ReCALL **27**(1): 58–83. 2015 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344014000251
First published online 23 May 2014

A statistical method of evaluating the pronunciation proficiency/intelligibility of English presentations by Japanese speakers

HIROSHI KIBISHI

Toyohashi University of Technology, Computer Science and Engineering, Japan (email: kibishi@slp.cs.tut.ac.jp)

KUNIAKI HIRABAYASHI

Toyohashi University of Technology, Computer Science and Engineering, Japan (email: kuniaki@slp.cs.tut.ac.jp)

SEIICHI NAKAGAWA

Toyohashi University of Technology, Computer Science and Engineering, Japan (email: nakagawa@slp.cs.tut.ac.jp)

Abstract

In this paper, we propose a statistical evaluation method of pronunciation proficiency and intelligibility for presentations made in English by native Japanese speakers. We statistically analyzed the actual utterances of speakers to find combinations of acoustic and linguistic features with high correlation between the scores estimated by the system and native English teachers. Our results showed that the best combination of acoustic features produced correlation coefficients of 0.929 and 0.753 for pronunciation and intelligibility scores, respectively, using open data for speakers at the 10-sentence level. In an offline test, we evaluated possibly-confusing pairs of phonemes that are often mispronounced by Japanese speakers of English. In addition, we developed an online real-time score estimation system for Japanese learners of English using offline techniques to evaluate the pronunciation and intelligibility scores in real-time with almost the same ability as English teachers. Finally, we show that both the objective and subjective evaluations improved after learning with our system.

Keywords: English learning, pronunciation evaluation, intelligibility evaluation, offline/online execution, Japanese speakers

1. Introduction

As internationalization continues, the ability to communicate in English is becoming increasingly important. Although private lessons are beneficial for language learning, such teaching of English is difficult at all schools, because of the cost. Recently many efforts have

applied speech technologies to language learning. For instance, many Computer Assisted Language Learning (CALL) systems or Computer Assisted Pronunciation Training (CAPT) systems have been released (Kawahara & Minematsu, 2011), some of which use speech recognition techniques (Nakagawa, Reyes, Suzuki & Taniguchi, 1997; Tsubota, Kawahara & Dantsuji, 2002; Eskenazi, Kennedy, Ketchum, Olszewski & Pelton, 2007). CAPT is a crucial component of CALL that focuses on evaluating pronunciation proficiency or correcting pronunciation errors.

The authors have developed a stressed syllable detector and an accentuation-habit estimator, where the estimated habits of individual learners accorded well with their English pronunciation proficiency and intelligibility rated by English teachers (Fujisawa, Minematsu & Nakagawa, 1998; Nakamura, Nakagawa & Mori, 2004). In this paper, we propose extended pronunciation proficiency/intelligibility estimation methods using an online system developed by the authors. This enabled us to evaluate the learning effect in pronunciation proficiency and showed improvement in the intelligibility of learners' utterances.

1.1. Three approaches to pronunciation assistance

Computer assisted pronunciation training (CAPT) has a compelling motivational effect (Stenson, Downing, J. Smith & K. Smith, 1992). Aist (1999) classified pronunciation assistance into three general approaches. The first approach is to use a program that analyzes a learner's utterance to extract acoustic features such as intonation (or pitch contour), loudness and spectrogram and then displays these features visually along with the teacher's (or reference's) features (visual feedback approach). The second approach is to compare a learner's utterance with a template or reference recorded by a native speaker and then to automatically score the pronunciation (template based approach). The third approach is to evaluate a learner's utterance by using statistical models trained by many native speakers (model-based approach).

Our approach adopts the model-based approach.

1.2. Related research on model-based CAPT

Many researchers have studied automatic methods of evaluating pronunciation proficiency. Neumeyer, Franco, Weintraub and Price (1996) proposed an automatic text-independent pronunciation scoring method that uses Hidden Markov Model (HMM) log-likelihood scores (see Appendix), segment classification error scores, segment duration scores, and syllabic timing scores for French. The evaluation by segment duration outperformed others. Ronen, Neumeyer and Franco (1997), who investigated evaluation measures based on HMM-based phone log-posterior probability scores and the combination of the above scores proposed the log-likelihood ratio scores of native acoustic models to non-native acoustic models and found that this measure outperformed the above posterior probability (Ramos, Franco, Neumeyer & Bratt, 1999). We also investigated posterior probability as an evaluation measure for Japanese (Nakagawa, Reyes, Suzuki & Taniguchi, 1997). Cucchiarini, Strik and Bovels (2000) 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*. Neri, Cucchiarini and Strik (2008)

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

compared three systems: an ASR-based CAPT system with automatic feedback, a CAPT system without feedback, and no CAPT system, and showed the effectiveness of computer-based speech corrective feedback. Wang and Lee (2012) integrated Error Pattern (EP)-based with Goodness-of-Pronunciation (GOP) -based mispronunciation detectors (Witt & Young, 1999) in a serial structure to improve a mispronunciation detection system.

Koniaris and Engwall (2011) described a general method that quantitatively measures the perceptual differences between a group of native speakers and many different kinds of non-native speakers; their system was verified by the theoretical findings in literature obtained from linguistic studies. To evaluate phoneme pronunciation, Yoon, Hasegawa-Johnson and Sproat (2009) utilized a Support Vector Machine (SVM) using Perceptual Linear Predictive (PLP) features and formant information as acoustic feature parameters. To automatically detect mispronounced phonemes, Li, Wang, Liang, Huang and Xu (2009) combined three methods: Neural Network (NN) & MLP-NN using TempoRAl Patterns (TRAP) features, SVM, and Gaussian Mixture Model (GMM). Smit and Kurimo (2011) recognized individual accent utterances using stacked transformations. For the speech recognition of non-native speakers, linear or nonlinear transformations are usually input to HMM-based acoustic models (Karafiat, Janda, Cernocky & Burget, 2012).

The above studies were evaluated for European languages or English uttered by European non-native speakers. Wu, Su and Liu (2012) presented an efficient approach to detecting personalized mispronunciation in Taiwanese-accented English. Holliday, Beckman and Mays (2010), who focused on fricative sounds like *shu* whose pronunciation is difficult for non-native speakers, distinguished between English and Japanese speakers. In contrast, we evaluated the Japanese spoken by foreign students (Ohta & Nakagawa, 2005). For non-Japanese, pronouncing the chocked sound and longer vowels of Japanese is very difficult. On the other hand, for Japanese, pronouncing consecutive consonants and discriminating between similar phonemes in English is very difficult, (e.g., "strike," " l and r" and "b and v"). These difficulties are caused by the differences in phonotactic structure and phoneme system between Japanese and English.

1.3. Our approach

In contrast with the above research studies, this paper focuses on the following points: (a) the target utterance is "presentation/spontaneous speech" at an international conference rather than "read speech" for given sentences; (b) the system estimates both pronunciation and intelligibility scores; (c) we transferred offline techniques to the online system; and (d) we introduced new acoustic/linguistic features to estimate pronunciation and intelligibility scores. We proposed a statistical method for estimating the pronunciation and intelligibility scores of presentations given in English by Japanese speakers (Nakagawa & Ohta, 2007; Hirabayashi & Nakagawa, 2010; Kibishi & Nakagawa, 2011). Then we investigated the relationship between two scores (pronunciation proficiency and intelligibility) rated by native English teachers and various measures used to estimate a score. To the best of our knowledge, the automatic estimation of intelligibility has not yet been studied except for the intelligibility of dysarthric speech (Falk, Chan & Shein, 2012). Furthermore, we developed an online real-time score estimation system, evaluated the system's interface, and showed its effectiveness for learning pronunciation. Finally, we show that certain combinations of acoustic measures can predict pronunciation and intelligibility scores with almost the same ability as English teachers.

