Statistical Method of Evaluating Pronunciation Proficiency/Intelligibility of English Presentation by Japanese speakers

Hiroshi Kibishi, Kuniaki Hirabayashi and Seiichi Nakagawa
Toyohashi University of Technology, Computer Science and Engineering, Toyohashi, Japan
e-mail: {kibishi, 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 speaker

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 and Minematsu 2011), some of which use speech recognition techniques (S. Nakagawa, A. A. Reyes, et al. 1997, Tsubota, Kawahara and Dantsuji 2002, Eskenazi, et al. 2007). CAPT is a crucial component of CALL that focuses on evaluating pronunciation proficiency or correcting pronunciation errors.

We 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 and Nakagawa 1998, Nakamura, Nakagawa and Mori 2004). In this paper, we propose extended pronunciation proficiency/intelligibility estimation methods and developed an on-line estimation system. Using this system, we evaluated the learning effect in pronunciation proficiency and showed the improvement in intelligibility of learner’s utterances.

1.1 Three approaches to pronunciation assistance

Computer assisted pronunciation training (CAPT) has a compelling motivational effect (Stenson, et al. 1992). Aist classified pronunciation assistance into three general approaches (Aist 1999). The first approach is 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 et al. proposed an automatic text-independent pronunciation scoring method (Neumeyer, et al. 1996) 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. Franco et al., who investigated evaluation measures based on HMM-based phone log-posterior probability scores and the combination of the above scores (Franco, Neumeyer, et al. 1997) 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, et al. 1999). We also investigated posterior probability as an evaluation measure for Japanese (S. Nakagawa, A. A. Reyes, et al. 1997). Cucchiarini 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 \( LR \) (Cucchiarini, Strik and Bovels 2000). They compared 3 CAPT systems; their 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 (Neri, Cucchiarini and Strik 2008). Wang and Lee integrated Error Pattern (EP)-based and Goodness-of-Pronunciation (GOP, (Witt and Young 1999))-based mispronunciation detectors in a serial structure to improve a mispronunciation
detection system (Wang and Lee 2012).

Koniaris and Engwall 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 (Koniaris and Engwall 2011). To evaluate phoneme pronunciation, Yoo et al. utilized a Support Vector Machine (SVM) using Perceptual Linear Predictive (PLP) features and formant information as acoustic feature parameters (Yoon, Hasegawa-Johnson and Sproat 2009). To automatically detect mispronounced phonemes, Li et al. combined three methods: Neural Network (NN) & MLP-NN using TempoRAL Patterns (TRAP) features, SVM, and Gaussian Mixture Model (GMM) (Li, et al. 2009). Smit and Kurimo recognized individual accent utterances using stacked transformations (Smit and Kurimo 2011). For the speech recognition of non-native speakers, linear or nonlinear transformations are usually input to HMM-based acoustic models (Karafiat, et al. 2012).

The above studies were evaluated for European languages or English uttered by European non-native speakers. Wu et al. presented an efficient approach to detecting personalized mispronunciation in Taiwanese-accented English (Wu, Su and Liu 2012). Holliday et al., who focused on fricative sounds like shu whose pronunciation is difficult for non-native speakers, distinguished between English and Japanese speakers (Holliday, Beckman and Mays 2010). In contrast, we evaluated the Japanese spoken by foreign students (Ohta and 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 to with the above research works, 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 propose a statistical method for estimating the pronunciation and intelligibility scores of presentations given in English by Japanese speakers (Nakagawa and Ohta 2007, Hirabayashi and Nakagawa 2010, Kibishi and 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 and 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.

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. Fig. 1 shows a block diagram of our evaluation system for pronunciation and intelligibility scores.

![Fig. 1 Block diagram of our estimation system for pronunciation and intelligibility scores](image)

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 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 and Ohta 2007, Hirabayashi and Nakagawa 2010, Kibishi and 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 21 speakers × 10–21 sentences for, 289 English sentences in presentations spoken by 21 male speakers at three skill levels of pronunciation proficiency: above average, average, or below average. 16 of 21 are Japanese speakers, and the remaining five are native speakers from the USA.

