

A STATISTICAL METHOD OF EVALUATING PRONUNCIATION PROFICIENCY FOR ENGLISH WORDS SPOKEN BY JAPANESE

Seiichi Nakagawa, Kazumasa Mori, Naoki Nakamura

Department of Information and Computer Sciences
Toyohashi University of Technology, Toyohashi

nakagawa@slp.ics.tut.ac.jp

Abstract

In this paper, we propose a statistical method of evaluating the pronunciation proficiency of English words spoken by Japanese. We analyze statistically the utterances to find a combination that has a high correlation between an English teacher's score and some acoustic features. We found that the likelihood ratio of English phoneme acoustic models to phoneme acoustic models adapted by Japanese was the best measure of pronunciation proficiency. The combination of the likelihood for American native models, likelihood for English models adapted by Japanese, the best likelihood for arbitrary sequences of acoustic models, phoneme recognition rate and the rate of speech are highly related to the English teacher's score. We obtained the correlation coefficient of 0.81 with open data for vocabulary and 0.69 with open data for speaker at the five words set level, respectively. The coefficient was higher than the correlation between humans' scores, 0.65.

1. INTRODUCTION

As internationalization progresses, the ability to communicate in English is becoming increasingly important. Although individual lessons are a good for language learning, it is difficult to teach English individually at public schools, etc.

Many efforts have therefore been made recently to apply speech technologies to language learning. Many CALL (Computer Assisted Language Learning) systems have been released. Some of these software packages use speech recognition techniques.

We have been investigating a CALL system which focuses on prosody and the effect of Japanese characteristics, and particularly on Japanese manners of generating English word stress[1,2]. In our previous studies, a stressed syllable detector and an accentuation habit estimator were developed, where the estimated habits of individual learners accorded well with their English pronunciation proficiency rated by English teachers.

In this paper, we propose a statistical method of evaluating the pronunciation proficiency for English words spoken by Japanese. Many researchers have studied automatic methods of evaluating pronunciation proficiency. Nuemeyer et al. proposed an automatic text-independent pronunciation scoring method. They used HMM log-likelihood score, segment classification error scores, segment duration scores, and syllabic timing scores for the French language[3]. The evaluation by segment duration was better than others. Furthermore, Franco et al. investigated the evaluation measure based on HMM-

based phone log-posterior probability score and combination of the above scores[4]. We also investigated the posterior probability as the evaluation measure[5]. Furthermore, they proposed the log-likelihood ratio score of native acoustic models to non-native acoustic models and found that this measure outperformed the posterior probability mentioned above[6].

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

The above studies were evaluated for European languages or English uttered by European non-native speakers. We also evaluated English uttered by Japanese.

We also compared acoustic measures of log-likelihood (native acoustic models and non-native acoustic models), likelihood ratio, phoneme recognition rate, rate of speech and best likelihood for arbitrary phoneme sequences and combined these measures by a linear regression model. The result showed that the best was the combination of the above five measures.

2. EXPERIMENTAL SETUP

We used the set of W5 of the English speech database read by Japanese learners[8] for evaluation test data. This set consists of 15 English words spoken by 14 Japanese male student speakers who have better, standard or worse pronunciation proficiency. We used the TIMIT/WSJ database for training native phoneme HMMs and another Japanese speech database for adapting them (non-native acoustic models)[9].

Table 1 shows a summary of the speech materials. The speech is downsampled to 12kHz, and preemphasized then a Hamming window with a width of 21.3 msec is applied every 8 ms. 14 dimensional LPC cepstrum coefficients are used as speech feature parameters for a frame. The acoustic features are 10 LPC based Cepstrum coefficients, Δ and $\Delta\Delta$ features. Acoustic models based on monophone HMMs were learned by the analyzed speech. The HMMs are composed of two to four states, each of which has four mixture Gaussian distributions with full covariance matrices.

Table 1: Speech materials for HMM’s training

Speaker	# speakers	# total sentences
Native (TIMIT)	326	3260
(WSJ)	50	6178
Japanese	76	1065

3. PRONUNCIATION EVALUATION BY ENGLISH TEACHERS

We divided the set W5 into three groups, that is, every group consists of five words. Such a five word group was assessed by four English teachers, two of them (C and D) were American native speakers and the others (A and B) were Japanese English teachers. They ranked every group on a scale ranging from 1 (poor) to 5 (excellent). The criterion of evaluation was focused on segmental proficiency, but not supra-segmental proficiency like accent.

