ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 1aSCb: Digital Speech Processing (Poster Session)

1aSCb16. An investigation of vowel substitution rules in the automatic evaluation system of English pronunciation

Kei Sato, Masaharu Kato and Tetsuo Kosaka*

*Corresponding author's address: Graduate School of Science and Engineering, Yamagata University, 4-3-16 Jonan, Yonezawa, 992-8510, Yamagata, Japan, tkosaka@yz.yamagata-u.ac.jp

We investigate the performance improvement of an automatic evaluation system of the English pronunciation of Japanese learners. In this system, Japanese and English acoustic models are used to detect mispronunciation at a phoneme level. Hidden Markov models (HMMs) are used as acoustic models. Mispronunciation is detected by comparing the output likelihoods of the two models. In order to improve the performance of this system, we investigate certain mispronunciation rules, which represent common mispronunciations among Japanese learners. We use four mispronunciation rules: vowel insertion (at the end of a word), vowel substitution, vowel insertion (between consonants), and consonant substitution. In this system, the accuracy of the mispronunciation rules is particularly important. The rules are determined on the basis of the knowledge of phonetics in our previous system. However, the effectiveness of the rules has not been analyzed quantitatively, and we do so in this work. A knockout procedure is used to select effective rules. By selecting effective rules, we found that the correlation coefficient between the subjective evaluation value and the system performance improved from 0.757 to 0.858.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

In this study, we investigate the performance improvement of an automatic system used to evaluate the English pronunciation of Japanese students. Many research approaches have been previously proposed in this field. For example, Kawai et al. proposed a pronunciation error detection system that utilizes a speech recognition technique [1]. In this system, errors can be detected using two types of acoustic models. One model is trained based on English sentences uttered by native speakers, and the other uses sentences uttered by non-native speakers. In addition, using various normalization methods, Kusumi et al. proposed a performance improvement technique for an error detection system [2]. They reported that a combination of cepstral mean normalization (CMN) and a histogram equation (HEQ) [3] can effectively reduce acoustic mismatches between an input utterance and the acoustic models.

The basic configuration of these systems is shown in Figure 1. First, the system displays an English sentence to be pronounced, and the learner utters the sentence as is. The input speech is then analyzed to obtain the feature vectors. The displayed sentence is then automatically translated into a phoneme sequence that represents the pronunciation of the sentence. This sequence contains phonemes for both correct and incorrect pronunciations. The incorrect sequences are derived from mispronunciation rules that represent mispronunciations common among Japanese learners. Next, the phoneme sequence is converted into a hidden Markov model (HMM) sequence that will be used for a process called “forced alignment.” This process determines the best assignment for the feature vectors generated from the speech analysis module using the Viterbi algorithm. During this procedure, the single best state path is determined by selecting either the Japanese or English HMMs. If the Japanese HMMs are selected, or if the correct English HMMs are not selected, pronunciation errors are detected.

As illustrated from this procedure, mispronunciation rules are essential to this system. Such rules represent mispronounced phoneme sequences that frequently occur by Japanese learners while speaking English. We evaluated eight types of mispronunciation rules [2], and then selected the following four types for use in the system: vowel insertion (at the end of a word), vowel substitution, vowel insertion (between consonants), and consonant substitution. We determined these rules by referring to existing knowledge and information from phonetics (e.g., [4], [5]). However, the effectiveness of these rules has not yet been analyzed quantitatively.

In this study, we will evaluate the effectiveness of the mispronunciation rules in an objective manner. Based on this evaluation, we will try to improve the performance of the system by selecting more effective rules. A knockout procedure will be used for this selection. The system performance will be measured by comparing the results of subjective assessments conducted by English teachers. The rules that do not correlate with the results of these subjective assessments will be judged as inappropriate. Such inappropriate rules will be removed to improve the performance. In this work, we focus on vowel substitution rules because they reportedly exhibit lower performance results than the others [2].

FIGURE 1. A block diagram of the proposed automatic pronunciation evaluation system
EVALUATION SYSTEM

Overview

As we noted in the previous section, the basic configuration of the evaluation system is shown in Figure 1. Evaluation is carried out using the forced alignment technique with both English and Japanese acoustic models. We use HEQ to reduce any mismatch between acoustic models and test data. Moreover, as many pronunciation errors are detected even in the speech of native speakers, we use a weighting method to reduce such errors.

Histogram Equalization

Another problem is that an acoustic mismatch between the training data and test data can have a negative influence on evaluation. Therefore, better performance can be achieved by using normalization techniques to remove only such mismatches. In this research, HEQ (Histogram equalization) is used for normalization. The transform function $HEQ()$ is given by

$$o'_i = HEQ(o_i) = C_T^{-1}(C_E(o_i)),$$

where $C_E$ and $C_T$ denote the CDFs estimated from the test and training data, respectively.