We used the TIMIT (Garofolo, et al. 1993)/WSJ (Garofalo, et al. 2007) database for training the native English phoneme HMMs, which is another Japanese speech database for adapting non-native English phoneme HMMs (Nakagawa, et al. 1997, in Japanese), and the ASJ (Itahashi, et al. 1997)/JNAS (Itou, et al. 1999) database for training the native Japanese syllable HMMs (strictly speaking, mora-unit HMMs).

Tables 1 and 2 summarize the speech materials.

<table>
<thead>
<tr>
<th>Table 1 Speech material training data for HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Japanese</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2 Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker (Database)</td>
</tr>
<tr>
<td>Native Japanese (TED)</td>
</tr>
<tr>
<td>Native Japanese (TED)</td>
</tr>
<tr>
<td>Total (TED)</td>
</tr>
</tbody>
</table>

Witt et al. 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, Ronen and Neumeyer 2001, Young and Witt 1999, Zhao and 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.

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

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Model</th>
<th>Del</th>
<th>Ins</th>
<th>Subs</th>
<th>Cor</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native</td>
<td>monophone</td>
<td>14.8</td>
<td>2.6</td>
<td>31.8</td>
<td>53.4</td>
<td>50.8</td>
</tr>
<tr>
<td></td>
<td>triphone</td>
<td>8.7</td>
<td>3.0</td>
<td>24.8</td>
<td>66.5</td>
<td>63.6</td>
</tr>
<tr>
<td>Japanese</td>
<td>monophone</td>
<td>11.9</td>
<td>3.1</td>
<td>51.5</td>
<td>36.6</td>
<td>33.5</td>
</tr>
<tr>
<td></td>
<td>triphone</td>
<td>4.4</td>
<td>11.5</td>
<td>52.2</td>
<td>43.4</td>
<td>31.9</td>
</tr>
</tbody>
</table>

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.

We assessed the test data by four out of the above five native English teachers for all 289 sentences and calculated the intelligibility. They 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:

\[ \text{Intelligibility} = \frac{A}{B} , \]
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

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

\[
\text{Intelligibility} = \frac{A}{B} = \frac{18}{19.75} = 0.911
\]

*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

\[
\text{Intelligibility} = \frac{A}{B} = \frac{10}{11.5} = 0.902
\]

Here, the mark “@” denotes a word/phrase by the speaker that the teacher heard a word 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}.

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

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Pronunciation score</th>
<th>Intelligibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B, C, D, E</td>
<td></td>
<td>A, B, C, D</td>
</tr>
</tbody>
</table>

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

(a) Among teachers  
(b) Among teachers and others

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Correlation</th>
<th>Teacher</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B</td>
<td>0.591</td>
<td>A, the others</td>
<td>0.921</td>
</tr>
<tr>
<td>A, C</td>
<td>0.570</td>
<td>B, the others</td>
<td>0.631</td>
</tr>
<tr>
<td>A, D</td>
<td>0.986</td>
<td>B, the others</td>
<td>0.936</td>
</tr>
<tr>
<td>A, E</td>
<td>0.933</td>
<td>C, the others</td>
<td>0.614</td>
</tr>
<tr>
<td>B, C</td>
<td>0.586</td>
<td>D, the others</td>
<td>0.519</td>
</tr>
<tr>
<td>B, D</td>
<td>0.588</td>
<td>E, the others</td>
<td>0.945</td>
</tr>
<tr>
<td>B, E</td>
<td>0.519</td>
<td>Average</td>
<td>0.683</td>
</tr>
<tr>
<td>C, D</td>
<td>0.597</td>
<td></td>
<td>0.794</td>
</tr>
<tr>
<td>C, E</td>
<td>0.520</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D, E</td>
<td>0.945</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.683</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, man2/4</td>
<td>0.750</td>
</tr>
<tr>
<td>B, man2/4</td>
<td>0.701</td>
</tr>
<tr>
<td>C, man2/4</td>
<td>0.649</td>
</tr>
<tr>
<td>D, man2/4</td>
<td>0.689</td>
</tr>
<tr>
<td>Average</td>
<td>0.697</td>
</tr>
</tbody>
</table>

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. Fig. 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. On the other hand, the higher the pronunciation score is, the more the person can correctly comprehend the speaker’s utterances.
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 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 and Mori 2004, Ohta and Nakagawa 2007). In this paper, we added new features to estimate the pronunciation and intelligibility scores. Although the previous work (Nakagawa and 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.

