t \frac{1}{K_i} \sum_{k=1}^{K_i} y_{int,k}^{(i)}(k, t) \quad (4)
\]

\[
Q = \sum_i \left( \frac{1}{K_i} \sum_{k=1}^{K_i} \alpha_{ik} y_{int,k}^{(i)}(k, \hat{t}_i) + \gamma - e_i \right)^2 \quad (5)
\]

Here, \(e_i\) is the human rating of the \(i\)-th utterance, where an utterance with the best intonation has the smallest value. \(\gamma\) is a bias to be optimized with \(\alpha_{ik}\). On determining \(\alpha_{ik}\), it is difficult to determine individual \(\alpha\) for each of distinct words in the sentences because the number of words in the training sentences are not sufficient. Therefore, we cluster words in all training sentences using a decision tree, and the coefficients \(\alpha_{ik}\) are determined by cluster. Figure 2 shows an example of the decision tree for clustering. First, we prepare questions concerning words or sentences. Then a decision tree is generated using the questions so that all words in a cluster share a word importance factor \(\alpha_{ik}\) and the estimation error \(Q\) becomes minimum by clustering using the decision tree.

Let \(C(i, k)\) be a mapping from the \(k\)-th word to the \(i\)-th utterance to the cluster to which the word belongs, \(\alpha_{C(i, k)}\) and \(\gamma\) be the coefficients that minimize the error \(Q\). Then the intonation score of an utterance is calculated as follows:

\[
\hat{t} = \arg\min_t \frac{1}{K} \sum_{k=1}^{K} y_{int}(k, t) \quad (6)
\]

\[
\hat{S} = \frac{1}{K} \sum_{k=1}^{K} (\alpha_{C(k)} y_{int}(k, \hat{t}) + \gamma) \quad (7)
\]

where \(K\) is the number of words in a sentence to be evaluated and \(C(k)\) is a mapping from the \(k\)-th word to a cluster determined by the decision tree.

### 2.5. Combining rhythm feature

In addition to the features for intonation, features for rhythm evaluation are also introduced into the intonation evaluation. Word duration ratio and DP distance are used as features of rhythm evaluation [7]. Using a rhythm score \(y_{rh}(k, t)\) calculated from the rhythm features, the intonation score is calculated as follows.

\[
\hat{u}_i = \arg\min_u \frac{1}{K_i} \sum_{k=1}^{K_i} y_{rh}^{(i)}(k, u) \quad (8)
\]

\[
S_i = \frac{1}{K} \sum_{k=1}^{K} (\alpha_{C(i,k)} y_{int}(k, \hat{t}_i) + \beta_{C(i,k)} y_{rh}^{(i)}(k, \hat{u}_i)) + \gamma \quad (9)
\]

\[
Q = \sum_i (S_i - e_i)^2 \quad (10)
\]

Then the parameters \(\alpha, \beta\) and \(\gamma\) are optimized so that the error \(Q\) is minimized.

The evaluation of accuracy of the intonation evaluation method is conducted by observing correlation between the human rating values and the scores calculated by the system, which was 0.48 [7].

### 3. Combination of scores using multiple regression

#### 3.1. Problems of the conventional framework

Although the accuracy of intonation evaluation can be improved by introducing the word importance factors, further improvement should be made. To this end, we propose a framework that calculates an intonation evaluation score for each teacher utterance and combines multiple scores using a multiple regression.

In the conventional framework, at first the teacher utterance with the minimum distance is determined utterance by utterance (Eq. (4) and (8)), and then a decision tree is generated so that the error by using the scores of the selected teachers (Eq. (10)) is minimized.
This framework has two problems. One problem is that only one teacher is considered for evaluating one utterance. As explained above, calculation of the intonation score is on the “nearest-neighbor” basis. Therefore, distances between the utterance and reference utterances by other than the nearest teacher are just ignored, though those distances involve some information that contributes to the intonation score. The second problem is concerned with combination of the intonation and rhythm features. When calculating the score, the nearest teachers for intonation feature and rhythm feature are chosen individually (Eq. (4) and (8)). However, it may cause improper evaluation of an utterance when the selected intonation pattern and rhythm pattern are inconsistent.

3.2. Intonation score calculation by combining multiple scores

To solve these problems, we propose a novel framework of calculating an intonation score. The overview of the proposed framework is shown in Fig. 3.