Weighting Method

In our previous study, we faced the problem of frequent detectable pronunciation errors even in the speech of native speakers. Therefore, in this study, a weighting method is used to reduce such errors. Weighting of the output likelihood is performed as follows. $b_i(o_i)$ denotes the output probability in state $S_i$. Weighting is carried out by calculating $\lambda b_i(o_i)$, where $\lambda$ represents the weight of a Japanese phoneme and is set at least less than 1.0. Through this method, the errors of native speakers can be reduced.

CONSIDERATIONS OF VOWEL SUBSTITUTION RULES

In the proposed system, pronunciation error rules are used to detect mispronunciation of Japanese learners. These rules are categorized into four groups.

- **Vowel insertion (at the end of word):**
  The rule in which a Japanese vowel is inserted after an English consonant at the end of a word
  Example: sing (s i η u)

- **Vowel substitution:**
  The rule in which an English vowel is replaced by a Japanese vowel
  Example: the (ð æ → ð æ)

- **Vowel insertion (between consonants):**
  The rule in which a Japanese vowel is inserted between English consonants
  Example: study (s u i t η d i)

- **Consonant substitution:**
  The rule in which an English consonant is replaced by a Japanese consonant
  Example: child (ʃ η ʃ a l d → ʃ η ʃ a r d)

Compared with other rules, the vowel substitution rules had the most detrimental results. Fourteen vowels (æ, æ, a, è, o, ë, u, æ, ë, o, ë, u, æ, ë, o, ë) are used to define vowel substitution rules. Table 1 shows substitution rules for each English vowel. For example, the English vowel /æ/ can be substituted for the Japanese vowel /æ/. These vowel rules are determined based on knowledge of phonetics [2], and the validity of the rules has not actually been examined. In this paper, we examine whether these vowel rules have evaluation validity, and aim to improve the accuracy of the system by using only effective rules. The knockout procedure is used to select effective rules. In the experiment, we determine which rules should be removed to achieve the greatest performance improvement possible. Rule elimination is executed repeatedly, until performance no longer improves.
### TABLE 1. Vowel substitution rules

<table>
<thead>
<tr>
<th>English vowel</th>
<th>Japanese vowel</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>æ</td>
<td>a</td>
<td>apple, cat</td>
</tr>
<tr>
<td>å</td>
<td>a</td>
<td>bgt, love</td>
</tr>
<tr>
<td>ø</td>
<td>a</td>
<td>about, police</td>
</tr>
<tr>
<td>øː</td>
<td>o</td>
<td>tep, hot</td>
</tr>
<tr>
<td>øː</td>
<td>aa</td>
<td>member, record</td>
</tr>
<tr>
<td>i</td>
<td>i</td>
<td>sit, king</td>
</tr>
<tr>
<td>iː</td>
<td>i or ii</td>
<td>meet, see</td>
</tr>
<tr>
<td>u</td>
<td>uu</td>
<td>book, cook</td>
</tr>
<tr>
<td>uː</td>
<td>uu or uuu</td>
<td>boot, blew</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
<td>head, beg</td>
</tr>
<tr>
<td>æː</td>
<td>æː</td>
<td>day, fail</td>
</tr>
<tr>
<td>ō</td>
<td>o or oo</td>
<td>call, walk</td>
</tr>
<tr>
<td>ōː</td>
<td>o or oo</td>
<td>home, blow</td>
</tr>
</tbody>
</table>

### EXPERIMENTAL CONDITIONS

#### Evaluation of System Performance

To evaluate system performance, we use 1,900 English sentences uttered by 190 Japanese speakers (95 males and 95 females) in the ERJ database [6]. Each Japanese speaker utters 10 sentences and English teachers use a five-grade evaluation to assess the speaker’s performance. Since four English teachers assign evaluation values to each sentence, the average of those values is calculated and used as a subjective evaluation value. The sentences are sorted in order of their subjective value, and every 50 sentences are grouped into one class. A pronunciation error rate for each class is calculated by dividing the mispronunciation count by the number of phonemes that may cause errors. This error rate is then subtracted from 1.0 to obtain the system’s accuracy. This accuracy is used as an objective evaluation value. The system performance can be evaluated based on the correlation between the subjective and objective evaluation values. Moreover, to check whether the system is performing correctly, accuracy rates are computed for 20 American English speakers. These Americans are native GA (General American) speakers. GA speakers have a standard American accent comparable to that generally used in television networks in the United States.