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

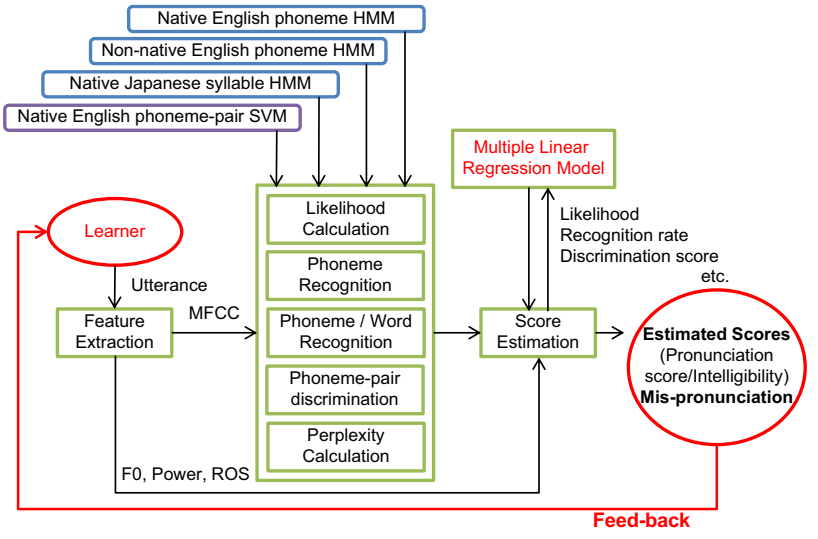

Fig. 1. Block diagram of our estimation system for pronunciation and intelligibility scores

2. System overview

In this paper, we propose a statistical method that evaluates pronunciation proficiency for presentations in English. We calculated acoustic and linguistic measures from presentations given during lectures and combined these measures by a linear regression model to estimate both scores. Figure 1 shows a block diagram of our evaluation system for pronunciation and intelligibility scores.

First, our system extracts the following phonetic/prosodic features from speaker utterances: Mel-Frequency-Cepstrum Coefficient (MFCC, which corresponds to the spectrogram envelope), *F0*, *Power*, and *ROS*. *F0*, *Power*, and *ROS* are directly used as prosodic features in score estimation. Next, using MFCC, it calculates many kinds of acoustic/linguistic measures as clues to estimate scores. For phoneme/word recognition, three types of HMMs are used for various likelihood calculations, and SVM is used in phoneme-pair discrimination. Then these measures are used in score estimation with *F0*, *Power*, and *ROS* and combined with multiple linear regression to estimate the scores. This statistical method is explained in Section 6, and the automatic speech-recognition method using HMM is explained in the Appendix.

3. Database

We used the Translanguage English Database (TED), which was presented at the International Conference on EuroSpeech, for the evaluation test data (Nakagawa & Ohta, 2007; Hirabayashi & Nakagawa, 2010; Kibishi & Nakagawa, 2011). Only part of TED is comprised of texts transcribed by a native speaker (not the speaker himself); the rest contains raw data. This set consists of 289 English sentences in presentations spoken by 21 male speakers, which are rated at three skill levels of pronunciation proficiency: above average, average, or below average. Sixteen of the 21 are Japanese speakers, and the remaining five are native English speakers from the USA.

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

Table 1 *Speech material training data for HMM*

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

Table 2 *Test data*

Speaker	(Database)	#Speakers	#Total sentences
Native	(TED)	5	63
Japanese	(TED)	16	226
Total	(TED)	21	289

Table 3 *English phoneme recognition result using monophone and triphone model trained for native data [%]*

Speaker	Model	Del	Ins	Subs	Cor	Acc
Native	monophone	14.8	2.6	31.8	53.4	50.8
	triphone	8.7	3.0	24.8	66.5	63.6
Japanese	monophone	11.9	3.1	51.5	36.6	33.5
	triphone	4.4	11.5	52.2	43.4	31.9

We used the TIMIT (Garofolo, Lamel, Fisher, Fiscus, Pallett, Dahlgren & Zue, 1993)/WSJ (Garofalo, Graff, Paul & Pallett, 2007) database for training the native English phoneme HMMs, which is another Japanese speech database for adapting non-native English phoneme HMMs (Nakagawa, Reyes, Suzuki, Reyes & Taniguchi, 1997, in Japanese), and the ASJ (Kobayashi, Itahashi, Hayamizu & Takezawa, 1992, in Japanese)/JNAS (Itou, Yamamoto, Takeda, Takezawa, Matsuoka, Kobayashi & Shikano, 1999) database for training the native Japanese syllable HMMs (strictly speaking, mora-unit HMMs).

Tables 1 and 2 summarize the speech materials.

Franco, Neumeyer, Kim and Ronen (1997) found that for the pronunciation evaluation of non-native English speakers, a triphone model performs worse than a monophone model if the HMMs are trained by native speech; less detailed (native) models perform better for non-native speakers (Franco *et al.*, 1997; Young & Witt, 1999; Zhao & He, 2001). We also confirmed this fact. A triphone model improved the performance for native speakers more than a monophone model, but not for Japanese speakers (see Table 3), because of the influence of Japanese phonotactics. Japanese cannot correctly pronounce consecutive syllable sequences, so a context-sensitive tri-phone model affects the recognition of English uttered by Japanese.

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

For example, the accuracy for native speakers using triphone models was 64.4% and 50.2% with the monophone models; for Japanese, the accuracy was 30.8% for the triphone models and 33.5% for monophone models.

4. Definition of estimating scores

In this paper, we defined two kinds of scores, pronunciation score and intelligibility, and calculated/estimated them using an automatic evaluation system.

4.1. Pronunciation score

The pronunciation score used in this paper is the average of two scores: a phonetic pronunciation score and a prosody (rhythm, accent, intonation) score. This score was assessed for each of 289 sentences by five native English teachers who ranked each utterance on a scale of 1 (poor) to 5 (excellent).

4.2. Intelligibility score

Typically, the physical measure that is highly correlated with speech intelligibility is called the Speech Intelligibility Index (SII) (Acoustical Society of America SII). SII is calculated from the acoustical measurements of speech and noise/reverberation. In contrast, the intelligibility that we used in this paper is defined as how well the pronunciation of utterances by non-natives is recognized or perceived by native English teachers.

The test data were assessed by four of the above five native English teachers for all 289 sentences and the intelligibility was calculated. The teachers transcribed each sentence by listening to all test data while scoring each speaker. The transcription by one native speaker was not used because it was unreliable. Four transcriptions from the same sentence by English teachers were compared, and if two or more English teachers transcribed the same word, we determined it to be an uttered word and called it **man2/4**. Once man2/4s for all utterances were extracted, the intelligibility (the correctly transcribed rate) was calculated:

$$Intelligibility = {}^{A}/_{B}, \tag{1}$$

where A represents the number of words in man2/4 and B represents the total number of words in the target sentences. We show two examples below, where the underlined words denote man2/4.

Example of transcription:

Teacher 1: and to work robustly since it's spontaneous input and also obviously because speech recognition is not perfect yet

Teacher 2: and to work robustly since it spontaneous input and also obviously because speech recognition isn't perfect

Teacher 3: and to work robustly since its spontaneous input and also obviously because speech recognition is not perfect yet

Teacher 4: and to work robustly and since its spontaneously input and since speech recognition is not perfect

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

man2/4: and to work robustly since its spontaneous input and also obviously because speech recognition is not perfect yet

$$Intelligibility = \frac{A}{B} = \frac{18}{19.75} = 0.911$$

Teacher 1: and then we uh estimate the @ parameters automatically from the sequence
Teacher 2: and then we estimate the @ parameters automatically from the sequence
Teacher 3: and then we estimate the armor @ automatically from the sequence
Teacher 4: and then we estimate the ARMA parameters automatically from the sequence
man2/4: and then we estimate the ¥ parameters automatically from the sequence

$$Intelligibility = \frac{A}{B} = \frac{10}{11.5} = 0.902$$

Here, the mark "@" denotes a word/phrase by the speaker that the teacher heard without understanding it completely, and "¥" denotes that a word is present. However, because we did not have correct transcriptions of the test data from the speakers themselves, we could not obtain the exact number of words in the sentences. Consequently, we assumed that the total number of words in a sentence is the sum of the words transcribed as man2/4 in the sentence and the average number of transcribed words that are not included in the man2/4 figures from the same sentence: in other words, the average number of transcribed words.

4.3. Scoring by English teachers

All five teachers scored pronunciation, and four of them transcribed and calculated intelligibility. Table 4 shows the set of teachers who scored pronunciation proficiency and/or intelligibility. Tables 5 and 6 summarize the correlation coefficients between each English teacher's pronunciation and intelligibility scores, where "A and the others" means the correlation between the score of A and the average score of {B, C, D, E}.

We can determine from Tables 5 and 6 that the target of our automatic evaluation for the correlation between the human's score and an automatically evaluated score is 0.683 ~ 0.794 and 0.697, respectively, to develop an automatic evaluation system with the same ability as human experts. Figure 2 shows the relationship between intelligibility and pronunciation scores rated by native English teachers, where this correlation was found to be 0.717. This evaluation was performed using a presentation's utterances having a duration of two minutes. Speakers with high pronunciation scores have high intelligibility, i.e., the higher the pronunciation score is, the more the person can correctly comprehend the speaker's utterances.

