Automatic Evaluation of English Pronunciation Based on Speech Recognition Techniques

Hiroshi HAMADA, Satoshi MIKI, and Ryohei NAKATSU

NTT Human Interface Laboratories
1-2356 Take, Yokosuka, Kanagawa 238-03, Japan

Abstract. A new method is proposed for automatically evaluating the English pronunciation quality of Japanese speakers. It is assumed that this can be determined based on three criteria: the static characteristics of phonetic spectra, the dynamic structure of spectrum sequences, and the prosodic characteristics of utterances. Evaluation is carried out by comparing English words pronounced by a Japanese with those pronounced by a native speaker using speech recognition techniques. Preliminary experiments show the evaluation results obtained using the proposed method to correspond well with human judgement of pronunciation quality.

1. Introduction

Several kinds of speech processing techniques, such as speech recognition, speech synthesis and speech coding, have been studied for communications or information input/output purposes. Recently, attention has been focused on education as one area in which speech processing techniques can be applied. For example, a speech trainer for the handicapped and an English pronunciation trainer for Japanese [1] have already been developed. With this type of speech trainer, the student attempts to reproduce the formant frequencies of vowels and fundamental frequencies of words or sentences vocalized by a native speaker and displayed on a screen.

This paper proposes a new quantitative method for automatically evaluating the pronunciation of English. English pronunciation quality is determined by the accuracy of phonetic pronunciation and English pronunciation fluency is determined by the smoothness of phonetic transitions and the accuracy of prosodic characteristics. The method proposed here evaluates these characteristics quantitatively by comparing English pronounced by a Japanese with that pronounced by a native speaker. A speaker adaptation technique is used for evaluating the static characteristics of the phonetic spectra. A pattern matching technique is used to evaluate the dynamic structure of spectral sequences and the prosodic characteristics. Although it would be ideal to evaluate pronunciation quality based on whole sentences, words are used for the evaluation in this preliminary study.

Section 2 describes the new method for evaluating static and dynamic characteristics and also prosodic characteristics. Section 3 gives the results of evaluation experiments using 441 words spoken by 10 native and 10 Japanese speakers. Finally, Section 4 provides a summary and some conclusions.

2. Method of Evaluating English Pronunciation

The pronunciation quality of English words pronounced by a Japanese is evaluated based on three characteristics using speech processing techniques: (1) the static characteristics of the phonetic spectra, (2) the dynamic characteristics of spectrum sequences, and (3) the prosodic characteristics. Since direct evaluation of these characteristics would be difficult, evaluation is carried out by comparing them in words spoken by Japanese and native speakers.

2.1 Evaluation of Static Characteristics of Phonetic Spectra

The static characteristics of phonemes are evaluated by measuring the stability of mapping vectors, which express the correspondence between the phonetic spectra of the Japanese speakers and those of the native speakers. Codebook vectors obtained by vector quantization are used as the phonetic spectra set. Mapping vectors are obtained through speaker adaptation using the codebook of the Japanese speakers and that of the native speakers. Figure 1 schematically shows the process for evaluating static characteristics. The process comprises the following steps:

- Subject (Japanese)
- Reference (Native Speaker)
- Codebook Generation (for Subject and Reference)
- Vector Quantization
- Original Codeword of Subject
- Speaker Adaptation
- Mapping Vector Calculation
- Mapped Codeword
- Evaluation of Mapping Vector
(1) Pronunciation of word set
A Japanese speaker (the subject speaker) and a native speaker (the reference speaker) pronounce a pre-defined word set.

(2) Codebook generation
Two codebooks are generated using the LBG method [2]. One codebook is generated using speech data obtained from the subject speaker and the other using speech data obtained from the reference speaker.

(3) Speaker adaptation
Codebook vectors of the subject are mapped onto the codebook vector space of the reference speaker. The speaker adaptation method used in this mapping process was that proposed by Shikano et al [3]. This method has two processing stages: (i) finding the vector correspondence between the two codebooks through DTW (dynamic time-warping) matching and generating a histogram of vector correspondence, and (ii) generating a mapped codebook using the histogram.