#### Acoustic Models and Speech Data

In the speech analysis module, a speech signal is digitized at a sampling frequency of 16 kHz with a quantization size of 16 bits. The length of the analysis frame is 32 ms, and the frame period is set at 8 ms. A 13-dimensional feature (with 12-dimensional MFCC and log power) is derived from the digitized samples for each frame. Further, Δ and ΔΔ features are calculated from the MFCC and log power. Thus, the total number of dimensions is 39. Each monophone HMM consists of three states and 16 mixture components per state. The training and test data are shown in Table 2. For native speakers, leave-one-out cross-validation is conducted to separate training and test data. In order to improve system performance further, normalization procedures are conducted for Japanese acoustic models and test data. For training Japanese acoustic models, each Japanese training utterance is normalized to the average parameters of English training data by HEQ. For evaluation, each test utterance spoken by Japanese and native speakers is normalized to the average parameters of English training data by HEQ.
TABLE 2. The training data and test data

<table>
<thead>
<tr>
<th>Training data</th>
<th>Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>- English acoustic model(GA)</td>
<td>- Japanese speakers</td>
</tr>
<tr>
<td>Speakers: 20 (8 males and 12 females)</td>
<td>Speakers: 190 (95 males and 95 females)</td>
</tr>
<tr>
<td>Number of sentences: 8,051</td>
<td>Number of sentences: 1900</td>
</tr>
<tr>
<td>- Japanese acoustic model(ASJ)</td>
<td>- Native speakers</td>
</tr>
<tr>
<td>Speakers: 204 (102 males and 102 females)</td>
<td>Speakers: 20 (8 males and 12 females)</td>
</tr>
<tr>
<td>Number of sentences: 31,511</td>
<td>Number of sentences: 399-403 per person</td>
</tr>
</tbody>
</table>

RESULTS AND DISCUSSIONS

The results of the vowel substitution rules are shown in Figure 2.

In the figure, a line parallel to the horizontal axis indicates a value of 0.757, which is the correlation coefficient of the vowel substitution rules defined by 14 vowels. This value is taken as the baseline of this system. Each bar graph shows the correlation coefficient when the evaluation was conducted by excluding a particular vowel from the vowel substitution rules. Moreover, the line graph shows the change in coefficients using a knockout procedure in which inappropriate vowels are removed in order of decreasing performance improvement.

The results of the bar graph indicate that performance improvement can be obtained by excluding /a/, /ɑ/, /i/, /ɛ/, or /u/. The results of the knockout procedure suggest that performance improvement can be obtained by excluding vowels from /a/ and /ɑ/ to /u/. However, performance declines when /ɑ/ is removed. The result indicates that the vowels /a/, /ɑ/, /i/, /ɛ/, /u/, /ou/, /œ/, and /u/ are not effective for pronunciation evaluation. Table 3 compares two rule sets. In one set, all 14 vowels are used, and, in the other, only six vowel rules are used (rules from /a/ to /u/ are removed). Weight in the table indicates the weight in relation to the output likelihood of Japanese phonemes. Since the difference in accuracy between Japanese and native speaker is very large, the system is thought to be working well.

From the results, it can be seen that selection of effective vowel substitution rules caused the correlation coefficient to improve from 0.757 to 0.858. We investigated the relationship between the knowledge from phonetics and our experimental results [4] [5]. Investigation shows that the vowels (e.g., /ɛ/, /æ/), which are reportedly difficult for Japanese speakers to pronounce accurately, appear to be effective for evaluating Japanese pronunciation.

**FIGURE 2.** Evaluation results obtained by omitting a certain vowel from the vowel substitution rule

![Graph showing correlation coefficient improvement](image)

**TABLE 3.** Comparison between the evaluation results of two rule sets

<table>
<thead>
<tr>
<th>Rule used</th>
<th>weight</th>
<th>accuracy for Japanese speakers [%]</th>
<th>accuracy for Native speakers [%]</th>
<th>correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>14 vowels</td>
<td>0.0100</td>
<td>85.39</td>
<td>97.08</td>
<td>0.757</td>
</tr>
<tr>
<td>/u, e, æ, œ, a, ə/</td>
<td>0.0085</td>
<td>79.53</td>
<td>96.99</td>
<td>0.858</td>
</tr>
</tbody>
</table>
CONCLUSIONS

In automatic pronunciation evaluation, the accuracy of mispronunciation rules is especially important. However, in our previous work, the effectiveness of the determined rules was not analyzed quantitatively.

In this work, we evaluated the effectiveness of the mispronunciation rules quantitatively to select effective rules. We focused on vowel substitution rules in the experiments. Owing to the selection of effective rules, the correlation coefficient improved from 0.757 to 0.858.

REFERENCES