Table 4 *Set of teachers for pronunciation score and intelligibility*

	Teacher
Pronunciation score	A, B, C, D, E
Intelligibility	A, B, C, D

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

Table 5 *Correlation coefficient of inter-teacher pronunciation scores*

(a) Among teachers		b) Among teachers and others	
Teacher	Correlation	Teacher	Correlation
A, B	0.591	A, the others	0.921
A, C	0.570	B, the others	0.631
A, D	0.986	C, the others	0.614
A, E	0.933	D, the others	0.936
B, C	0.586	E, the others	0.868
B, D	0.588	Average	0.794
B, E	0.519		
C, D	0.597		
C, E	0.520		
D, E	0.945		
Average	0.683		

Table 6 *Correlation of inter-teacher intelligibility scores: based on man2/4*

Teacher	Correlation
A, man2/4	0.750
B, man2/4	0.701
C, man2/4	0.649
D, man2/4	0.689
Average	0.697

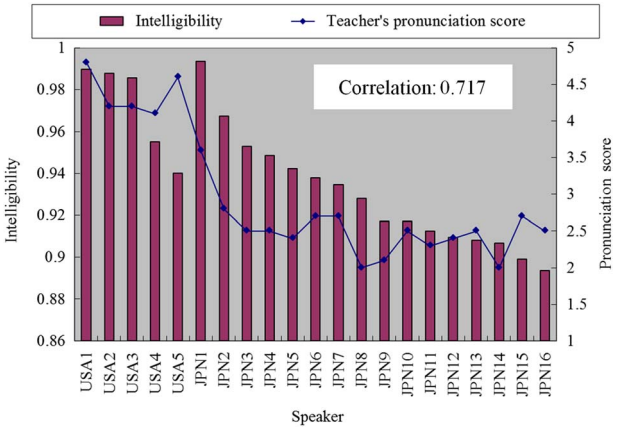

Fig. 2. Relationship of intelligibility and teacher pronunciation scores (for utterances having a duration of about two minutes); USA means native English speakers and JPN means Japanese English speakers

5. Definition of measures and evaluation

As described in Section 4. 3, since English teachers evaluated the pronunciation and intelligibility scores (or transcription) for a set of utterances, we assumed for convenience that

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

the human score for the set of sentences in two minutes was the same score for every sentence in the set. We previously proposed effective acoustic features (Nakamura, Nakagawa & Mori, 2004; Ohta & Nakagawa, 2005). In this paper, we added new features to estimate the pronunciation and intelligibility scores. Although the previous work (Nakagawa & Ohta, 2007) used read-sentence utterances as test sets, this work used presentation (spontaneous) utterances in English. The terminology of log-likelihood and posterior probability using HMM is defined in Appendix A.

5.1. Acoustic measures

(A) Log-likelihood using native and non-native English HMMs and the learner's native language HMM (Nakamura, Nakagawa & Mori, 2004)
We calculated the correlation rate between the averaged English teacher scores and the log-likelihood (LL) for a pronunciation dictionary sequence based on the concatenation of phone HMMs at the 1-sentence level. The likelihood was normalized by length in the frames. We used native English phoneme HMMs (LL_{native}), non-native English phoneme HMMs that are adapted by Japanese utterances ($LL_{non-native}$), and native Japanese syllable HMMs (LL_{mother}).

(B) Best log-likelihood for arbitrary phoneme sequences (Nakamura, Nakagawa & Mori, 2004)
The best log-likelihood for arbitrary phoneme sequences is defined as the likelihood of free phoneme (syllable) recognition without using phonotactic language models. We used native English phoneme HMMs (LL_{best}).

(C) Likelihood ratio (Nakamura, Nakagawa & Mori, 2004)
We used the likelihood ratio (LR) between native English HMMs and non-native English HMMs, which were defined as the difference between the two log-likelihoods: $LR = LL_{native} - LL_{non-native}$.

Figure 3 illustrates the Gaussian distributions for native English HMMs and non-native English HMMs/Japanese HMMs. Note that the likelihood is associated with the inverse of distance. A denotes a sample from a typical native English speaker, B denotes a sample from an outlying native English speaker and C denotes a Japanese utterance sample from a non-native speaker. Even if a native English speaker utters his/her mother language, the likelihood, using native English HMMs, is distributed widely from a high to a low value. Therefore, absolute value LL_{native} is not suitable for outlying speakers. However, the difference in the likelihoods between $LL_{native} - LL_{non-native}$ or $LL_{native} - LL_{mother}$ compensates/normalizes this phenomenon. In Figure 3, LL_{native} for samples B and C is almost similar. On the other hand, $LL_{native} - LL_{non-native}$ for B is larger than that for C, and B is considered a better English utterance sample than C. We assume that this measure has a speaker normalization function as well as a similar effect of Vocal Tract Length Normalization (VTLN); in other words, it is a mother-language-independent English evaluation measure.

(D) Posterior probability (Nakamura, Nakagawa & Mori, 2004)
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}), which means

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

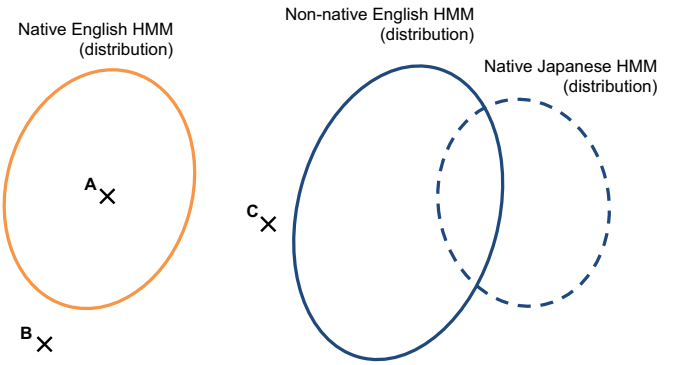

Fig. 3. Illustration of Gaussian distributions corresponding to native HMMs and non-native HMMs/native Japanese HMMs; A: typical native sample; B: outlying native sample; C: non-native Japanese utterance sample

the logarithm of *posterior* probability: $LR' = LL_{native} - LL_{best}$ (Nakagawa, Reyes, Suzuki & Taniguchi, 1997, in Japanese; Nakagawa, Reyes, Suzuki & Taniguchi, 1997).

(E) Likelihood ratio for phoneme recognition (Nakamura, Nakagawa & Mori, 2004)
We used the ratio of the likelihood of free phoneme recognition between native and non-native English HMMs (LR_{adap}), which were defined as the difference between the two log-likelihoods: $LR_{adap} = LL_{best\_native} - LL_{best\_non-native}$.

We also used the ratio of the likelihood of the free phoneme (syllable) recognition between native English and native Japanese HMMs (LL_{mother}), which were defined as the difference between the two log-likelihoods: $LL_{mother} = LL_{best\_native} - LL_{best\_mother}$.

(F) Phoneme recognition result (Nakamura, Nakagawa & Mori, 2004)
We used the results of the free phoneme recognition. The test data were restricted to the correctly transcribed parts based on man2/4, because this measure requires correct transcription of utterances.

(G) Word recognition result
We used the correct rate of word recognition with a language model called Large Vocabulary Conversational Speech Recognition (LVCSR). We used the WSJ database (WSJ) or Eurospeech '93 paper (EURO) for training the bigram language models (Ohta & Nakagawa, 2005). This measure also requires the correct transcription of utterances.

(H) Standard deviation of powers and pitch frequencies
The standard deviations of powers (*Power*) and fundamental (*Pitch*) frequencies (*FO*) are calculated for every utterance.

(I) Rate of speech
We used the rate of speech (*ROS*) of the sentence. Silences in an utterance were removed. We calculated the *ROS* of each sentence as the number of phonemes divided by the duration in seconds.