![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](image)

5.1 Acoustic measures
(A) Log-likelihood using native and non-native English HMMs and the learner's native language HMM (Nakamura, Nakagawa and 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 ($LL_{\text{native}}$), non-native English phoneme HMMs that adapted by Japanese utterances ($\star$), and native Japanese syllable HMMs ($\dagger$).

**B) Best log-likelihood for arbitrary phoneme sequences** (Nakamura, Nakagawa and 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 ($LL_{\text{best}}$).

**C) Likelihood ratio** (Nakamura, Nakagawa and 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:

$$LR = LL_{\text{native}} - LL_{\text{non-native}}.$$  

**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**

Fig. 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 $LL_{\text{native}}$ is not suitable for outlying speakers. However, the difference in the likelihoods between $B$ and $C$ compensates/normalizes this phenomenon. In Fig. 3, $LL_{\text{native}}$ for samples $B$ and $C$ is almost similar. On the other hand, 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 and Mori 2004)

We used the likelihood ratio ($LR^p$) between the log-likelihood of native English HMMs ($LL_{\text{native}}$) and the best log-likelihood for arbitrary phoneme sequences ($LL_{\text{best}}$), which means the logarithm of posterior probability: (S. Nakagawa, H. Reyes, et al. 1997, in Japanese, S. Nakagawa, A. A. Reyes, et al. 1997).

(E) Likelihood ratio for phoneme recognition (Nakamura, Nakagawa and Mori 2004)

We used the ratio of the likelihood of free phoneme recognition between native and non-native English HMMs ($LR_{\text{free}}$), which were defined as the difference between the two log-likelihoods: .

We also used the ratio of the likelihood of the free phoneme (syllable) recognition between native English and native Japanese HMMs ($LR_{\text{syll}}$), which were defined as the difference between the two log-likelihoods: .

(F) Phoneme recognition result (Nakamura, Nakagawa and 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 and Nakagawa 2005). This measure also requires the correct transcription of utterances.

(H) Standard deviation of powers and pitch frequencies
The standard deviations of powers \((P_{\text{power}})\) and fundamental (pitch) frequencies \((F_0)\) 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.

(J) Perplexity and entropy
Perplexity can be used to evaluate an utterance complexity. 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 and Ohta 2007). Entropy \(H\) and perplexity \(PP\) can be calculated for word sequence in a test set, where the word in an out-of-vocabulary (OOV) is classified as an UNKNOWN word (Hirabayashi and Nakagawa 2010):

\[
PP = 2^H. \tag{2}
\]

\[
\text{Cases of out-of-vocabulary and adjusted perplexity can be calculated:}
\]

\[
\text{,} \tag{3}
\]

where \(n_\mu\) 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:
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, \( x_t^i \) represents the MFCC of the \( i \)-th dimension at the \( t \)-th frame, and \( x_{t-1}^i \) represents the MFCC in the previous frame of the \( i \)-th dimension.

(L) Phoneme-pair discrimination score

Using SVM, we discriminated 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 ΔMFCC.

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 human’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 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.

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

* represents features calculated without correct transcription.
Fairly high correlations are evident from Tables 7 and 8 for most of the likelihood measures (ex. $LL_{\text{native}}$, $LR$, $LL_{\text{non-native}}$, $LL_{\text{best}}$, $LL_{\text{mother}}$, $LL_{\text{adapt}}$). 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 $LL_{\text{best}}$ 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. 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 /θ/, /b/ and /v/, /b/ and /d/ and /z/ and /ð/ 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, \(LL_{best}\) and \(LL_{mother}\), 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. With 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.

### Table 8 Correlation between acoustic/linguistic measures and intelligibility;

* represents features calculated without correct transcription.