(4) Mapping vector calculation
Mapping vectors are calculated by subtracting the vectors in the original codebook from the vectors in the mapped codebook. Mapping vectors are defined as follows:

\[ V_i = B_i - A_i \]

where
\[ V_i : \text{the i-th mapping vector}, \]
\[ N : \text{the number of vectors in the codebook}, \]
\[ B_i : \text{the i-th vector of the original subject codebook}, \]
\[ A_i : \text{the i-th vector of the mapped codebook}. \]

Some examples of mapping vectors obtained through the above process are provided in Fig. 2. Fig.2(a) is an example of mapping vectors obtained when the subject mispronounces a phoneme (for example, pronouncing /r/ instead of /l/, and vice versa). If the subject’s pronunciation is unstable and words are frequently mispronounced, the mapping vectors take the form shown in Fig.2(b) and the mapped codebook space degenerates. The mapping vectors in Fig.2(c) show the subject’s pronunciation to be correct but not clear. Whereas, Fig.2(d) shows what the mapping vectors look like when the subject’s pronunciation is good. The criteria for evaluating mapping vectors, therefore, should be the parameters which distinguish case (d) from cases (a)-(c).

(5) Evaluation of mapping vectors
The following two parameters are used to evaluate mapping vectors:

(i) Standard deviation of mapping vectors
Standard deviation of mapping vectors is used as one evaluation criterion, based on the idea that all mapping vectors should comprise the same elements if the word is pronounced correctly. Accordingly, evaluation criterion \( E_{Sl} \) is defined as

\[ E_{Sl} = \frac{1}{N} \sum_{i=1}^{N} (V_i - V_{ave})^2 \]

where \( V_{ave} \) is the average vector of \( V_i \) (i=1,...,N).

(ii) Ratio of vector variance
The ratio of original codebook variance to mapped codebook variance is used as the other evaluation criterion. That is,

\[ E_{S2} = \frac{\text{min}(D_A, D_B)}{\text{max}(D_A, D_B)}, \]

where \( D_A \) and \( D_B \) are the variances of original and mapped codebooks, respectively.

Evaluation criteria \( E_{S1} \) and \( E_{S2} \) are used in the evaluation of the static characteristics of phonetic spectra.

2.2 Evaluation of Dynamic Characteristics of Spectral Sequences
Dynamic characteristics of spectral sequences are evaluated based on the spectral distance between the subject and reference speaker’s pronunciation of words. Spectral distance is calculated by DTW matching the subject’s pronunciation of words through a mapped codebook and the reference speaker’s pronunciation of the same words. In addition to the difference in the dynamic structure of spectral sequences, DTW matching also reveals distortion in speaker adaptation. However, this distortion will be treated and compensated for in the dynamic evaluation stage. A block diagram of the process for evaluating dynamic characteristics is provided in Fig. 3. The following procedure is used to evaluate the dynamic characteristics of spectral sequences:

(1) Pronunciation of word set
The subject speaker and the reference speaker pronounce a pre-defined word set. It is better to use a word set different from that used in the evaluation of the static characteristics of phonetic spectra.

(2) Vector quantization
Vector quantization of the pronunciation of the subject and of the reference speaker is carried out using the codebooks obtained in Section 2.1.

(3) DTW matching and evaluation
DTW matching is performed using the Double SPLIT method [4]. In the distance calculation, the mapped codebook obtained in Section 2.1(3) is used for the subject speaker in place of the original codebook. The average spectral distance \( E_{D} \) over all the words in the word set is used as the criterion for evaluating the dynamic characteristics of spectral sequences.

2.3 Evaluation of Prosodic Characteristics
Several quantitative analyses have been conducted on the prosodic characteristics of English uttered by Japanese speakers. In this study, in order
to achieve a quantitative evaluation of prosodic characteristics, the prosodic features of English uttered by Japanese speakers were evaluated in comparison to the prosodic features of utterances by native speakers.

