An Online Evaluation System for English Pronunciation Intelligibility for Japanese English Learners

Kibishi Hiroshi* and Nakagawa Seiichi*
* Department of Computer Science and Engineering Toyohashi University of Technology, Japan
E-mail: {kibishi, nakagawa}@slp.cs.tut.ac.jp

***Abstract*—We have previously proposed a statistical method for estimating pronunciation proficiency and intelligibility of presentations delivered in English by Japanese speakers. In an offline test, we also evaluated possibly-confused pairs of phonemes that are often mispronounced by Japanese native speakers [7][8]. In this study, we developed an online evaluation system for English spoken by Japanese speakers using offline techniques and carried out an evaluation to obtain the effect thereof based on experimental results. The results showed that both the objective and subjective evaluations improved when using this system.**

We have previously proposed a statistical method for estimating pronunciation proficiency and intelligibility of presentations delivered in English by Japanese speakers. In an offline test, we also evaluated possibly-confused pairs of phonemes that are often mispronounced by Japanese native speakers [7][8]. In this study, we developed an online evaluation system for English spoken by Japanese speakers using offline techniques and carried out an evaluation to obtain the effect thereof based on experimental results. The results showed that both the objective and subjective evaluations improved when using this system.

I. Introduction

Many researchers have investigated automatic methods for evaluating pronunciation proficiency. For example, Neumeyer et al. proposed an automatic text-independent pronunciation scoring method. They used an HMM (hidden Markov model) log-likelihood score, segment classification error scores, segment duration scores, and syllabic timing scores for the French language [1]. They found that evaluation using segment duration showed better results than the other metrics. Franco et al. investigated an evaluation method based on an HMM phone log-posterior probability score and a combination of the scores [2]. We also previously investigated the use of posterior probability as an evaluation metric [3]. Furthermore, Franco et al. proposed using the log-likelihood ratio of native to non-native acoustic models and found that this measure outperformed posterior probability evaluation [4].

Cucchiarini et al. compared the acoustic scores as measured by *TD* (total duration of speech including 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 highly with human ratings than *LR* [5]. All the studies mentioned above focused on either European languages or English uttered by European non-native speakers. Li et al. utilized the weighted combination of three methods, that is, n eural networks (NNs) and multilayer perceptron NNs using TRAP, support vector machines (SVMs) and generalized Markov models, to automatically detect pronunciation mistakes [6]. We also evaluated phoneme pronunciation using SVMs [7]. Furthermore, in our earlier work, we proposed a statistical method to evaluate the pronunciation proficiency of presentations given in English by Japanese speakers [7][8].

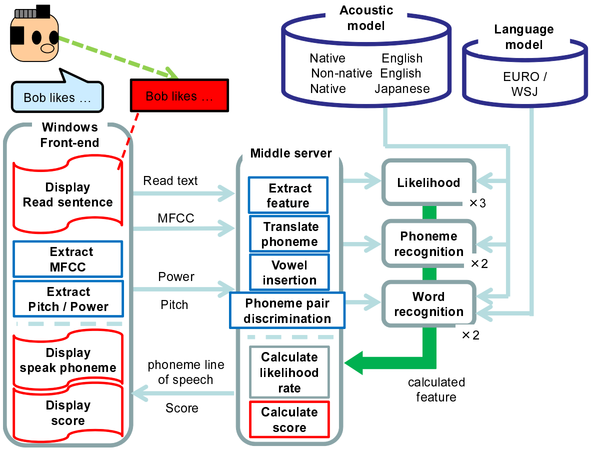


Fig. 1: Configuration of the system.

We built on this previous research by proposing a statistical method for estimating the pronunciation score and intelligibility of presentations given in English by Japanese speakers. Because automatic transcription rates in phoneme and word recognition are not directly related to intelligibility, we investigated the relationship between pronunciation score / intelligibility and various acoustic measures, and then combined these measures using a linear regression model. These studies were conducted offline.

