
Kuniaki Hirabayashi and Seiichi Nakagawa

Department of Computer Science and Engineering
Toyohashi University of Technology, Japan
{kuniaki, nakagawa}@slp.cs.tut.ac.jp

Abstract
In this paper, we propose a method for estimating a score for English pronunciation.
Scores estimated by the proposed method were evaluated by correlating them with the learner’s pronunciation scores which was scored by native English teachers. The average correlation between the estimated pronunciation scores and the learner’s pronunciation scores over 1, 5, and 10 sentences was 0.807, 0.873, and 0.921, respectively. When a text of spoken sentence was unknown, we obtained a correlation of 0.878 for 10 utterances.

For English phonetic evaluation, we classified English phoneme pairs that are difficult for Japanese speakers to pronounce, using SVM, NN, and HMM classifiers. The correct classification ratios for native English and Japanese English phonemes were 94.6% and 92.3% for SVM, 96.5% and 87.4% for NN, 85.0% and 69.2% for HMM, respectively. We then investigated the relationship between the classification rate and a proficiency score of non-native learner’s English pronunciation, and obtained a high correlation of 0.6 ~ 0.7.

Index Terms: pronunciation evaluation, English, Japanese, HMM, SVM, Neural Network

1. Introduction
We have been investigating a CALL (Computer Assisted Language Learning) system that focuses on prosody and the effect of Japanese characteristics, and particularly on Japanese mannerisms in generating the correct emphasis for English words [1,2].

Many researchers have studied automatic methods for evaluating pronunciation proficiency. Neumeyer et al. proposed an automatic text-independent pronunciation scoring method for the French language, using HMM log-likelihood scores, segment classification error scores, segment duration scores, and syllabic timing scores [3]. The evaluation by segment duration performed better than the other methods. Furthermore, Franco et al. investigated an evaluation measure based on HMM-based phoneme log-posterior probability scores and a combination of the above scores [4]. We also investigated the posterior probability as an evaluation measure [5]. In addition, Franco et al. proposed a log-likelihood ratio score of native acoustic models to non-native acoustic models and found that this measure outperformed the posterior probability previously considered [6].

Cucchiarelli et al. compared the acoustic scores by TD (total duration of speech plus pauses), ROS (rate of speech: total number of segments/TD), and LR (likelihood ratio, corresponding to the posterior probability) and showed that TD and ROS correlated more strongly with the human ratings than LR [7].

All of the above studies considered European languages or English uttered by European non-native speakers. In addition, we evaluated Japanese uttered by foreign students [10]. Based on our previous work, Ohta et al. proposed a statistical method for evaluating the pronunciation proficiency of Japanese speakers when presenting in English [11].

In this paper, we propose a statistical method to estimate the pronunciation score for spoken English using new acoustic measures and pattern recognition techniques.

Regarding the new acoustic features, we used log-likelihood (forced alignment) based on the native English phoneme acoustic model for a given utterance, log-likelihood based on the Japanese English phoneme acoustic model, the log-likelihood ratio of these two features, English phoneme recognition likelihood from the English phoneme acoustic model, the ratio of log-likelihood and recognition log-likelihood from the English phoneme acoustic model, the recognition log-likelihood ratio of a native English phoneme acoustic model and Japanese English phoneme acoustic model, the recognition log-likelihood ratio of a native English phoneme acoustic model and Japanese syllabic acoustic model, the phoneme recognition ratio, the word recognition ratio, standard deviation of pitch and power, variation of spectral feature, and perplexity.

Scores estimated by the proposed methods are evaluated by their correlation with a learner’s pronunciation scores which were scored by native English teachers. The average correlation between the estimated scores and learner’s actual pronunciation scores over 1, 5, and 10 sentences was 0.807, 0.873, and 0.921, respectively. When a text of spoken sentence is unknown, we obtained a correlation of 0.878 over 10 utterances.

For English phoneme evaluation, we classified English phoneme pairs that are difficult for Japanese speakers to pronounce, using SVM (Support Vector Machine), NN (Neural Network), and HMM (Hidden Markov Model) classifiers. The correct classification ratios for native English and Japanese English phonemes were 94.6% and 92.3% for SVM, 96.5% and 87.4% for NN, 85.0% and 69.2% for HMM, respectively. We then investigated the relationship between the classification rate and a proficiency score of non-native learner’s English pronunciation, and obtained a high correlation of 0.6 ~ 0.7.