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

### Table 9. English phoneme-pair discrimination result using SVM [%]
<table>
<thead>
<tr>
<th>Phoneme-pair</th>
<th>Native</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>l and r</td>
<td>97.2</td>
<td>69.3</td>
</tr>
<tr>
<td>m and n</td>
<td>92.5</td>
<td>87.7</td>
</tr>
<tr>
<td>s and sh</td>
<td>97.8</td>
<td>90.7</td>
</tr>
<tr>
<td>s and th</td>
<td>97.0</td>
<td>82.5</td>
</tr>
<tr>
<td>b and v</td>
<td>84.1</td>
<td>70.6</td>
</tr>
<tr>
<td>b and d</td>
<td>89.2</td>
<td>79.8</td>
</tr>
<tr>
<td>z and dh</td>
<td>98.3</td>
<td>83.6</td>
</tr>
<tr>
<td>z and d</td>
<td>98.2</td>
<td>89.6</td>
</tr>
<tr>
<td>d and dh</td>
<td>92.0</td>
<td>86.5</td>
</tr>
<tr>
<td>Average</td>
<td>94.0</td>
<td>82.3</td>
</tr>
</tbody>
</table>


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:

\[
Y = \sum_{i} (\alpha_i \times x_i) + \varepsilon
\]

where \( \varepsilon \) is a residue (Ohta and Nakagawa 2005). The coefficients \( \{\alpha_i\} \) are determined by minimizing the square of \( \varepsilon \). 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 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.

**Table 10** Correlation between combination of acoustic/linguistic measures and pronunciation score rated by humans; boldface denotes best value. Last row denotes result with only features calculated without correct transcription.
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).

Fig. 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.

![Diagram](image1)

(a) Pronunciation score

![Diagram](image2)

(b) Intelligibility score

**Fig. 4** Relationship between estimated and teacher scores (ten sentences)

---

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

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

*Boldface denotes best value. Last row denotes result with only features calculated without correct transcription.*
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, Computer Assisted Language Learning system (CALL) is eagerly anticipated. 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 calculation servers, and three phoneme recognition servers (Fig. 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.

![Fig. 5 Configuration of the execution of the system](image)

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

$$\frac{MFCC(i)}{\beta + i} = \frac{MFCC_0 \times \beta + MFCC_{\text{tot}}} {\beta + i}$$

where $i$ represents the current frame of an utterance, $\beta$ represents the weight of $MFCC_0$ (set to 50), $\beta + i$ represents the total MFCC from the first frame to frame $i$, 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) (ERJ English Speech Database Read by Japanese Students) recorded in quiet rooms.

Fig. 6 depicts the front-end interface, where the “legend box” is not displayed. These are English translation 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 good 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.
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 and 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: 15 learning times. 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>
<thead>
<tr>
<th>Phase</th>
<th>#Sentences</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>100</td>
<td>Tactics for TOEIC (Included with a native speaker's voice)</td>
</tr>
<tr>
<td>Test</td>
<td>20</td>
<td>ERJ in mind phoneme learning (10) in mind intonation (5) in mind accent and rhythm (5)</td>
</tr>
</tbody>
</table>

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 English Read by Japanese (ERJ) (Minematsu, et al. 2002) 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.

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).

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

<table>
<thead>
<tr>
<th>Phase</th>
<th>Score</th>
<th>Corpus</th>
<th>#male</th>
<th>#sentence</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>Pronunciation score</td>
<td>Recording</td>
<td>8</td>
<td>480</td>
<td>Same sentences at every learner</td>
</tr>
<tr>
<td></td>
<td>Intelligibility</td>
<td>Recording</td>
<td>8</td>
<td>48</td>
<td>different sentences</td>
</tr>
</tbody>
</table>
Regarding intelligibility, we transcribed 48 sentences (eight learners × three times × two sentences) selected independently from the training set. Users uttered sentences prepared \textit{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, “△” 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 and Nakagawa 2012, in Japanese), so the score is 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 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.”

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
Teacher & Correlation \\
\hline
F, the others & 0.664 \\
G, the others & 0.670 \\
H, the others & 0.595 \\
I, the others & 0.475 \\
J, the others & 0.487 \\
K, the others & 0.336 \\
\hline
Average & 0.540 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
Teacher & Correlation \\
\hline
F, the others & 0.845 \\
G, the others & 0.873 \\
H, the others & 0.731 \\
I, the others & 0.772 \\
J, the others & 0.808 \\
K, the others & 0.752 \\
\hline
Average & 0.800 \\
\hline
\end{tabular}
\end{table}

\textbf{Table 15 Scores of system (estimated score) and native teacher for every test}

The values in parentheses denote the mis-perception reduction rate.
For the pronunciation score/intelligibility, Table 16 shows the correlation coefficient for the estimated and native teacher scores.