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

(J) Perplexity and entropy

Perplexity can be used to evaluate the complexity of an utterance. This measure corresponds to the average number of words that can appear in a given left context. We used WSJ and Eurospeech '93 papers (EURO) for training the bigram language models (Nakagawa & Ohta, 2007). Entropy H and perplexity PP can be calculated for word sequence $w_1 w_2 \cdots w_{n-1} w_n$ in a test set, where the word in an out-of-vocabulary (OOV) is classified as an UNKNOWN word (Hirabayashi & Nakagawa, 2010):

$$H = -\tfrac{1}{n}\log_2 P(w_1 w_2 \cdots w_{n-1} w_n) \tag{2}$$

$$PP = 2^H \tag{3}$$

Cases of out-of-vocabulary and adjusted perplexity can be calculated:

$$APP = (P(w_1 w_2 \cdots w_{n-1} w_n) m^{n\mu})^{\frac{1}{n}}, \tag{4}$$

where n_μ represents the number of out-of-vocabulary words, and m represents the number of out-of-vocabulary items in a test set.

(K) Spectrum changing rate

Since a native speaker's English utterances are spontaneous, the spectrum's changing rate may vary rapidly. It can be calculated:

$$d(x(t), x(t-1)) = \sqrt{\sum_{i=1}^{n} (x_i(t) - x, (t-1))}. \tag{5}$$

We examined the Euclid distance between the adjacent frames of the calculated MFCC and used the standard variation and variance, where i represents the i-th index, x_i(t) represents the MFCC of the i-th dimension at the t-th frame, and $x_i(t-1)$ represents the MFCC in the previous frame of the i-th dimension.

(L) Phoneme-pair discrimination score

Using SVM, we identified and discriminated between the following nine pairs of phonemes that are often mispronounced by Japanese native speakers: */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/* (ATR Institute of Human Information, 2000, ATR Institute of Human Information, 1999).

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

The phoneme-pair discrimination score is a value that reflects a quantized distinction rate from 1 (poor) to 4 (excellent) for every sentence. Each sentence includes an average of 37 phoneme pairs.

5.2. Correlation between acoustic measure and the teacher's score

Tables 7 and 8 summarize the correlation between each acoustic or linguistic measure for every sentence and their averaged English teacher scores.

The number of sentences of each speaker was not constant. Additionally, to keep as many samples as possible, we computed a 5- and 10-sentence level as the following example; to

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

Table 7 *Correlation between acoustic/linguistic measures and pronunciation score*

Measure	1 sentence	5 sentences	10 sentences
LLnative	−0.466	−0.625	−0.669
LLnon-native	−0.638	−0.771	−0.804
LR	0.800	0.859	0.880
**LLbest*	−0.473	−0.613	−0.660
**LLmother*	0.719	0.804	0.811
**LRadap*	0.772	0.827	0.822
LR′	0.214	0.273	0.390
Phone recog(*Sub.*)	−0.298	−0.567	−0.662
Phone recog(*Del.*)	0.056	0.116	0.220
Phone 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	−0.020
**Pitch(F0)*	0.495	0.638	0.691
ROS	0.523	0.692	0.773
PP (WSJ)	−0.068	−0.151	−0.203
PP (EURO)	−0.077	−0.187	−0.257
APP (WSJ)	−0.051	−0.112	−0.145
APP (EURO)	−0.077	−0.187	−0.256
H (WSJ)	−0.298	−0.574	−0.719
H (EURO)	−0.007	−0.029	−0.077
Spectrum changing rate	0.320	0.339	0.329
Spectrum rate(SD)	0.400	0.517	0.578
Spectrum rate(variance)	0.413	0.532	0.592
Phoneme-pair	0.241	0.462	0.590

*Represents features calculated without correct transcription.

compute a 5-sentence level, we averaged the first 5 sentences' acoustic/linguistic measure values and averaged from the 2nd to the 6th sentence and so on to build a new averaged list. The correlation between each acoustic measure and human's score is shown in Tables 7 and 8.

Fairly high correlations are evident from Tables 7 and 8 for most of the likelihood measures (ex. $LL_{non\text{-}native}$, LR, LL_{mother}, LR_{adap}). The correlations between the intelligibility and acoustic/linguistic measures given in Table 8 improved considerably at levels with more than five sentences. These results show that utterances with more than five sentences are necessary to estimate intelligibility and the automatic word recognition performance is only slightly related to intelligibility. The correlation between intelligibility and LL_{best} gives the highest negative correlations: − 0.551 and − 0.731 at 5- and 10-sentence levels. Concerning perplexity, we expected that a speaker with good pronunciation might utter a complicated sentence and unfamiliar words for which a positive correlative value would be observed, but the results showed a negative value. This result indicates that pronunciation and intelligibility scores worsen when a speaker utters a complicated sentence and unfamiliar words.

Table 9 shows the phoneme-pair discrimination result.

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

Table 8 *Correlation between acoustic/linguistic measures and intelligibility*

Measure	1 sentence	5 sentences	10 sentences
LLnative	−0.180	−0.497	−0.684
LLnon-native	−0.202	−0.518	−0.690
LR	0.184	0.432	0.539
**LLbest*	−0.257	−0.551	−0.731
**LLmother*	0.177	0.317	0.353
**LRadap*	0.165	0.367	0.408
LR′	0.337	0.473	0.617
Phone recog(*Sub.*)	−0.052	−0.235	−0.254
Phone recog(*Del.*)	0.152	0.149	0.059
Phone recog(*Cor.*)	0.117	0.093	0.208
Word recog(WSJ, *Cor.*)	−0.013	0.047	0.071
Word recog(EURO, *Cor.*)	0.009	0.229	0.204
**Power*	−0.022	−0.051	−0.052
**Pitch(F0)*	0.196	0.405	0.527
ROS	0.166	0.397	0.513
PP (WSJ)	0.041	−0.006	0.024
PP (EURO)	−0.113	−0.188	−0.121
APP (WSJ)	0.045	0.024	0.085
APP (EURO)	−0.113	−0.188	−0.120
H (WSJ)	−0.047	−0.234	−0.461
H (EURO)	−0.052	−0.080	−0.047
Spectrum changing rate	0.197	0.339	0.404
Spectrum rate(SD)	0.098	0.160	0.245
Spectrum rate(variance)	0.101	0.168	0.255
Phoneme-pair	0.132	0.340	0.503

*Represents features calculated without correct transcription.

Table 9 *English phoneme-pair discrimination result using SVM [%]*

Phoneme-pair	Native	Japanese
l and *r*	97.2	69.3
m and *n*	92.5	87.7
s and *sh*	97.8	90.7
s and *th*	97.0	82.5
b and *v*	84.1	70.6
b and *d*	89.2	79.8
z and *dh*	98.3	83.6
z and *d*	98.2	89.6
d and *dh*	92.0	86.5
Average	94.0	82.3

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

The average correct discriminative rates of native English and Japanese English speakers using SVM were 94.0% and 82.3%. Among the nine phoneme-pairs, the pronunciation of */l* and */r*, */s* and */th*, */b* and */v*, */b* and */d*/ and */z* and */dh*/ was especially difficult for Japanese speakers in comparison with native speakers who can correctly pronounce them. Using these discrimination results, we can evaluate Japanese English pronunciation for individual phonemes.

The mark "*" represents a feature that is calculated without correct transcription. For the pronunciation scores, LL_{best}, LL_{mother}, and LR_{adap}, which are calculated using the likelihood rate, have high correlation. Both *Pitch(F0)* and the spectrum rate capture accent in English and have good correlation. *ROS* is also high. Comparing Tables 7 and 8, the correlation of intelligibility for all features except for *LR'* was lower than that of pronunciation, because we used man2/4 as the correct transcription, but it might be unstable.

6. Statistical method for evaluating each score and result

(Nakamura, Nakagawa & Mori, 2004)

For estimating the pronunciation and intelligibility scores, we proposed a linear regression model that was derived from the relationship between acoustic/linguistic measures and the scores of the English teachers. We established independent variables $\{x_i\}$ for the parameters and value Y for each English teacher's score and defined the linear regression model as:

$$Y = \sum_i (a_i \times x_i) + \varepsilon, \tag{6}$$

where ε is a residue (Ohta & Nakagawa, 2005). The coefficients $\{a_i\}$ are determined by minimizing the square of ε. Next, we experimented with open data for speakers by investigating whether our proposed method is independent of the speaker. In an open experiment with the speakers, we estimated the regression model using the utterances of 20 of the 21 speakers (the remaining speaker's score was estimated) at all 1-, 5- and 10-sentence levels. We repeated this experiment for each speaker.