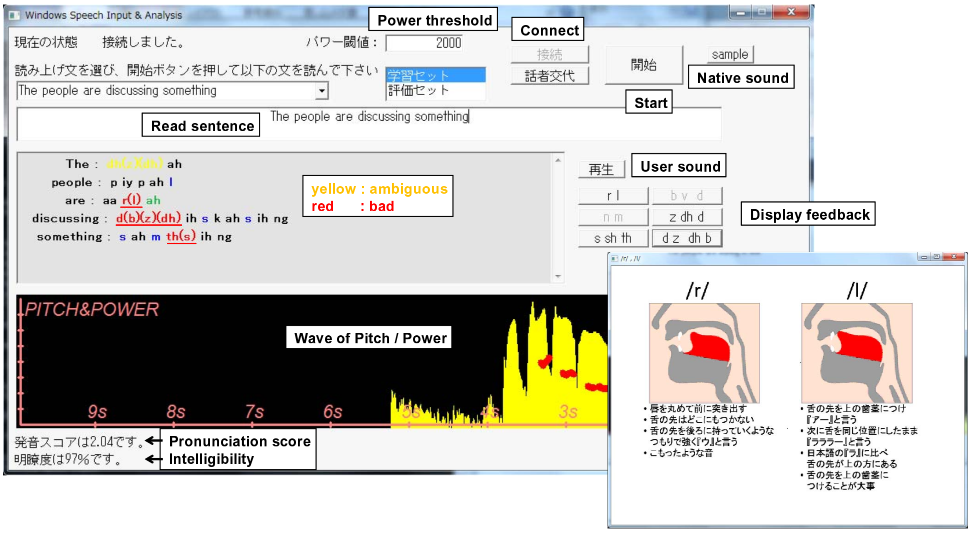
In this study, we developed an online evaluation system for English by Japanese speakers using these techniques. The results of our experiments show that in both the objective and subjective evaluations as well as the English pronunciation scores of the subjects improved when using this system.

II. System overview

An overview of the developed online evaluation system for English is illustrated in Fig. 1. The system is divided into three parts; Front-end, Middle server and Calculation servers.

A. *Front-end*

Based on the configuration in Fig. 1, a running example of the system is shown in Fig. 2. We describe below the input



現在の状態　接続しました。

パワー閾値：2000

読み上げ文を選び、開始ボタンを押して以下の文を読んで下さい

The people are discussing something

The people are discussing something

The : dh(z)(dh) ah
people : p iy p ah l
are : aa r(l) ah
discussing : d(b)(z)(dh) ih s k ah s ih ng
something : s ah m th(s) ih ng

yellow : ambiguous
red : bad

発音スコアは2.04です。
明瞭度は97%です。

- 唇を丸めて前に突き出す
- 舌の先はどこにもつかない
- 舌の先を後ろに持っていくようなつもりで強く『ウ』と言う
- こもったような音

- 舌の先を上の歯茎につけ『アー』と言う
- 次に舌を同じ位置にしたまま『ラララー』と言う
- 日本語の『ラ』に比べ舌の先が上の方にある
- 舌の先を上の歯茎につけることが大事

Fig. 2: Illustrations of the execution of the system.

and output flow in the front-end: (1) Enter the threshold of *power* of the voice and start the process. (2) Push the “Connect Button” and then communicate with the middle server. (3) Select the sentence to read. (4) Push the “Start Button” and then start reading the sentence. (5) Display pronunciation score, intelligibility, and results of the evaluated pronunciation units, i.e., phonemes. (6) Repeat (1) to (5) for the next utterance.

1) Evaluation indicators: In this system, users can obtain contours of their power and pitch, the results of the evaluated pronunciation phoneme units, and pronunciation and intelligibility scores as evaluation indicators.

