Automatic Evaluation of English Pronunciation by Japanese Speakers Using Various Acoustic Features and Pattern Recognition Techniques

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 score 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 $0.6 \sim 0.7$.

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].

Cucchiarini 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 (a 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 leaner'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 $0.6 \sim 0.7$.

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 $21(speakers) \times 10 \sim 21(sentences)$ 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

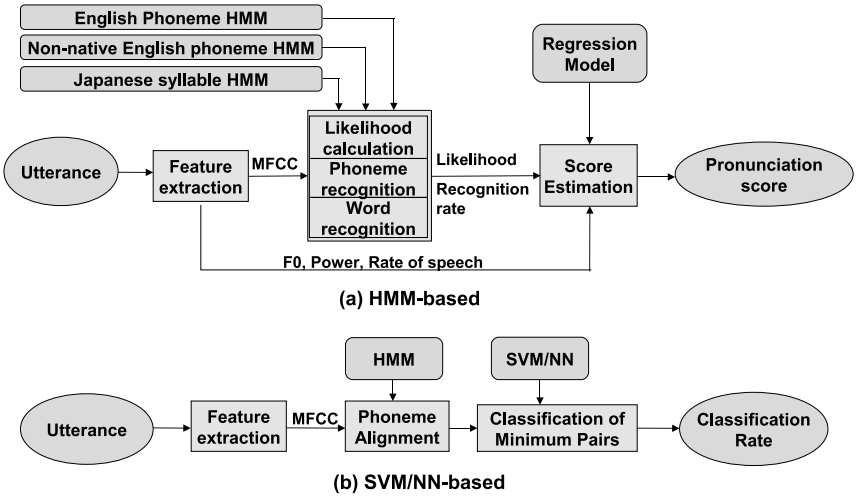

Figure 1: Block diagram of our estimation system for pronunciation score

of the other four is 0.794.

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. The acoustic features include 12 MFCCs, Δ and $\Delta\Delta$ features. 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.

HMM	speaker (database)	♯ speakers	total ♯ sentences
English	Native (TIMIT)	326	3260
	(WSJ)	50	6178
	Japanese students	76	1065
Japanese	Native (ASJ)	30	4518
	(JNAS)	125	12703

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, $LL_{native} - LL_{best}$ [8].

(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_{adap}), defined as the difference between the two log-likelihoods, that is, $LL_{best\_native} - LL_{best\_non-native}$. We also used the ratio of the likelihood of arbitrary phoneme (syllable) recognition between native English HMMs and native Japanese HMMs (LR_{mother}), defined as the difference between the two log-likelihoods, that is, $LL_{best\_native} - LL_{best\_mother}$.

Table 2: Correlation between acoustic measures and pronunciation score ("*" denotes a text-independent measure)

Measure	1 sentence	5 sentences	10 sentences
LL_{native}	-0.466	-0.625	-0.669
$\mathbf{LL_{non-native}}$	**-0.638**	**-0.771**	**-0.804**
LR	**0.800**	**0.859**	**0.880**
* LL_{best}	-0.473	-0.613	-0.660
* $\mathbf{LR_{mother}}$	**0.719**	**0.804**	**0.811**
* $\mathbf{LR_{adap}}$	**0.772**	**0.827**	**0.822**
LR'	0.214	0.273	0.349
Phoneme recog($Sub.$)	-0.298	-0.567	-0.662
Phoneme recog($Del.$)	0.056	0.116	0.220
Phoneme recog($Cor.$)	0.299	0.461	0.483
Word recog(WSJ, $Cor.$)	0.102	0.163	0.261
Word recog(EURO, $Cor.$)	0.113	0.256	0.281
* $Power$	-0.066	-0.057	-0.002
* $Pitch(F_0)$	0.495	0.638	0.691
Rate of speech	**0.523**	**0.692**	**0.773**

(f). Phoneme recognition result

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.

(g). Word recognition result

We used the correct rate for word recognition with a language model. The WSJ database (WSJ) or Eurospeech'93 paper (EURO) was used to train the bigram language models[11]. The test data are limited to the correctly transcribed parts by man2/4.

(h). Standard deviation of powers and F_0

The standard deviation of powers ($Power$) and fundamental(pitch) frequencies (F_0) were calculated.

(i). Rate of speech

We used the rate of speech of the sentence. Silences in the utterance were removed. We calculated each sentence's ROS as the number of phonemes divided by the duration in seconds.

3.2. Classification Methods

We used three classification methods for minimum phoneme pairs. Two were discriminative models based on an SVM and NN (Feed-forward Neural Network), respectively, while the third was a generative model based on an HMM-based method[16][17].

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.

(a). HMM

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.

Table 4: Correlation between phoneme pair classification rate and pronunciation score (20 Japanese speakers)

Correlation	intonation	rhythm	segmental
SVM	0.693	0.514	0.605
NN	0.734	0.471	0.567
HMM	0.508	0.418	0.124

(b). SVM/NN

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, LR_{mother}, LR_{adap}, ROS).