Tables 10 and 11 summarize the evaluation results of the pronunciation and intelligibility scores obtained at the levels of 1-, 5-, and 10-sentences for the open data, which means that

Table 10 *Correlation between combination of acoustic/linguistic measures and pronunciation score rated by humans*

Acoustic/Intelligibility measures	1 sentence	5 sentences	10 sentences
LLnative, LLnon-native, LR, *LLmother*, Phone recog (*Del.*), *Power*, *H* (WSJ), Phoneme-pair	**0.807**	0.862	0.867
LR, Word recog (WSJ, Del.), Word recog (WSJ, Cor.), Word recog (EURO, Cor.), *Power*, *PP* (EURO), *APP* (EURO), *H* (WSJ), Phoneme-pair	0.751	**0.881**	**0.929**
**LLbest, *LRadap, *Power*	0.779	0.853	0.878

Boldface denotes best value. Last row denotes result with only features calculated without correct transcription.

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

Table 11 *Correlation between combination of acoustic/linguistic measures and intelligibility rated by humans*

Acoustic/Intelligibility measures	1 sentence	5 sentences	10 sentences
LLnon-native, *LLbest*, *LRmother*, *LRadap*, Pitch (F0), Phone recog (*Cor.*), *APP* (WSJ)	**0.476**	0.518	0.499
LRadap, *LR'*, Phone recog (*Cor.*), Word recog (EURO, *Cor.*), *Power*, Spectrum changing rate, Phoneme-pair	0.356	**0.652**	0.752
LLnon-native, *LR'*, Phone recog (*Sub.*), *PP* (WSJ), *PP* (EURO), *APP* (EURO), Phoneme-pair	0.129	0.537	**0.753**
**LLbest*	0.230	0.516	0.693

Boldface denotes best value. Last row denotes result with only features calculated without correct transcription.

for each speaker the test set is different from the training set. Here, boldface text denotes the best value from among the many combinations of feature parameters for every set of 1-, 5- and 10-sentences.

By combining certain acoustic/linguistic measures, we obtained correlation coefficients of 0.929 and 0.753 for the pronunciation and intelligibility scores using open data with each speaker at the 10-sentence levels. If we only use the features calculated without correct transcription, the correlation becomes 0.878 and 0.693 for the pronunciation and intelligibility scores showing that we can get sufficient performance for any utterance in comparison with native English teachers (Tables 5 and 6).

Figure 4 illustrates the relationship between the estimated pronunciation score/intelligibility and that of the English teachers based on the open data for a set of 10-sentence levels.

These results confirm that our proposed method for automatic estimation of pronunciation and intelligibility scores has approximately the same effectiveness as actual evaluations performed by English teachers.

7. Real-time estimation system

We designed an online real-time system based on the above method to learn English pronunciation. Since language/pronunciation learning that requires human intervention is expensive and time/space dependent, a Computer Assisted Language Learning system (CALL) is eagerly anticipated by second language learners.

We experimentally investigated whether it is possible to obtain the same performance in real-time both online and offline.

7.1. Description of online real-time system

We designed a real-time system for estimating the pronunciation and intelligibility scores in our laboratory for English pronunciation learning. We calculated the measures of the acoustic features in real-time to show the scores soon after a speaker finishes reading a specific sentence. Our real-time pronunciation and intelligibility scores estimation system consists of a front-end, an intermediate server, two word recognition servers, two likelihood

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

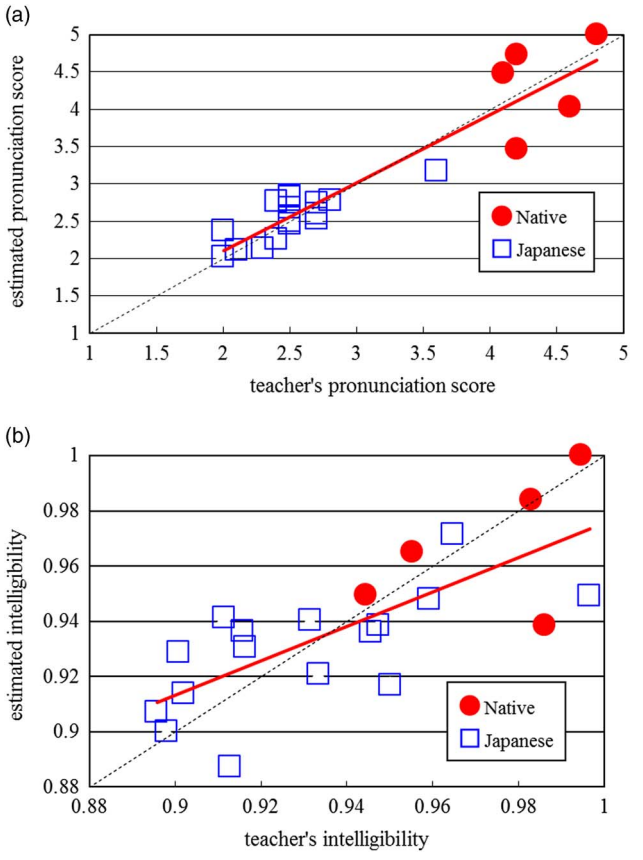

Fig. 4. Relationship between estimated and teacher scores (ten sentences). (a) Pronunciation score; (b) Intelligibility score

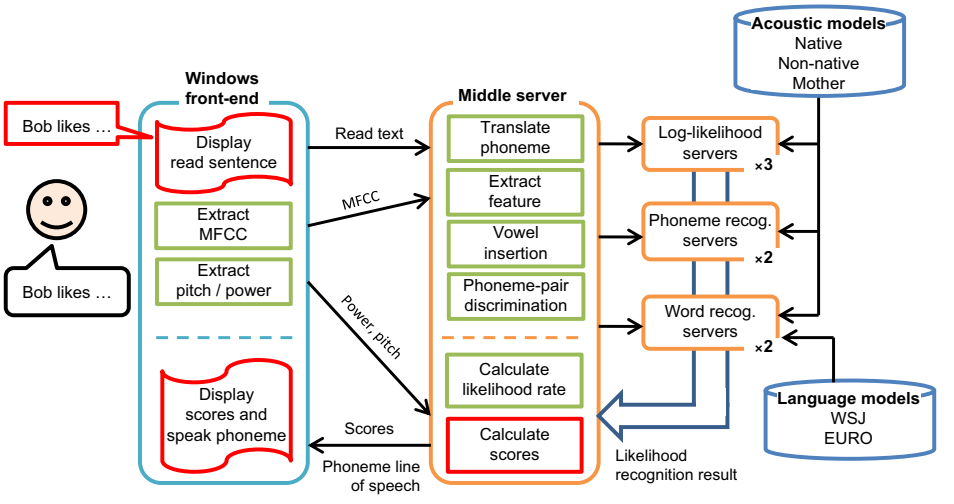

Fig. 5. Configuration of the execution of the system

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

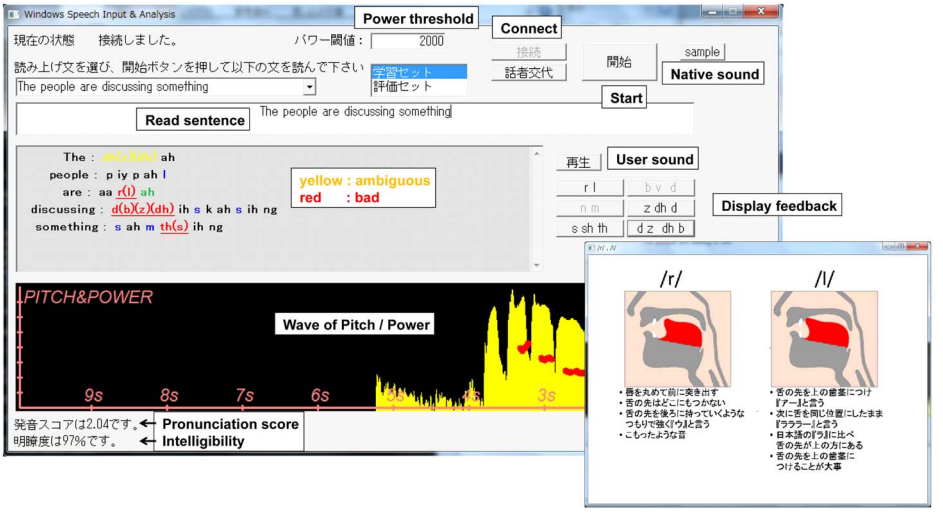

Fig. 6. Illustration of system execution

calculation servers, and three phoneme recognition servers (Figure 5). All servers were connected by a network to exchange data to obtain pronunciation and intelligibility scores in real-time. The front-end runs under a Windows operating system while all other programs run under Linux.