Fundamental frequency (pitch) and speech power are used as parameters for evaluating prosodic features. Frame-by-frame difference between reference and subject pronunciation patterns are obtained through DTW matching. The evaluation criterion is, then, obtained by accumulating the difference between corresponding frames.

This evaluation process, illustrated in Fig. 4, is described below:

(1) Utterance for evaluation
The same word set used for evaluating spectral characteristics is employed in the evaluation of prosodic features. It is, therefore, not necessary to have the speakers pronounce another word set for the purpose of this evaluation.

(2) Extraction of prosodic parameters
Fundamental frequency and speech power are extracted from the pronunciation of the subject and reference speakers using the following procedure:

(i) Fundamental frequency is extracted using the partial correlation method.

(ii) The extracted pitch pattern is smoothed to remove such errors as half or double pitch.

(iii) The smoothed pattern is transformed into logarithmic form and then normalized by subtracting the average value.

(iv) The speech power pattern is also transformed into logarithmic form and normalized by subtracting the average value.

(3) Calculation of prosodic evaluation values
Two types of evaluation methods are proposed. One takes the difference in dynamic range between the two patterns into consideration. The other does not. Procedures for both of these evaluation methods are described below.

(i) Absolute difference value of prosodic features without normalization
The absolute difference value of prosodic features is obtained for each corresponding frame of words spoken by the Japanese and native speakers, and the average value of the absolute difference values is obtained for the hole speech period of each word uttered by a subject speaker. Then, these values are accumulated to obtain the average value of all words uttered by the subject. As the result, evaluation criteria for fundamental frequency \((E_{pf1})\) and for speech power \((E_{pp1})\) are obtained.

(ii) Absolute difference value of prosodic features with normalization
To eliminate the effect of dynamic range from the evaluation of prosodic features, it is necessary to make the deviation value of the words to be compared coincide. Each of the prosodic parameters obtained in (2) above is divided by its deviation value. Then, the same evaluation calculation process described in (i) above is carried out. Finally, evaluation criteria for fundamental frequency \((E_{pf2})\) and for speech power \((E_{pp2})\) are obtained.

3. Evaluation Experiments
Pronunciation evaluation experiments were conducted to confirm the validity of the proposed method. This section describes the experimental conditions and gives the experimental results obtained.

3.1 Experimental Conditions
(1) Speech database
The word set used 441 words taken from an English textbook used in Japanese junior high schools. Phoneme balance and phonetic variation were taken into consideration. Ten Japanese male speakers and ten native male speakers pronounced the word set as subject speakers and reference speakers, respectively. One half of the word set (221 words) was used for speaker adaptation and evaluation of the static characteristics of phonetic spectra, and the other half (220 words) was used for evaluation of the dynamic characteristics of spectral sequences and the prosodic characteristics.

(2) Speech analysis
Speech data was passed through a 5.2 kHz low pass filter, sampled at a rate of 12,000 samples/s and then converted into 12-bit digital signals. Then, after passing through 16-msec Hamming window, 16-order LPC cepstrum parameters were obtained through LPC analysis, which was carried out every 8 msec. Fundamental frequency and speech power were obtained every 8 msec through 32 msec analysis window.

The LBG method was used for codebook generation, and 256 vectors were generated.

(3) DTW matching method
A symmetric DTW matching method with a slope constraint of 1/3 to 3 was employed in the DTW stage of speaker adaptation and word matching processes [5].

3.2 Human Judgement of Pronunciation Quality
Pronunciation quality was also evaluated by a native speaker who is an expert in English education. The evaluation was carried out in 3 steps:

(1) The word set pronounced by each Japanese speaker was divided into 5 subsets. (2) The expert graded each subset of pronounced words on a scale of up to 100 points. (3) The average score over the 5 subsets was used as the human judgement value of
each Japanese speaker’s English pronunciation quality. The evaluation values by the expert were distributed from 73 to 91 points, and the average value of the 10 Japanese speakers was 78.7 points.