2. Database and System Overview
We used the Translanguage English Database (TED), presented at EuroSpeech, for the evaluation test data. Only a part of the TED has transcribed texts, consisting of 21(speakers) x 10 ~ 21(sentences) giving a total of 289 English sentences spoken by 21 male speakers who have good, average, or bad pronunciation proficiency. 16 of the 21 are Japanese speakers while the other 5 are native speakers from the USA. The pronunciation score used in this paper is the average of 2 scores, i.e., the phonetic pronunciation score and prosody (rhythm, accent, and intonation) score, as determined by five English teachers for the 289 sentences. The correlation between the English raters is 0.683, while that between a single English rater and the average...
We also used the ERJ (English Speech Database Read by Japanese) for the evaluation[12]. For this database, utterances of only 20 of the 76 Japanese speakers were assigned pronunciation scores by native English teachers. Scores for rhythm and intonation were allocated for every 5 utterances while those for segmental pronunciation were allocated for every 10 utterances. We used the TIMIT/WSJ database for training the native English phoneme HMMs, another Japanese speech database for adapting them (non-native English phoneme HMMs)[8] and the ASJ/JNAS database for training the native Japanese syllable HMMs (strictly speaking, mora-unit HMMs).

Table 1 gives a summary of the speech materials. The speech is downsampled to 16kHz and preemphasized, and then a Hamming window with a width of 25 ms is applied every 10 ms. A 12 dimensional MFCC (Mel Frequency Cepstrum Coefficient) is used as the speech feature parameter for each frame. Acoustic models based on monophone HMMs were trained by the analyzed speech. The English HMMs are composed of three states, each of which has four mixed Gaussian distributions with full covariance matrices, while the Japanese HMMs are composed of four states, each of which has four mixed Gaussian distributions with full covariance matrices.

Witt et al. found that for the pronunciation evaluation of non-native English speakers, triphones perform worse than monophones if the HMMs are trained by native speech; that is, less detailed (native) models perform better for non-native speakers[13][14][15].

Figure 1 presents a block diagram of our evaluation system for pronunciation score. Acoustic feature measures are extracted from the utterance and the pronunciation score estimated by corresponding regression models and phoneme-pair classification rates.

### Table 1: Speech materials used for training HMMs.

<table>
<thead>
<tr>
<th>HMM</th>
<th>speaker (database)</th>
<th># speakers</th>
<th>total # sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Native (TIMIT) (WSJ)</td>
<td>326</td>
<td>3260</td>
</tr>
<tr>
<td>Japanese students</td>
<td>50</td>
<td>6178</td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>Native (ASJ) (JNAS)</td>
<td>76</td>
<td>1065</td>
</tr>
<tr>
<td>Japanese</td>
<td>30</td>
<td>4518</td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>125</td>
<td>12703</td>
<td></td>
</tr>
</tbody>
</table>

3. Acoustic Feature Measures and Classification Methods for Minimum Phoneme - Pair

3.1. Explanation of Acoustic Measures

(a). Log-likelihood by native English HMM, non-native English HMM

We calculated the correlation rate between scores and the log-likelihood ($LL$) for a pronunciation dictionary sequence based on the concatenation of phoneme HMMs every 1, 5 and 10 sentences. The likelihood was normalized by the length in frames. We used native English phoneme HMMs ($LL_{native}$) and non-native English phoneme HMMs adapted by Japanese utterances ($LL_{non-native}$).

(b). Best log-likelihood for arbitrary phoneme sequences

The best log-likelihood for arbitrary phoneme sequences is defined as the likelihood of arbitrary phoneme (syllable) recognition without using phonotactic language models. We used native English phoneme HMMs ($LL_{best}$).

(c). Log-likelihood ratio

We used the log-likelihood ratio ($LR$) between native English HMMs and non-native English HMMs, defined as the difference between the two log-likelihoods, that is, $LL_{native} - LL_{non-native}$.

(d). A posteriori probability

We used the likelihood ratio ($LR'$) between the log-likelihood of native English HMMs ($LL_{native}$) and the best log-likelihood for arbitrary phoneme sequences ($LL_{best}$), giving the a posteriori probability, that is, $LL_{native} - LL_{best}$.