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

$$Y = \Sigma_i \alpha_i \times x_i + \varepsilon, \quad (1)$$

where ε is a residual [9][10]. The coefficients $\{\alpha_i\}$ were determined by minimizing the square of ε. 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. From 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

Table 3: Correlation between combination of acoustic measures and learner's pronunciation score by human raters

Number of sentences for evaluation	1 sentence		5 sentences		10 sentences	
Acoustic measures	CLOSED	SP.OPEN	CLOSED	SP.OPEN	CLOSED	SP.OPEN
$LL_{non-native}$, LR, LR_{mother}, $Power$, Phoneme recog($Del.$)	0.851	0.804	0.910	0.851	0.927	0.864
Word recog(EURO, $Cor.$), LR, $Power$, Word recog(WSJ, $Cor.$)	0.815	0.770	0.902	0.866	0.929	0.884
Word recog(EURO, $Cor.$), LR, $Power$	0.814	0.771	0.893	0.858	0.918	0.887
LL_{best}, LR_{mother}, $Power$	0.819	0.779	0.891	0.853	0.912	0.878

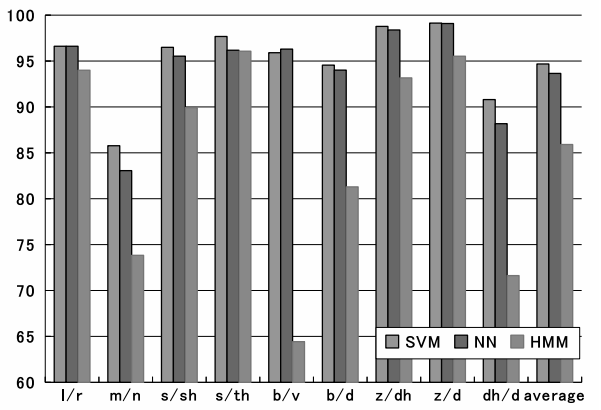

Figure 2: Classification rates for native speakers by SVM, NN and HMM

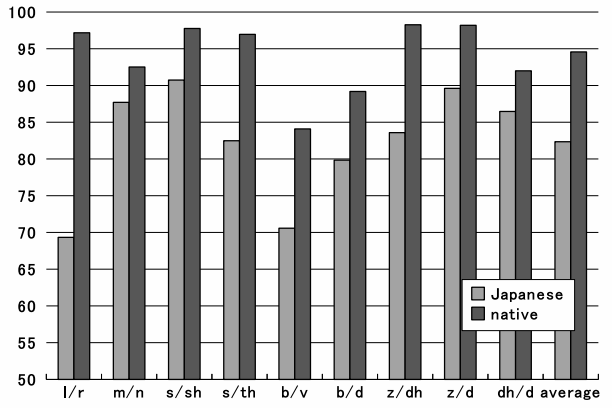

Figure 3: Classification rates for ERJ database by SVM

speakers. The correlation rates between classification rates and intonation, rhythm or segmental phoneme for every 10 utterances are summarized in Table 4. The correlation between the classification rate and segmental score is lower than that between the rate and intonation score. This shows that the evaluation of segmental pronunciation is very difficult even for native English speakers.

The correlation rate for segmental phonetics is 0.605, which is worse than the rate obtained using the statistical method for acoustic measures. Note however, that the evaluation data sets were different.

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

[1] Y. Fujisawa, N. Minematsu, and S. Nakagawa, "Evaluation of Japanese manners of generation word accent of English based on a stressed syllable detection technique," in *Proc. ICSLP*, pp.3103-3106, 1998.

[2] N. Nakamura, N. Minematsu, and S. Nakagawa, "Instantaneous estimation of accentuation habits for Japanese students to learn English pronunciation," in *Proc. EuroSpeech*, pp.2811-2814, 2001.

[3] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," in *Proc. ICSLP*, pp.1457-1460, 1996.

[4] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *Proc. ICASSP*, pp.1471-1474, 1997.

[5] Y. Taniguchi, A.A. Reyes, H. Suzuki, and S. Nakagawa, "An English conversation and pronunciation CAI system using speech recognition technology," in *Proc. EuroSpeech*, pp.705-708, 1997.

[6] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Proc. EuroSpeech*, pp.851-854, 1999.

[7] C. Cucchiarini, H. Strik, and L. Boves, "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms," in *Speech Communication*, 30(2-3), pp.109-119, 2000.

[8] S. Nakagawa, Allen A. Reyes, H. Suzuki, and Y. Taniguchi, "An English conversation CAI system using speech recognition technology," Trans. Information Processing Society in Japan. Vol.38 No. 8, pp. 1649-1657 (1997, in Japanese)

[9] S. Nakagawa, N. Nakamura, and K. Mori, "A statistical method of evaluating pronunciation proficiency for English words spoken by Japanese," IEICE Trans. Inf. & Syst., vol.E87-D, no.7, pp1917-1922, July 2004

[10] K. Ohta and S. Nakagawa, "A Statistical Method of Evaluating Pronunciation Proficiency for Japanese Words," in *Proc. Interspeech*, pp.2233-2236, 2005.

[11] S. Nakagawa and K. Ohta, "A Statistical Method of Evaluating Pronunciation Proficiency for Presentation in English," in *Proc. Interspeech*, pp.2317-2320, 2007.

[12] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji and S. Makino, "Development of English Speech Database Read by Japanese to Support CALL Research", in *Proc. ICA*, Vol. 1, pp.557-560, 2004.

[13] He,X., Zhao, Y., "Model complexity optimization for nonnative English speakers.", in *Proc. EuroSpeech*, pp.1461-1463, 2001.

[14] Ronen, O., Neumeyer, L.,Frando, H. "Automatic detection of mispronunciation for Language Instruction.", in *Proc. EuroSpeech*, Vol. 2, pp.649-652, 1997.

[15] Witt, S., Young, S. "Offline acoustic modeling of nonnative accents.", in *Proc. EuroSpeech*, Vol. 3, pp.1367-1370, 1999.

[16] Hongyan Li. "High Performance Automatic Mispronunciation Detection Method Based on Neural Network and TRAP Features", in *Proc. Interspeech*, pp.1911-1914, 2009.

[17] Su-Youn Y., Mark J., Richard S. "Automated Pronunciation Scoring using Confidence Scoring and Landmark-based SVM", in *Proc. Interspeech*, pp.1903-1906, 2009.