3.3 Experimental Results

In the experiment, 7 evaluation criteria (Es1, Es2, Ed, Epf1, Epp1, Epf2, Epp2) were obtained by comparing each of the Japanese speakers with each of the 10 native speakers. Next, the average evaluation value of 10 native speakers and the minimum evaluation value of the 10 native speakers were calculated. Finally, each Japanese speaker was assigned 14 evaluation values. Table 1 shows the correlation coefficients between the evaluation criteria and human judgement of pronunciation quality.

Evaluation criterion Es1 correlates with human judgement, whereas Es2, on the other hand, shows no correlation.

Evaluation criterion Ed shows fairly good correlation with human judgement which is explained by the fact that it combines the evaluation factors for adaptation distortion, phoneme mispronunciation, and fluency of phoneme-to-phoneme transition.

Among the evaluation criteria based on the fundamental frequency, Epf2 shows good correlation with human judgement. The evaluation criterion Epf1, on the other hand, shows only a slight correlation. With regard to the evaluation criteria using speech power characteristics, Epp1, an evaluation criterion without normalization, shows good correlation, while Epp2, a criterion with normalization, does not.

Multi-regression analysis was conducted on the results obtained, through which a correlation coefficient of 0.90 was obtained when the evaluation criteria Es1,min, Ed,ave, Epf2,ave, and Epp1,ave were used. Figure 5 plots the relation between human judgement of pronunciation quality and the evaluation results obtained through the above described process.

The experimental results show the automatic evaluation of English pronunciation quality through the use of speech processing techniques to be promising.

4. Conclusion

A new method is proposed for evaluating the English pronunciation ability of Japanese speakers. The method evaluates pronunciation quality based on the static characteristics of phonetic spectra, the dynamic characteristics of speech sequences, and the prosodic characteristics. Evaluation of the static characteristics of the phonetic spectra is achieved by calculating the standard deviation of the mapping vectors, which map the codebook vectors of the Japanese speaker onto the vector space of the native speaker. The spectral distance between words pronounced by the Japanese speaker and those pronounced by the native speaker is used to evaluate the dynamic characteristics of speech sequences. The differences in fundamental frequency and speech power between the pronunciation of the native and Japanese speaker are used as the evaluation criteria for prosodic characteristics.

Evaluation experiments were carried out using 441 words spoken by 10 Japanese speakers and 10 native speakers. The results show the correlation between the evaluation scores, obtained through a combination of the above three factors, and the scores obtained through human judgement to be 0.90.

Although the Japanese subjects used in the experiment had a similar level of pronunciation ability and the amount of speech data was not sufficient to obtain statistically reliable results, the evaluation results obtained using the proposed method did show good correspondence with human judgement of pronunciation quality.

![Figure 5 Relation between Evaluation Results and Human Judgements](image)

Table 1 Correlation Coefficients between Evaluation Criterion and Human Judgement

<table>
<thead>
<tr>
<th>Evaluation Criterion</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Es1,ave</td>
<td>0.57</td>
</tr>
<tr>
<td>Es1,min</td>
<td>0.55</td>
</tr>
<tr>
<td>Es2,ave</td>
<td>0.06</td>
</tr>
<tr>
<td>Es2,min</td>
<td>0.19</td>
</tr>
<tr>
<td>Ed,ave</td>
<td>0.57</td>
</tr>
<tr>
<td>Ed,min</td>
<td>0.66</td>
</tr>
<tr>
<td>Epf1,ave</td>
<td>0.22</td>
</tr>
<tr>
<td>Epf2,ave</td>
<td>0.77</td>
</tr>
<tr>
<td>Epp1,ave</td>
<td>0.57</td>
</tr>
<tr>
<td>Epp2,ave</td>
<td>0.86</td>
</tr>
<tr>
<td>Epp2,min</td>
<td>0.76</td>
</tr>
<tr>
<td>Epp3,ave</td>
<td>0.34</td>
</tr>
<tr>
<td>Epp3,min</td>
<td>0.27</td>
</tr>
</tbody>
</table>

References