(e). Likelihood ratio for phoneme recognition

We used the ratio of the likelihood of arbitrary phoneme recognition between native English HMMs and non-native English HMMs ($LR_{non-native}$), defined as the difference between the two log-likelihoods, that is, $LL_{best, native} - LL_{best, non-native}$.
We used the correct rate, substitution rate, and deletion rate for arbitrary phoneme recognition. The test data are limited to the correctly transcribed parts by man2/4, which means that two teachers out of 4 transcribed the same label.

We used the correct rate for word recognition with a language model. The WSJ database (WSJ) or Eurospeech93 paper (EURO) was used to train the bigram language models[11].

The standard deviation of powers (Power) and fundamental (pitch) frequencies (F0) were calculated.

We used the correct rate of speech recognition with a language model. The WSJ database (WSJ) or Eurospeech93 paper (EURO) was used to train the bigram language models[11]. The test data are limited to the correctly transcribed parts by man2/4.

The standard deviation of powers (Power) and fundamental (pitch) frequencies (F0) were calculated.

We used the correct rate of speech recognition with a language model. The WSJ database (WSJ) or Eurospeech93 paper (EURO) was used to train the bigram language models[11]. The test data are limited to the correctly transcribed parts by man2/4.

We used the correct rate for word recognition with a language model. The WSJ database (WSJ) or Eurospeech93 paper (EURO) was used to train the bigram language models[11]. The test data are limited to the correctly transcribed parts by man2/4.

We used the correct rate of speech recognition with a language model. The WSJ database (WSJ) or Eurospeech93 paper (EURO) was used to train the bigram language models[11]. The test data are limited to the correctly transcribed parts by man2/4.

We chose 9 phoneme pairs for the evaluation of pronunciation: /l/r, /m/n, /s/sh, /s/th, /b/v, /b/d, /z/dh, /z/d, and /d/dh.

The target phoneme in an utterance was extracted by a forced Viterbi alignment based on HMMs. Then, the extracted part is classified into the target phoneme or the rival phoneme by the likelihood HMM, calculated on a by-frame-by-frame basis.

Five successive frames in the center of the extracted part were used as the input pattern for an SVM or NN classifier.

### 4. Estimating Pronunciation score

#### 4.1. Statistical Method for Acoustic Measures

Table 2 summarizes the correlation between each acoustic measure and the learner’s pronunciation score which was scored by native English teachers. Fairly high correlations were obtained for most of the acoustic feature measures (e.g. \( LL_{non-native} \), \( LR_{mother} \), \( LR_{adap} \), \( ROSS \)).

A linear regression model derived from the relationship between the acoustic measures and the learner’s scores was proposed for estimating the pronunciation score. We established various independent variables \( \{ x_i \} \) as parameters and the value \( Y \) as the learner’s score, and defined the linear regression model as

\[
Y = \sum \alpha_i x_i + \varepsilon, \quad (1)
\]

where \( \varepsilon \) is a residual [9][10]. The coefficients \( \{ \alpha_i \} \) were determined by minimizing the square of \( \varepsilon \). We experimented with both closed and open data for the speakers. Next, we investigated whether or not our proposed method was independent of the speaker. For the open experiment on speakers, we estimated the regression model using utterances from 20 speakers and estimated the score of the remaining speakers. We repeated this experiment for every speaker.

Table 3 summarizes the results of the pronunciation score for closed and open data at 1, 5, and 10 sentence levels. By combining certain acoustic measures, we obtained a correlation coefficient of 0.887 for pronunciation scores using open data at the 10 sentence level.

This confirms that the outcome of the proposed automatic estimation method for pronunciation score is almost the same as the evaluation by English teachers.

#### 4.2. Classification Method by HMM, SVM, and NN

Figures 2 and 3 illustrate the classification rates of minimum phoneme pairs by HMM, SVM, and NN. According to Figure 2, the average classification rates are about 95% by SVM, 94% by NN, and 82% by HMM for 8 native speakers. Figure 3, it can be seen that the average classification rates by SVM are about 95% for 8 native speakers and about 83% for 25 Japanese
5. Conclusion

We have proposed a statistical method for estimating the pronunciation score for non-native English speakers based on a linear regression model and a classification method for minimum phoneme pairs. By combining the measures, we are able to evaluate the pronunciation score with almost the same accuracy as English teachers. This approach is better than the classification based approach. In the future, we aim to combine an acoustic measure based approach with a phoneme pair classification approach.

As a next step in the development of the system, we aim to included hints or advice to the speakers to improve their pronunciation scores. For this purpose, the classification based approach should be effective.

6. References