Table 2 summarizes the correlation coefficients between human raters at the 5 words group level and at the 15 words level. The average correlation was 0.65 at the 5 words group level and 0.84 at the 15 words level. Note that the correlation between ratings by two Japanese English teachers is 0.52 and the correlation between ratings by two native speakers is 0.73. This difference of coefficients is relatively large. The average correlation between native speaker and Japanese English teacher is 0.66, which is approximately equal to the average correlation between raters.

Our purpose is to evaluate the pronunciation proficiency at every word. The evaluation for every word is more difficult than that for every five words. Therefore we can say from the table that the target of our automatic evaluation for the correlation between the human score and an automatically evaluated score is about 0.66.

Table 2: Correlation between human raters

Raters	5 words	15 words level
A, B	0.516	0.700
A, C	0.739	0.870
A, D	0.670	0.946
B, C	0.636	0.819
B, D	0.602	0.784
C, D	0.730	0.922
Average	0.652	0.840

4. CORRELATION BETWEEN ACOUSTIC FEATURE MEASURE AND ENGLISH TEACHER’S RATING SCORE

As described in Section 3, English teachers rated the word pronunciation proficiency for a set of five words, whereas acoustic features were measured every one word. Therefore we assumed the human’s rating for the set of five words as the same score for every word in the set for convenience’s sake.

4.1. Log-likelihood

We calculated the correlation rate between English teacher’s score and the log-likelihood (LL) for a pronunciation dictionary based on concatenation of phone HMMs at the word level. The likelihood was normalized by the length in frames. The average correlation coefficient at the 5 words set level was 0.30 for native acoustic HMMs (LL_{native}) and -0.11 for non-native acoustic HMMs adapted by Japanese utterances ($LL_{non-native}$). Since the range of log-likelihood varies from speaker to speakers, it is not useful for the evaluation of pronunciation proficiency.

4.2. Likelihood ratio

Next, we used the likelihood ratio (LR) between native HMMs and non-native HMMs, which were defined as the difference between the two log-likelihoods, that is, $LL_{native} - LL_{non-native}$. The average correlation at the 5 words set level was 0.50, hence the likelihood ratio is useful for the evaluation.

4.3. Best log-likelihood for arbitrary phoneme sequences

The best log-likelihood for arbitrary phoneme sequences (LL_{best}) is defined as the likelihood of free phoneme recognition without using phonotactic language models. We used native phoneme HMMs with four Gaussian mixture distributions having full covariance matrices per state. The average correlation at the 5 words set level was 0.35.

4.4. *a posteriori* probability

We used the likelihood ratio (LR') between the log-likelihood of native HMMs (LL_{native}) and the best log-likelihood for arbitrary phoneme sequences (LL_{best}), which means *a posteriori* probability, that is, $LL_{native} - LL_{best}$ [9]. The average correlation at the 5 words set level was 0.24.

4.5. Phoneme recognition result

We used the results of free phoneme recognition. The average correlations at the 5 words set level of substitution rate, insertion rate, deletion rate, correct rate and accuracy rate were -0.14, -0.09, -0.35, 0.67 and 0.65, respectively. The correct rate ($Cor.$), which is defined as $1.0 - \text{substitution rate} - \text{deletion rate}$, was the most useful for the evaluation among them and the next most useful was the accuracy rate. However, these measures are unreliable for the word level.

4.6. Rate of speech

We defined the rate of speech (ROS) as the ratio of the number of phonemes in a spoken word to the duration (length in frames). The average correlation at the 5 words set level was 0.40. The speech rate is thus very useful for the evaluation at the 5 words set (or sentence) level[7], and it is also useful at the word level.