Although the word recognition server already had its own protocol for online real-time word recognition, we did not use it for score estimation since the word recognition speed was too slow for real-time estimation. All other servers finished their calculations at the same time as the speaker finished reading.

We normalized the feature parameter Mel Frequency Cepstrum Coefficient (MFCC) by the following equation to normalize the environmental condition in the recording and the speaker's differences:

$$\overline{MFCC(i)} = \frac{MFCC_0 \times \beta + MFCC_{total}(i-1) + MFCC(i)}{\beta + i} \tag{7}$$

where i represents the current frame of an utterance, β represents the weight of $MFCC_0$ (set to 50), $\mathrm{MTCC}_{\mathrm{total}}(i-1)$ represents the total MFCC from the first frame to frame, and $MFCC_0$ is an initial value of the standard MFCC average, which is obtained by processing 24 minutes' worth of read passage recordings by eight members in our laboratory and 24 minutes worth of recordings by 16 individuals from the "English Speech Database Read by Japanese Students" (ERJ) (see Minematsu, Tomiyama, Yoshimoto, Shimizu, Nakagawa, Dantsuji & Makino, 2002) recorded in quiet rooms.

Figure 6 depicts the front-end interface, where the "legend box" is not displayed. These are English translations of Japanese captions for readers. The following are the system's outputs: (a) pronunciation score, (b) intelligibility score, (c) pitch/power wave, (d) phoneme-pair recognition result, where blue denotes a well-pronounced word, yellow denotes an ambiguous one, red denotes a badly pronounced one, and green denotes an inserted vowel error between consonants (in Japanese, there is no phonotactics of consecutive consonants), (e) sounds of native speakers or users, and (f) articulation display of confusable consonants.

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

Table 12 *Contents of read sentences*

Phase	#Sentences	Contents	
Training	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)

Table 13 *Training and test data for on-line evaluation*

Phase	Score	Corpus	#male	#sentence	Comments
Test	Pronunciation score	Recording	8	480	Same sentences at every learner
	Intelligibility	Recording	8	48	different sentences

7.2. Conditions

We carried out an experiment with eight male Japanese students at our university to examine how our proposed system affects the learning of English pronunciation (Kibishi & Nakagawa, 2011). The eight students participated in the experiment as volunteers, but we paid them an allowance of 1,000 Yen/hour. The English ability of these students was not high, so they were not proficient in English conversation. The learning period lasted for about three weeks, 20 minutes per day, five times per week: fifteen learning sessions. To evaluate the system's effectiveness, evaluation test data were recorded before the experiment (pre), after ten learning intervals (mid), and again after 15 learning intervals (post). Table 12 gives the details of the read sentence set.

Table 13 summarizes the test data set for our on-line learning evaluation. The regression coefficient value was determined from Table 2 using all 21 speakers.

For the learning process, a training set of 100 sentences was prepared from the Tactics for Test of English for International Communication (TOEIC) (Grant, 2008). The sentences in this learning set were spoken by a native speaker, so the system presented a native voice to users. In addition, subjects learned the system's basic operation, so that at the time of learning they could easily use it. 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 sentences were selected from ERJ while taking three factors into consideration: pronunciation of phonemes (10 sentences), intonation (5 sentences), and rhythm (5 sentences) as shown in the following examples.

Examples of ERJ:

- Irish youngsters eat fresh kippers for breakfast.
- He told me that there was an accident.
- Legumes are a good source of vitamins.

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

Table 14 *Correlation coefficient between one native teacher's pronunciation score/intelligibility and average score of others*

(a) Pronunciation score		(b) Intelligibility	
Teacher	Correlation	Teacher	Correlation
F, the others	0.664	F, the others	0.845
G, the others	0.670	G, the others	0.873
H, the others	0.595	H, the others	0.731
I, the others	0.475	I, the others	0.772
J, the others	0.487	J, the others	0.808
K, the others	0.336	K, the others	0.752
Average	0.540	Average	0.800

For each of the 480 test sentences (eight learners × three times × 20 sentences), we obtained scores focusing on phoneme pronunciation, fluency, and prosody by six native English teachers (F, G, H, I, J, K).

7.3. Evaluation result

Table 14 shows that the correlation between the score for one native speaker and the average score of the others is moderately high, but not high enough. However, intelligibility's correlation is high. These correlations show large differences between Tables 5 and 6 and Table 14, because of the different spontaneous speech (off-line) vs. read speech (on-line), a special field content (including unfamiliar words, off-line) vs. a general field content (on-line), and speakers having better English skill (off-line) vs. speakers having standard English skill (on-line).

Regarding intelligibility, the teachers transcribed 48 user utterances (eight learners × three times × two sentences), which are selected randomly from the training set. Users uttered sentences prepared *a priori*. Let the number of words in the read sentence be A and the number of words correctly transcribed be B. Intelligibility is calculated as $\frac{A}{B}$ (see Section 4.2).

Table 15 gives the experimental results of our online system experiment. Here, the "◎" mark denotes that the post-test achieves the best score of the three tests, "○" denotes that the post-test relatively outperforms the pre- and mid-tests, "Δ" denotes that the pre-test is comparable with the post- and mid-tests, and "×" denotes that the pre-test achieves the best score among the three tests. In this result, intelligibility score is different from that of the test set (Kibishi, Hirabayashi & Nakagawa, 2012, in Japanese), so the score is a little low. We used the (badly pronounced comparatively) utterances in pronunciation practice, because we needed many independent utterances to be transcribed by native teachers. The subject "h" has strong Kansai dialect, therefore his performance was different from other subjects.

We defined the rate of improvement as (difference in score of mid-test or post-test and pre-test)/(score of pre-test). For almost all cases of the on-line pronunciation learning results, the post-test results surpassed those of the pre-test. The rate was about 10% for intelligibility; moreover, the perceptual error reduction rate (improvement rate of intelligibility/mis-perception rate) was about 30% as shown in parentheses. Five to seven out of the eight learners improved their pronunciation and intelligibility scores in the objective and

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

Table 15 *Scores of system (estimated score) and native teacher for every test*

Score/Subject			*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	Average	Rate of improve [%]
System	pronunciation score (1 ~ 5)	pre	2.60	2.76	3.22	2.62	2.89	2.43	2.70	2.67	2.74 -	–
		mid	2.78	2.57	3.23	3.23	3.36	2.24	2.95	2.78	2.89 -	5.5
		post	3.50	3.07	3.52	3.11	3.37	2.58	3.23	2.34	3.08 -	12.4
			◎	◎	◎	○	◎	◎	◎	△	◎	
	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	100.0	84.9	14.3 (41.2)
		post	83.9	100.0	87.9	52.9	69.7	84.7	79.0	77.1	79.4	6.9 (19.8)
			×	◎	○	△	○	◎	×	○	○	
Estimated by teachers	pronunciation score (1 ~ 5)	pre	2.67	2.79	2.46	2.72	2.70	2.45	2.90	3.12	2.73 -	–
		mid	2.77	2.86	2.69	2.98	2.67	2.63	3.01	2.80	2.80 -	2.6
		post	2.96	2.96	2.63	2.77	2.36	2.60	3.10	2.38	2.72 -	–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	11.1 (34.3)
		post	88.1	95.9	85.6	77.2	74.5	83.3	79.5	69.2	81.7	8.2 (24.5)
			○	◎	◎	×	◎	◎	△	△	○	

The values in parentheses denote the mis-perception reduction rate.

Table 16 *Correlation coefficient between automatically estimated and averaged teacher scores*

(a) Pronunciation score		(b) Intelligibility	
Time	Correlation	Time	Correlation
pre	0.539	pre	0.808
mid	0.510	mid	0.688
post	0.476	post	0.712
all	0.492	all	0.747

subjective evaluations. Although the improvement in pronunciation was only a few percentage points, the rate increased to 3.7–4.9% in the case of all learners except for subject "h."

For the pronunciation score/intelligibility, (Table 16 shows the correlation coefficient for the estimated and native teacher scores.

From Tables 14 and 16, since the system estimated the pronunciation score/intelligibility with a correlation of 0.492/0.747 between the estimated and average native scores, we believe that the estimation is adequate, because the correlation among the native teacher scores was 0.540/0.800 (Table 14).

For intelligibility, since Table 14(b) shows that the correlation among teachers is high, teachers stably calculated intelligibility for read speech. Table 16(b) also shows a high correlation between teachers and the system. Our proposed system stably calculated intelligibility in a manner that resembled that of the teachers.