1) Contours of power and pitch: In general, a stress accent is used by native English speakers. Therefore, we speculated that the sound of a good English speaker should include a good power accent. Thus, the system displays the power contour to judge whether the English utterance is good. The system also displays the contour of the pitch from the fundamental frequency.
2) Phoneme pair discrimination[7][8]: According to the determination results of the SVM for confusable phoneme pairs, the system displays which phoneme pronunciations are incorrect for users. At the same time, if there is any incorrect pronunciation, the button that corresponds to the phoneme pair is activated and the system presents the feedback of articulation shown in Fig. 2 when the button is pressed.
3) Pronunciation score and intelligibility: As indicators of the evaluation of a user utterance, we use pronunciation score and intelligibility.

   The pronunciation score used in this paper is the average of two scores: a phonetic pronunciation score and a prosody (rhythm, accent, intonation) score, which are estimated by statistical analysis based on the data rated by five American English teachers for each of the 289 sentences in the Translanguage English Database (TED).

   Intelligibility in this paper is defined as how well English teachers interpret (recognize) the pronunciation of non-native speakers. Four out of the five American English teachers transcribed each sentence, while scoring the speaker’s pronunciation proficiency. The four transcriptions of the same sentence were compared, and any word transcribed by two or more English teachers as being the same word was referred to as **man2/4**. After computing all the man2/4 values for all utterances, the intelligibility was calculated as:

$$Intelligibility = A/B, \qquad (1)$$

   where A represents the number of words transcribed as man2/4 in each sentence, that is, how many words a teacher recognized / interpreted correctly, and B represents the total number of words in each sentence. However, because we did not have correct transcriptions of the test data from the speakers themselves, we were not able to obtain the exact number of words in the sample sentences. Consequently, we assumed the total number of words in a sentence to be the sum of the number of words transcribed as man2/4 in the sentence combined with the average number of transcribed words not included in the man2/4 figures from the same sentence.

B. Middle server

To estimate the pronunciation score and intelligibility, we developed a linear regression model, derived from the relationship between the observed acoustic measures and the scores of the English teachers. We used the likelihood score, likelihood ratio, phoneme and word recognition rates, phoneme pair discrimination rate, and vowel insertion rate as acoustic features. Let the independent variables $\{x_i\}$ denote the parameters and the value Y denote the English teacher scores, then the linear regression model is defined as:

$$Y = \Sigma_i \alpha_i \times x_i + \varepsilon, \qquad (2)$$

where ε is the residue [7][8]. The coefficients $\{\alpha_i\}$ are determined by minimizing the square of ε. We conducted

TABLE I: Contents of the read sentences

phase	No.	contents	
train	100	Tactics for TOEIC (Included with a native speaker's voice)	
test	20	ERJ	in mind Phoneme learning (10) in mind Intonation(5) in mind Accent and Rhythm(5)

experiments with open data for speakers and investigated whether our proposed method can function independently of the speaker. For the open experiment using these speakers, we calculated a regression model using the utterances of 20 of the speakers and estimated the score of the remaining speaker.

Using an SVM, the system discriminated nine pairs of phonemes that are often mispronounced by native Japanese speakers and get the phoneme pair discrimination rate. These pairs are / l and r /, / m and n /, / s and sh /, / s and th /, / b and v /, / b and d /, / z and dh /, / z and d /, and / dh and d /.

The SVM input data comprised fixed length frames, that is, five consecutive frames beginning from the -2nd frame of the central frame of the phoneme segment. The features are MFCC (12) and Δ MFCC (12).

C. Calculation servers

1) Acoustic and language models: We used the TIMIT / WSJ (Wall Street Journal) database for training the native English phoneme HMMs (native English model), the English speech database read by Japanese speakers (ERJ) [14] for adapting them (non-native English model), and the ASJ / JNAS database for training native Japanese syllable HMMs (strictly speaking, the mora-unit Japanese model). Features used are MFCC (12), ΔMFCC (12), $\Delta\Delta$MFCC (12), ΔPower and $\Delta\Delta$Power; 38 dimensions in total.

We used the correct rate of word recognition obtained from the SPOJUS large-vocabulary continuous speech recognition (LVCSR) system [11] with a language model. We used the WSJ database and Eurospeech'93 papers for training bigram language models.