5. STATISTICAL METHOD OF EVALUATING PRONUNCIATION PROFICIENCY

A linear regression model that is derived from the relationship among acoustic measures and the score of English teachers is proposed for estimating the evaluation score of pronunciation proficiency. We establish some independent variables $\{x_i\}$ for the parameters and the value Y for English teacher's score, and define the linear regression model as

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

,where ε is a residue. The coefficients $\{\alpha_i\}$ are determined by minimizing the square of ε . We experiment with closed data, open data for speakers and open data for vocabulary, respectively. Next, we investigated whether our proposed method is independent of speaker and vocabulary or not. For the open experiment on speakers, we estimated the regression model by using utterances of 13 speakers and estimated the score of the remaining speaker. We repeated this experiment for every speaker.

For the open experiment on the vocabulary, we estimated the regression model using utterances of only 10 words \times 14 speakers and estimated the remaining five words. This procedure was repeated for every 5 words set.

Table 3 summarizes the results for closed data and open data at the 5 words set level. As shown in Table 4, the correlation between an estimated score and native rater's score is larger than that between an estimated score and Japanese English teacher's score. We also find that the correlation between an estimated score and each rater is smaller than that between an estimated score and that of average score by four raters (all). These differences in correlation are due to the difference in the amount of evaluation data.

We estimated the linear model in the case of combining the log-likelihood for native HMMs (LL_{native}), the likelihood for non-native HMMs ($LL_{non-native}$), the rate of speech (ROS), the best likelihood for arbitrary phoneme sequences (LL_{best}) and the correct rate of recognition results ($Cor.$), and obtained the following model:

$$Y = 3.22 + 0.38 \times LL_{native} - 0.20 \times LL_{non-native} + 0.23 \times LL_{best} + 0.29 \times Cor. + 0.54 \times ROS. \quad (2)$$

By using this model, we obtained the correlation of 0.806 on open vocabulary and 0.690 on open speaker for a set of five words as shown in Table 4(b). These coefficients were higher than that between human raters as shown in Table 3. This shows that an automatic evaluation method is superior to the evaluation by Japanese English teachers. Figure 3 illustrates the distribution of estimated scores and native teacher's scores.

6. Conclusion

We proposed a statistical method of evaluating English pronunciation proficiency, which was based on a linear regression model. Although we also investigated a non-linear regression model with a logistic function, there was no difference between

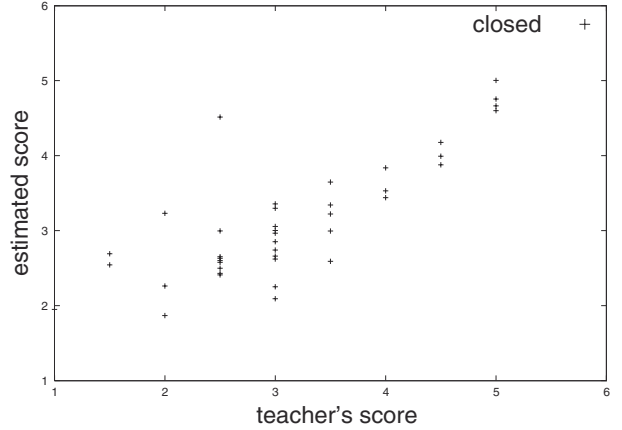


Figure 1: Distribution of estimated score and native teacher's score in closed data for a set of 5 words (correlation=0.80)

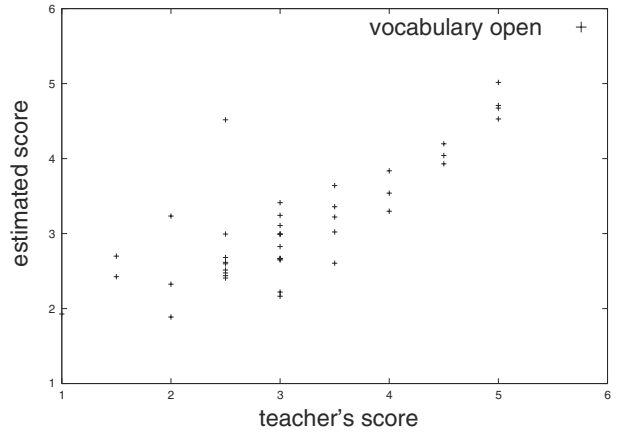


Figure 2: Distribution of estimated score and native teacher's score in open data on vocabulary for a set of 5 words (correlation=0.81)