7.4. Evaluation by questionnaire

Finally, at the end of our experiment, students completed a questionnaire with the following main questions (5: excellent ~ 1: bad on the average). The feedback in this system denotes functions to indicate mispronounced phonemes, listening to user or native utterances, and showing correct pronunciation. The questions and averaged answers are as follows:

(Q.I)	Was the pronunciation estimation valid?	==>	3.13
(Q.II)	Was the intelligibility estimation valid?	==>	3.00
(Q.III)	Was this system useful for learning English pronunciation?	==>	3.75
(Q.IV)	Was its feedback valid?	==>	3.25
(Q.V)	Did the feedback improve your pronunciation?	==>	4.00
(Q.VI)	Did listening to a native speaker's voice improve your pronunciation?	==>	4.13
(Q.VII)	Do you want to use this system for learning English pronunciation?	==>	3.63
(Q.VIII)	Overall, were you satisfied with this system?	==>	3.75

The detailed answers for the functions are summarized in Table 17. From the answers, we found that the pronunciation score, listening to a native speaker's voice, indication of mispronounced phonemes, and listening to user utterances facilitated learning and simultaneously provided motivation to practice. Almost all of the learners reported that it was better to see one's own mispronounced phonemes checked by this proposed system. Six out of eight learners reported that they wanted to use this system for learning English pronunciation.

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

Table 17 *Questionnaire: Whether each functions was useful*

Function \ Learner	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	Ave.	Rate of over 4 [%]
Power/pitch contours	3	2	2	1	2	1	1	3	1.9	0
Pronunciation score	5	4	4	3	4	4	5	4	4.1	87.5
Intelligibility	4	4	4	1	3	3	1	3	2.9	37.5
Display mispronounced phonemes	5	4	3	4	3	4	5	3	3.9	62.5
Feed-back	3	3	2	4	3	3	5	4	3.4	37.5
Sound of user's utterance	3	5	5	5	3	3	1	4	3.6	50.0
Sound of native utterance	4	5	5	5	5	5	5	5	4.9	100
Overall system	5	4	4	4	4	4	5	5	4.4	100
Want to continue to use	4	3	4	4	4	4	3	5	3.9	75.0

Regarding the intelligibility scores, however, the subjects did not know how to use them for measuring how much native speakers could comprehend their English. The power and pitch contours were also not used during the pronunciation training because users did not know how to practice with them.

In addition, from question responses that compared our system with self-directed learning (e.g., repeating utterance using CDs), it is clear that the estimated scores and display of mispronounced phonemes determined by phoneme-pair discrimination provide motivation for continued practice. From the answers concerning self-directed learning, we ascertained that the information about the quality of the subject's own pronunciation and which phonemes must be improved is very important.

At the same time as scoring and transcription were performed, native English teachers marked the mispronounced words. We investigated the relationship between the number of mispronounced words, pronunciation score, and the number of incorrectly discriminated phoneme pairs. The correlation between estimated pronunciation score by the system and the number of mispronounced words was found to be 0.493; furthermore, a higher correlation of 0.610 was obtained between the pronunciation score by teachers and the number of mispronounced words. The number of mispronounced words had an even higher correlation, at 0.629, with the rate of incorrectly discriminated phoneme-pairs. These findings support the validity of our system.

Table 17 shows the answers from the questionnaire. The pronunciation score and the display of mispronounced phonemes obtained a good rating as seen by their high average scores. However, some subjects reported that they could not understand how to use the intelligibility score, which denotes how well native teachers perceive the words of the learner's utterance. They also could not understand how to use power/pitch contours for learning pronunciation. These issues remain for future study. Compared with self-directed learning (e.g., repeating utterances using CDs), it is apparently beneficial to see one's own mispronounced phonemes using this proposed system.

8. Conclusion

In this paper, we proposed a statistical method for estimating the pronunciation and intelligibility scores of presentations made in English by non-native speakers based on a

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

linear regression model offline. By combining acoustic and linguistic measures, our proposed method evaluated pronunciation and intelligibility scores with almost the same accuracy and effectiveness as native English teachers. Our evaluation system could also estimate these functions without a correct transcription of the learner's utterances.

We also developed an online learning evaluation system for English pronunciation targeted at Japanese speakers. Through experiments using this system to practice English pronunciation, we confirmed its positive learning effects: The pronunciation proficiency and intelligibility of learners were improved by using the proposed on-line system. From questionnaires, we ascertained that the pronunciation scores, listening to a native voice, indications of mispronounced phones, and listening to the user's own utterances provided motivation to practice. Six learners out of the eight subjects reported that they wanted to use this system for learning English pronunciation.

Future work will integrate more useful functions into our online system, in particular, graphical user feedback with more emphasis on interpersonal skills. Finally, based on the knowledge obtained here, we want to improve the performance of non-native speech recognition.

References

Acoustical Society of America SII. Speech Intelligibility Index. http://www.sii.to/index.html

Aist, G. (1999) Speech recognition in computer-assisted language learning. In: Cameron, K. (ed.), *Computer Assisted Language learning; Media, Design and applications*. Lisse, The Netherlands: Swets & Zeitlinger, 165–181.

ATR Institute of Human Information. (2000) *Full version Scientific Progress Method for English speaking*. Tokyo, Japan: Kodansha.

ATR. (1999) *Full version Scientific Progress Method for English Speaking*. Tokyo, Japan: Kodansha.

Cucchiarini, C., Strik, H. and Bovels, L. (2000) Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, **30**(2–3): 109–119.

Eskenazi, M., Kennedy, A., Ketchum, C., Olszewski, R. and Pelton, G. (2007) The native accent pronunciation tutor: measuring success in the real world. *Proceedings of SIG-SlaTE*. Baixas, France: ISCA, 124–127.

Falk, T. H., Chan, W. and Shein, F. (2012) Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication*, **54**(5): 622–631.

Franco, H., Neumeyer, L., Kim, Y. and Ronen, O. (1997) Automatic pronunciation scoring for language instruction. *Proceedings of ICASSP*. New York: IEEE, 1471–1474.

Fujisawa, Y., Minematsu, N. and Nakagawa, S. (1998) Evaluation of Japanese manners of generation word accent of English based on a stressed syllable detection technique. *Proceedings of ICSLP*. Baixas, France: ISCA, 3103–3106.

Garofalo, J. D., Graff, D. Paul, and Pallett, D. (2007) CSR-I (WSJ0) *Complete Linguistic Data Consortium*. Philadelphia, USA: LDC.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L. and Zue, V. (1993) *TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium*. Philadelphia, USA: LDC.

Grant, T. (2008) *Tactics for TOEIC Listering and Reading Test Student Book*. Oxford, UK: Oxford University Press.

Hirabayashi, K. and Nakagawa, S. (2010) Automatic evaluation of English pronunciation by Japanese speakers using various acoustic features and pattern recognition techniques. *Proceedings of Interspeech*. Baixas, France: ISCA, 598–601.

Holliday, J. J., Beckman, M. E. and Mays, C. (2010) Did you say susi or shushi? measuring the emergence of robust fricative contrasts in English- and Japanese-acquiring children. *Proceedings of Interspeech*. Baixas, France: ISCA, 1886–1889.

Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T. and Shikano, K. (1999) JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of the Acoustical Society of Japan (E)*, **20**(3): 199–206.

Karafiat, M., Janda, M., Cernocky, J. and Burget, L. (2012) Region dependent linear transforms in multilingual speech recognition. *Proceedings of ICASSP*. New York: IEEE, 4885–4888.

Kawahara, T. and Minematsu, N. (2011) Tutorial on Computer-assisted language learning (CALL) based on speech technologies. *Proceedings of APSIPA Tutorial session*. Hong Kong: APSIPA.

Kibishi, H. and Nakagawa, S. (2011) New feature parameters for pronunciation evaluation in English presentations at international conferences. *Proceedings of Interspeech*. Baixas, France: ISCA, 1149–1152.

Kibishi, H., Hirabayashi, K. and Nakagawa, S. (2012) Development of Online Evaluation System of English Pronunciation Score/Intelligibility for Japanese. *Proceedings of Acoustical Society of Japan* (in Japanese), Tokyo, Japan: ASJ, 499–502.

Kobayashi, T., Itahashi, S., Hayamizu, S. and Takezawa, T. (1992) ASJ continuous speech corpus for research. *Journal of the Acoustical Society of Japan (J)* (in Japanese), **48**(12): 888–893.