First, the intonation score of a word \( y_{int}^{(i)}(k,t) \) and the rhythm score \( y_{rhythm}^{(i)}(k,t) \) are calculated for all words of all training sentences with respect to all teacher utterances. Then a decision tree for determining word classes is trained teacher by teacher. Determination of the decision tree is based on Eq. (9) and (10), but the teachers \( t_i \) and \( \hat{t_i} \) are substituted to the teacher to which the decision tree is trained.

After training decision trees, the scores for the teacher utterances are combined. Let \( S_i(t) \) be an intonation score of the learner’s \( i \)-th utterance calculated using the teacher utterance \( t \). We tested the following four methods:

Minimum: This method is similar to the conventional framework, where the best score is chosen as the final score.

\[
\hat{S}_i = \min_t S_i(t)
\]  

(11)

Maximum: This method chooses the worst score as the final score.

\[
\hat{S}_i = \max_t S_i(t)
\]  

(12)

Median: This method chooses the median of the all scores.

\[
\hat{S}_i = \text{median}[S_i(1), S_i(2), \ldots, S_i(T)]
\]  

(13)

Regression: In this method, all scores are linearly combined [8].

\[
S_i = \sum_t a_t \hat{S}_i(t)
\]  

(14)

The coefficients \( a_t \) are optimized for minimizing the errors of the intonation scores with respect to the human rating values.

4. Experiments

4.1. Experimental conditions

We conducted experiments for evaluating the proposed intonation evaluation. The experimental conditions are shown in Table 1. We used Julian [9] for segmenting the input utterances into words. When training decision trees, we prepared 61 questions concerning a word and a sentence, as shown in Table 2. As the methods need training, we conducted 4-fold cross validation. The evaluation metric of the methods were correlation coefficient between the rating values by human evaluators and the system’s scores. As there are four raters for each utterance, we first normalized the rating values rater by rater so that the average became zero and the variance became one. Then all of the normalized rating values for an utterance were averaged, which was used as a human-rating value for calculation of correlation.

4.2. Effect of score combination

At first, we conducted an experiment for investigating effect of score combination. In this experiment, Maximum number of word classes was limited to 20, and minimum number of words in one class was limited to 3.

Table 3 shows the experimental result. In this table, the “Training” column denotes how to train the decision tree. In
Table 3: Correlation coefficients by all methods

<table>
<thead>
<tr>
<th>Training</th>
<th>Evaluation</th>
<th>Correlation (w/o rhythm)</th>
<th>Correlation (with Rhythm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimum</td>
<td>minimum</td>
<td>0.442</td>
<td>0.434</td>
</tr>
<tr>
<td>each teacher</td>
<td>minimum</td>
<td>0.447</td>
<td>0.500</td>
</tr>
<tr>
<td>each teacher</td>
<td>maximum</td>
<td>0.440</td>
<td>0.468</td>
</tr>
<tr>
<td>each teacher</td>
<td>median</td>
<td>0.451</td>
<td>0.502</td>
</tr>
<tr>
<td>each teacher</td>
<td>regression</td>
<td>0.474</td>
<td>0.539</td>
</tr>
</tbody>
</table>

![Figure 4: Effect of the number of nodes on correlation coefficient](image_url)

From the results shown in Table 3, we can confirm that the proposed method gave the greater correlation than the conventional method. The combination method using the multiple regression was the best one among the four kinds of score combination methods. The best correlation was 0.539 when using the rhythm score and score combination by the multiple regression. We conducted experiments for investigating performance of the proposed method, and we obtained significant improvement over the conventional method. The resulting performance of the proposed method was comparable to a human rater.

4.3. Effect of number of classes

Next, we changed the maximum number of word classes for training of decision trees. In this experiment, five methods (one conventional method and four proposed methods) were examined. Rhythm features were incorporated in this experiment. Figure 4 shows the correlation coefficients with respect to the maximum number of word classes. This result reveals that the combination method based on the multiple regression gave the best method regardless of the number of classes. The optimum number of classes was around 20.

5. Conclusion

In this paper, we proposed a new method for evaluating intonation of a learner’s English utterance. Our method is based on the evaluation method by Suzuki et al. that uses a decision tree for estimating word importance factor. We extended Suzuki’s method so that multiple regression trees were trained and the scores were combined using the multiple regression. We conducted experiments for investigating performance of the proposed method, and we obtained significant improvement over the conventional method. The resulting performance of the proposed method was comparable to a human rater.

One of remaining problems is how to feedback the evaluation results to the learner. The visualization of pitch movement is the most basic way [4], but it seems also important to show what part of intonation of an utterance is bad and what part is important for improving naturalness of the intonation. More research is needed to design the best way of showing a learner these kinds of information.

6. Acknowledgments

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

7. References