Table 16 Correlation coefficient between automatically estimated and averaged teacher scores

<table>
<thead>
<tr>
<th>Time</th>
<th>Correlation score</th>
<th>Intelligibility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre 0.539</td>
<td>Pre 0.808</td>
</tr>
<tr>
<td>Mid</td>
<td>0.510</td>
<td>Mid 0.688</td>
</tr>
<tr>
<td>Post</td>
<td>0.476</td>
<td>Post 0.712</td>
</tr>
<tr>
<td>All</td>
<td>0.492</td>
<td>All 0.747</td>
</tr>
</tbody>
</table>

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, we carried out 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:
(QI) Was the pronunciation estimation valid?  ==>  3.13
(QII) Was the intelligibility estimation valid?  ==>  3.00
(QIII) Was this system useful for learning English pronunciation?  ==>  3.75
(QIV) Was its feedback valid?  ==>  3.25
(QV) Did the feedback improve your pronunciation?  ==>  4.00
(QVI) Did listening to a native speaker’s voice improve your pronunciation?  ==>  4.13
(QVII) Do you want to use this system for learning English pronunciation?  ==>  3.63
(QVIII) 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.

Regarding the intelligibility scores, however, the subjects did not know how to use them for measuring how much natives 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 scoring and transcription were performed, native English teachers marked the mispronounced words. We investigated the relationship among the number of mispronounced words, pronunciation score, and the number of incorrectly discriminated phoneme-pair. 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 answer 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 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 learn pronunciation. These issues remain for future study. Compared self-directed learning (e.g. repeating utterance using CDs), it is apparently beneficial to see one’s own mispronounced phonemes using this proposed system.
Table 17 Questionnaire: Whether each functions was useful.

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

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 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 implement more useful functions in our online system, in particular, graphical feedback to users that considers more interpersonal skills. Finally, based on the knowledge obtained here, we want to improve the performance of non-native speech recognition.

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 △ (velocity, time derivation) and △△
(acceleration, 2\textsuperscript{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 $W$ that has maximum posterior probability $P(W|A)$ according to Bayes’ theorem as follows:

$$W = \underset{W}{\text{arg max}} \; [W \log (P(A|W) \cdot P(W))]\ ,$$

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 word sequence. Fig. 7 shows the diagram of a general ASR system. $P(A)$ is calculated approximately by $P(A) = \underset{W}{\text{max}} \; P(A,W)$

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

![Fig. 7 Diagram of a general ASR system](image)

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 $n$ preceding words. Given word history, word-based $n$-gram predicts the current word according to the following equation:

$$w_t = \arg\max_{w_t} P(w_t | w_{t-n+1}^{t-1})$$

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.
References


Garofalo, J S, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data
Hirabayashi, K., and S. Nakagawa. Automatic evaluation of English pronunciation by Japanese
speakers using various acoustic features and pattern recognition techniques. Proc. Interspeech,
Holliday, Jeffrey J., Mary E. Beckman, and C. Mays. Did you say susi or shushi? measuring the
emergence of robust fricative contrasts in English- and Japanese-acquiring children. Proc
Itou, K, et al. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition
Karafiat, M., M. Janda, J. Cernocky, and L. Burget. Region dependent linear transforms in
Kawahara, T., and N. Minematsu. Tutorial on Computer-assisted language learning (CALL) based
Kibishi, H., and S. Nakagawa. New feature parameters for pronunciation evaluation in English
Koniaris, C., and O. Engwall. Perceptual differentiation modeling explains phoneme
Li, H,, S. Wang, J. Liang, S. Huang, and B. Xu. High performance automatic mispronunciation
detection method based on neural network and TRAP features. Proc. Interspeech,
Minematsu, N., et al. English Speech Database Read by Japanese Learners for CALL System
Development. Proc. of the International Conference on Language Resources and Evaluation
Nakagawa, S., Allen A. Reyes, H. Suzuki, and Y. Taniguchi. An English conversation and
pronunciation CAI system using speech recognition technology. Proc. Eurospeech,
Nakagawa, S., and K. Ohta. A statistical method of evaluating pronunciation proficiency for
system using speech recognition technology." (Trans. Information Processing Society in