Koniaris, C. and Engwall, O. (2011) Perceptual differentiation modeling explains phoneme mispronunciation by non-native speakers. *Proceedings of ICASSP*. New York: IEEE, 5704–5707.

Li, H, Wang, S., Liang, J., Huang, S. and Xu, B. (2009) High performance automatic mispronunciation detection method based on neural network and TRAP features. *Proceedings of Interspeech*. Baixas, France: ISCA, 1911–1914.

Minematsu, N., Tomiyama, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M. and Makino, S. (2002) English Speech Database Read by Japanese Learners for CALL System Development. *Proceedings of of the International Conference on Language Resources and Evaluation (LREC 2002)* Paris, France: ERLA, 896–903.

Nakagawa, S., Reyes, A. A., Suzuki, H. and Taniguchi, Y. (1997) An English conversation and pronunciaiton CAI system using speech recognition technology. *Proceedings of Eurospeech*. Baixas, France: ISCA, 705–708.

Nakagawa, S. and Ohta, K. (2007) A statistical method of evaluating pronunciation proficiency for presentation in English. *Proceedings of Interspeech*. Baixas, France: ISCA, 2317–2320.

Nakagawa, S., Reyes, A., Suzuki, A., Reyes, H., Allen, A. and Taniguchi, Y. (1997) An English conversation CAI system using speech recognition technology, (in Japanese). *Trans. Information Processing Society in Japan*, **38**(8): 1649–1658.

Nakamura, N., Nakagawa, S. and Mori, K. (2004) A statistical method of evaluating pronunciation proficiency for English works spoken by Japanese. *IEICE Trans. Information and Systems*, **E87-D**(7): 1917–1922.

Neri, A., Cucchiarini, C. and Strik, H. (2008) The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCall*, **20**(2): 225–243.

Neumeyer, L., Franco, H., Weintraub, M. and Price, P. (1996) Automatic text-independent pronunciation scoring of foreign language student speech. *Proceedings of ICSLP*. Baixas, France: ISCA, 1457–1460.

Ohta, K. and Nakagawa, S. (2005) A statistical method of evaluating pronunciation proficiency for Japanese words. *Proceedings of Interspeech*. Baixas, France: ISCA, 2233–2236.

Ramos, M., Franco, H., Neumeyer, L. and Bratt, H. (1999) Automatic detection of phone-level mispronunciation for language learning. *Proceedings of EuroSpeech*. Baixas, France: ISCA, 851–854.

Ronen, O., Neumeyer, L. and Franco, H. (1997) Automatic detection of mispronunciation for language instruction. *Proceedings of Eurospeech*. Baixas, France: ISCA, 645–648.

Smit, P. and Kurimo, M. (2011) Using stacked transformations for recognizng foreign accented speech. *Proceedings of IEEE*. New York: IEEE, 5008–5111.

Stenson, N., Downing, B., Smith, J. and Smith, K. (1992) The effectiveness of computer-assisted pronunciation training. *CALICO Journal*, **9**(4): 5–19.

TED Translanguage English Database. http://www.elda.org/catalogue/en/speech/S0031.html

Tsubota, Y., Kawahara, T. and Dantsuji, M. (2002) Recognition and verification of English by Japanese students for computer-assisted language learning system. *Proceedings of ICSLP*. Baixas, France: ISCA, 1205–1208.

Wang, Y.-B. and Lee, L.-S. (2012) Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training. *Proceedings of ICASSP*. New York: IEEE, 5049–5052.

Witt, S. and Young, S. (1999) Computer-Assisted pronunciation teaching based on automatic speech recognition. In: Jager, S., Nerbonne, J. and Essen, A. V. (eds.), *Language Teaching and Language Technology*. Lisse, The Netherlands: Swets & Zeitlinger, 25–35.

Wu, C., Su, H. and Liu, C. (2012) Efficient personalized mispronunciation detection of Taiwanese-accented English speech based on unsupervised model adaptation and dynamic sentence selection. *Computer Assisted Language Learning*, **23**(5): 446–467.

Yoon, S.-Y, Hasegawa-Johnson, M. and Sproat, R. (2009) Automated pronunciation scoring using confidence scoring and landmark-based SVM. *Proceedings of Interspeech*. Baixas, France: ISCA, 1903–1906.

Young, S. and Witt, S. (1999) Offline acoustic modeling of nonnative accents. *Proceedings of Eurospeech*. Baixas, France: ISCA, 1367–1370.

Zhao, Y. and He, X. (2001) Model complexity optimization for nonnative English speakers. *Proceedings of Eurospeech*. Baixas, France: ISCA, 1461–1463.

Appendix

A.1. Speech Analysis

The speech was down-sampled to 16 kHz and pre-emphasized, and then a 25-ms wide Hamming window was applied every 10-ms. 12-dimensional MFCCs were used as speech feature parameters for a frame. The acoustic features were 12 MFCCs and their Δ (velocity, time derivation) and $\Delta\Delta$ (acceleration, $2^{\text{-nd}}$ order time derivation) features, in total of 36 dimensions.

A.2. Formal Model of Speech Recognition

Automatic speech recognition (ASR) task is to find the corresponding word sequence for a given acoustic signal. Given a speech signal A, ASR systems find the corresponding word sequence $\hat{W}$ that has maximum *posterior probability P(W/A)* according to Bayes' theorem as follows:

$$\begin{aligned}\hat{W} &= \arg\max_{w} P(W/A)\\ W &= \text{are}\max_{w} \frac{P(A/W)P(W)}{P(A)}\\ W &= \text{are}\max_{w}\ P(A/W)P(W)\\ W &= \text{are}\max_{w}(\log P(A/W) + \log P(W)),\end{aligned}$$

where $P(A/W)$ is the probability of A given W based on acoustic model, $P(W)$ is the probability of W based on LM. In general, the LM task is to assign a probability to a

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press

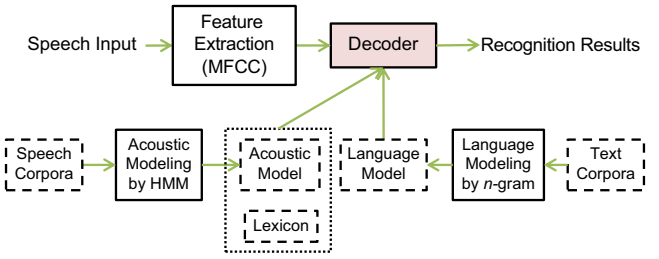

Fig. 7. Diagram of a general ASR system

word sequence. Figure 7 shows the diagram of a general ASR system. $P(A)$ is calculated approximately by

$$P(A) = \sum P(A, W) \approx \max P(A, W)$$

This corresponds to recognition likelihood for arbitrary word (phoneme) sequence without a language model. We call these log- probabilities as *log likelihoods*.

A.3. Acoustic Model by HMM

P(A/W) is calculated by an acoustic model, which has been represented by HMM Acoustic models based on monophone HMMs were learned by the analyzed speech. The English HMMs were composed of three states, each of which has four Gaussian mixture distributions with full covariance matrices. The number of monophones was 39. The Japanese syllable-based HMMs were composed of four states, each of which has four Gaussian mixture distributions with full covariance matrices. The number of syllables was 117.

In the proposed system, we used three different HMMS as follows (see Figure 1):

/Native English phoneme HMM trained by native English data.
/Non-native English phoneme HMM trained by English data uttered by Japanese.
/Native Japanese syllable HMM trained by Japanese data.

A.4. Language Model by n-gram

The word-based n-gram LM is the most common LM currently used in ASR systems. It is a simple yet quite powerful method based on the assumption that the current word depends only on the preceding words. This LM predicts the current word based on preceding words. Given word history $w_{i-n+1}^{i-1} = w_{i-n+1}, \ldots, w_{i-1}$, word-based n-gram predicts the current word w_i according to the following equation:

$$P\left(w_1^n\right) = \prod_{i=1}^{n} P(w_i \mid w_1 w_2 \cdots w_{i-1}) \approx \prod_{i=1}^{n} P\left(w_i \mid w_{i-n+1}^{i-1}\right)$$

for some $n \geq 1$. The number of n is closely related with the parameter number in the LM. We used the word-level n-gram LM for LVCSR, and the phoneme-level n-gram LM for phoneme recognition.

https://doi.org/10.1017/S0958344014000251 Published online by Cambridge University Press