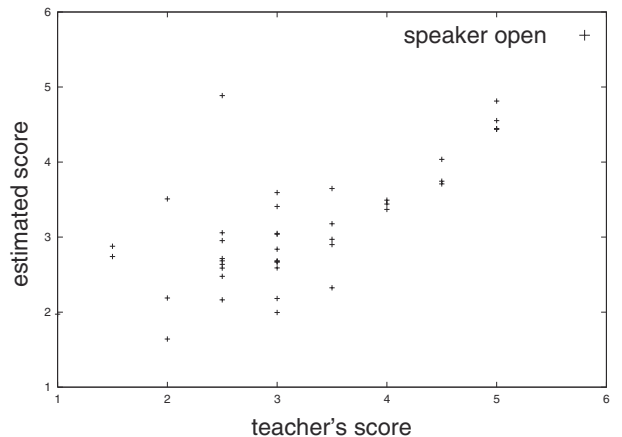


Figure 3: Distribution of estimated score and native teacher's score in open data on speaker for a set of 5 words (correlation=0.69)

Table 3: Correlation between combination of acoustic measures and human rater’s score at the word set level

(parenthesis : the standard deviation of estimation errors)

— at the five words set level —

Acoustic measures	CLOSED	VOC. OPEN	SPK. OPEN
$LL_{native}, LL_{non-native}$	0.529 (0.787)	0.530 (0.787)	0.309 (0.925)
LL_{native}, LL_{best}	0.467 (0.830)	0.462 (0.833)	0.143 (1.426)
$LL_{native}, LL_{non-native}, LL_{best}$	0.645 (0.717)	0.639 (0.723)	0.454 (0.856)
$LL_{native}, LL_{non-native}, ROS$	0.676 (0.684)	0.674 (0.686)	0.514 (0.811)
$LL_{native}, LL_{non-native}, LL_{best}, ROS$	0.712 (0.659)	0.713 (0.659)	0.527 (0.820)
$LL_{native}, LL_{non-native}, LL_{best}, Cor., ROS$	0.785 (0.582)	0.789 (0.578)	0.672 (0.701)
$LL_{native}, LL_{non-native}, LL_{best}, Acc., ROS$	0.725 (0.651)	0.730 (0.647)	0.521 (0.838)

Table 4: Correlation between combination of acoustic measures and human rater’s score with the best result (parenthesis: the standard deviation of estimation errors)

(a) at every word level

Data	Rater		
	Japanese	native	all
CLOSED	0.598 (0.777)	0.647 (0.757)	0.642 (0.722)
VOC. OPEN	0.534 (0.822)	0.574 (0.816)	0.584 (0.766)
SPK. OPEN	0.487 (0.848)	0.530 (0.845)	0.536 (0.797)

(b) at five words level

Data	Rater		
	Japanese	native	all
CLOSED	0.733 (0.665)	0.803 (0.589)	0.785 (0.582)
VOC. OPEN	0.733 (0.665)	0.806 (0.585)	0.789 (0.577)
SPK. OPEN	0.614 (0.777)	0.690 (0.723)	0.672 (0.701)

(c) at 15 words level

Data	Rater		
	Japanese	native	all
CLOSED	0.956 (0.258)	0.953 (0.269)	0.956 (0.258)
VOC. OPEN	-	-	-
SPK. OPEN	0.853 (0.484)	0.863 (0.475)	0.853 (0.484)

the two models. We found the best combination measures for the automatic evaluation was in the following:

- Log-likelihood of a spoken word for native acoustic phoneme HMMs.
- Log-likelihood of a spoken word for non-native acoustic phoneme HMMs adapted by Japanese speakers.
- Best log-likelihood of arbitrary phoneme sequences for native acoustic phoneme HMMs.
- Rate of speech, which is defined as the ratio of the number of phonemes in a spoken word to the length in frames.
- Phoneme recognition rate (correct rate).

By combining these measures, we could evaluate the pronunciation proficiency with almost the same ability as English

teachers, and it was better than the evaluation by Japanese English teachers. If we use one or two sentences for the pronunciation test in practical use, we can evaluate it with high reliability.

7. 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. Cucchiari, 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] N. Minematsu et. al., “Development of English speech database spoken by Japanese learners,” in *Reprint of the COCODA Workshop*, pp.76-81, 2001.
- [9] 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)