2) Calculation: Three types of calculation servers are used as described below:

1) Likelihood: We used three HMMs (native, non-native, Japanese) to calculate the likelihood, which was normalized by the length in frames.
2) Phoneme recognition: We used the correct rate, substitution rate, and deletion rate obtained from an arbitrary phoneme recognition system.
3) Word recognition: We used the correct rate of word recognition with a language model. We used the correct rate, substitution rate, and deletion rate obtained from the LVCSR of arbitrary words.

III. Experiment

A. Conditions

To examine whether using the proposed systems has any effect on learning English pronunciation, we carried out an experiment. Eight male Japanese students were used as subjects. The learning period lasted for about 3 weeks, 20 minutes per day, five times per week; a total of 15 learning intervals. To evaluate the effectiveness of the system, the evaluation test data were recorded before the experiment (pre), after 10 learning intervals (mid), and again after 15 learning intervals (post).

TABLE II: Correlation Coefficient (CC) between one native scorer and average of other pronunciation scorers

(a) pronunciation

Scorer	CC
1 and other	0.664
2 and other	0.670
3 and other	0.595
4 and other	0.475
5 and other	0.487
6 and other	0.336
average	0.540

(b) intelligibility

Scorer	CC
1 and other	0.845
2 and other	0.873
3 and other	0.731
4 and other	0.772
5 and other	0.808
6 and other	0.752
average	0.800

For the learning process, a training set of 100 statements was prepared from Tactics for TOEIC (Test of English for International Communication) [16]. The statements in this learning set were spoken by a native speaker, so that the system could present a native voice to users. In addition, we taught the subjects the basic operation of the system, so that at the time of learning, they could use the system freely.

The test was performed three times with different sets, each consisting of 20 sentences, from the training set. We used the same sentences for all three test iterations. The 20 statements were selected from the ERJ [14] taking the following three aspects into consideration: pronunciation of phonemes, intonation, and rhythm. Table I gives the details of the read sentence set. The experimental results are discussed in the next section. For each of the 480 (8 speakers $\times$ 3 times $\times$ 20 sentences) test sentences, we obtained scores focusing on phoneme pronunciation, fluency, and prosody by six native speakers. Table II (a) shows that the correlation between the score for one native speaker and the average score of the others is moderate. Regarding intelligibility, the six native speakers were requested to transcribe 48 sentences (8 speakers $\times$ 3 times $\times$ 2 sentences) selected independently from the training set, to be transcribed. Users uttered *a priori* prepared sentences. Let the number of words in the read sentence be A, and the number of words correctly transcribed be B, then intelligibility is calculated as follows: $\frac{A}{B} \times 100$. Table II (b) shows that the correlation is high.

B. Results

1) Pronunciation score / intelligibility: Table III gives the experimental results. Here, “” denotes that the post-test achieves the best score of the three tests, “” that the post-test is relatively better than the pre- & mid-tests, “” that the pre-test is relatively better than the post- & mid-tests, and “ ” that the pre-test achieves the best score of the three tests. According to the results, with the exception of E, the post-test results of the others were better than the pre-test.

For the pronunciation score, Table IV shows the correlation coefficient for the estimated score and the native speaker's score. From Tables II and IV, since the system could estimate the pronunciation score / intelligibility with a correlation of 0.492 / 0.747 between the estimated score and the native average score, we can state that the estimation is adequate, because the correlation between native scores is 0.540 / 0.800, respectively. And for intelligibility, a correlation of

TABLE III: Scores of system (estimated score) and native scorers for every test

			A	B	C	D	E	F	G	H	average	improve rate [%]
system	pronunciation score	pre	2.60	2.76	3.22	2.62	2.89	2.43	2.70	2.67	2.7	-
		mid	2.78	2.57	3.23	3.23	3.36	2.24	2.95	2.78	2.9	6.9
		post	3.50	3.07	3.52	3.11	3.37	2.58	3.23	2.34	3.1	15.6
	intelligibility	pre	87.4	58.6	79.5	90.6	47.5	74.1	93.5	63.1	74.3	-
		mid	85.2	79.8	89.2	97.7	74.1	71.3	82.3	100	84.9	41.5
		post	83.9	100	87.9	52.9	69.7	84.7	79.0	77.1	79.4	19.9
estimated by scorer	pronunciation score	pre	2.67	2.79	2.46	2.72	2.70	2.45	2.90	3.12	2.7	-
		mid	2.77	2.86	2.69	2.98	2.67	2.63	3.01	2.80	2.8	3.3
		post	2.96	2.96	2.63	2.77	2.36	2.60	3.10	2.38	2.7	-0.3
	intelligibility	pre	82.5	78.6	80.6	96.3	37.4	75.0	81.7	71.7	75.5	-
		mid	90.7	81.9	84.7	93.2	52.7	83.3	86.9	97.6	83.9	34.3
		post	88.1	95.9	85.6	77.2	74.5	83.3	79.5	69.2	81.7	25.3

TABLE IV: Correlation Coefficient (CC) between estimated score by system and native score

(a) pronunciation

time	CC
pre	0.539
middle	0.510
post	0.476
all	0.492

(b) intelligibility

time	CC
pre	0.808
middle	0.688
post	0.712
all	0.747

0.747 between the estimated score and the native average score. The scores for improved rate of pronunciation and intelligibility of the system are 6.9 $\sim$ 15.6 % and 19.9 $\sim$ 41.5 %, respectively. The scores for improved rate of pronunciation and intelligibility evaluated by the native speakers are -0.3 $\sim$ 3.3 % and 25.3 $\sim$ 34.3 %, respectively. Five to seven out of eight learners improved the pronunciation and intelligibility for the objective and subjective evaluations.

2) Questionnaire: We carried out a survey based on a questionnaire at the end of the experiment. From the answers, we found that the pronunciation score, listening to a native speaker's voice, indication of mispronounced phonemes, and listening to the user utterances facilitated learning and at the same time provided the motivation to practice. Regarding the intelligibility score, the subjects did not know how to use the intelligibility score that measures how much natives understand correctly. The contours of power and pitch were not used during pronunciation training because users did not know how to practice using these contours.

In addition, from the responses to a question comparing this system with self-practice, it is clear that the estimated score and display of mispronounced phonemes provide the motivation to continue practicing. From the answers concerning self-directed learning, we ascertained that information about the quality of the subject's own pronunciation and which phoneme could be improved is very important.

IV. Conclusions

In this paper, we developed an online evaluation system for English pronunciation targeted at Japanese speakers. Through experiments using this system to practice English pronunciation, we confirmed the positive learning effect thereof. From the questionnaire, we ascertained that the pronunciation score, listening to a native voice, indication of mispronounced phonemes, and listening to the user's own utterance provided the motivation to practice.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] K. Hirabayashi,and S. Nakagawa, "Automatic Evaluation of English Pronunciation by Japanese Speakers Using Various Acoustic Features and Pattern Recognition Techniques," in *Proc. Euro Speech*, pp.598-601, 2010.

[8] H. Kibishi, and S. Nakagawa, "New feature parameters for pronunciation evaluation in English presentations at international conferences.", in *Proc. Euro Speech*, pp.1149-1152, 2011.

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

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

[11] Y. Fujii, K. Yamamoto, and S. Nakagawa, "Large vocabulary Speech Recognition System: SPOJUS++.", in *Proc. MUSP*, pp.110-118, 2011.

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

[13] Acoustical Society of America "SII: Speech Intelligibility Index", http://www.sii.to/index.html

[14] Advanced Utilization of Multimedia to Promote Higher Education Reform "English Speech Database Read by Japanese Students", http://research.nii.ac.jp/src/eng/list/detail.html

[15] "TED Translanguage English Database", http://www.elda.org/catalogue/en/speech/S0031.html

[16] Grant Trew, "Tactics for TOEIC Listening and Reading Test Student Book", Oxford Univ Pr (Sd), 2008.