Practical Use of English Pronunciation System for Japanese Students in the CALL Classroom

Yasushi Tsubota, Tatsuya Kawahara, Masatake Dantsuji

Academic Center for Computing and Multimedia Studies, Kyoto University, Japan
{tsubota, kawahara, dantsuji}@media.kyoto-u.ac.jp

Abstract

We have developed an English pronunciation learning system which estimates the intelligibility of students' speech and ranks their errors from the viewpoint of improving their intelligibility.

We have begun using this system in a CALL class at Kyoto University. We have evaluated system performance through the use of questionnaires and analysis of speech data logged in the server, and will present our findings in this paper.

1. Introduction

We have developed a CALL (Computer-Assisted Language Learning) system for the practice of English speaking. In Japan, there are few lessons given on speaking – even in the classroom setting – since there are few teachers who can teach pronunciation. Moreover, it is logistically difficult even for a teacher who has the necessary knowledge and experience because teaching pronunciation is essentially a one-on-one activity and can be quite time-consuming. It is practically impossible in large classes consisting of 40 or more students.

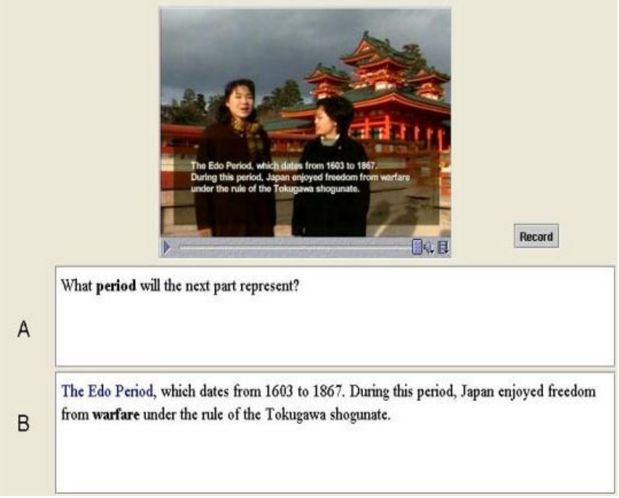
To deal with this problem, we have been conducting research on CALL systems which make use of speech recognition technology for English speaking practice. There are some systems using speech recognition technology on the market, but there are few which provide instruction and feedback on pronunciation to the user. We have developed original teaching materials for speaking [1] and pronunciation error detection technology specialized for Japanese students [2][3].

Relying on the use of pronunciation error detection technology specialized for Japanese students of English, we designed our system to estimate the intelligibility of students' speech as well as rank their errors in terms of improving their intelligibility to native speakers of English. Error diagnosis is important in self-study since students tend to spend time on aspects of pronunciation that do not noticeably affect intelligibility. For example, errors such as vowel insertion and non-reduction which are related to prosodic features, such as syllable structure and stress, are considered to be more crucial to intelligibility than purely segmental errors [4].

2. Overview of CALL System

The system covers English learning in two phases: (1) role-play conversation and (2) practice of individual pronunciation skills.

During role-play (shown in Figure 1), students play the role of a guide who provides information on famous events and/or landmarks in Kyoto. As the guide, the student (B) answers questions asked by a native English speaker (A). Each question is presented to the students in video format at the beginning of the practice session. The student records his/her spoken answers by following the script and recording prompts which appear on the screen. After the student finishes the first question, the system automatically proceeds to the next question.



The Edo Period, which dates from 1603 to 1867. During this period, Japan enjoyed freedom from warfare under the rule of the Tokugawa shogunate.

What **period** will the next part represent?

The Edo Period, which dates from 1603 to 1867. During this period, Japan enjoyed freedom from **warfare** under the rule of the Tokugawa shogunate.

Figure 1: Screenshot of role-play session

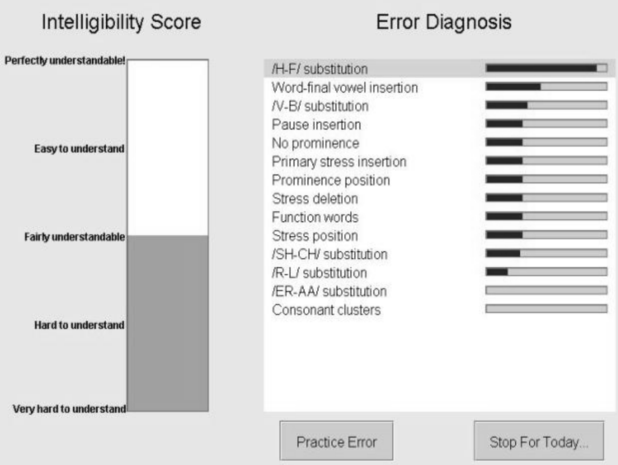


Figure 2: Screenshot of pronunciation error diagnosis

During the recording, the system works in the background to detect the student's pronunciation errors and stores a profile of his/her pronunciation skills. However, at this stage, the system does not inform the student of his/her errors because we want students to focus on the flow of the conversation. Instead, we added pronunciation models and a dictionary function for difficult words to facilitate the practice.

At the end of the role-play session, the system provides a pronunciation profile for the student. It consists of two parts: (1) an intelligibility score and (2) priority scores for the various pronunciation skills addressed. An example of the profile is shown in Figure 2. The intelligibility score is a score showing how well a student's pronunciation is understood by native speakers of English. It is computed from the error rates for each of the pronunciation errors students made. To determine the order in which the errors should be studied by a given learner, we determined the priority of each error. This value is calculated as the difference between the learner's error rate and the average error rate of students of the same intelligibility level.

In the second phase, the student practices correcting the individual pronunciation errors detected during the role-play session. The errors are categorized by type and contain the specific words or phrases which the student incorrectly pronounced during the role-play. Thus, a student is able to practice further by focusing on these words or phrases, which are a shorter form than the sentences that appeared in the conversation. During this stage, results of the error detection for the words and phrases and further instructions for correcting the errors are provided.

3. Technology of the System

3.1. Intelligibility Assessment

Intelligibility assessment based on error rates of pronunciation skills
To assess intelligibility, we adopted a probabilistic approach [5]. Given observed error rates O, our goal is to obtain the probability that the learner's intelligibility level is $i(i \in 1...5))$. This probability, noted $P(i|O)$, can be computed using Bayes formula:

$$P(i|O) \propto P(i)P(O|i) \tag{1}$$

where probability $P(i)$ is the ratio of level-i students in the considered population and $P(O|i)$ is the probability distribution of the error rates for the level-i speakers. Under the assumption that all error rates are statistically independent given the intelligibility level, the overall probability distribution is given by $P(O|i) = \Pi_j P(r_j|i)$, where $P(r_j|i)$ is the probability distribution of the j-th error rate among students of level i. We model each $P(r_j|i)$ by a Beta distribution, defined on $[0, 1]$ by:

$$\beta_{(a,b)}(x) = B(a,b)x^{(a-1)}(1-x)^{(b-1)} \tag{2}$$

where a and b are parameters and $B(a, b)$ is a normalizing constant. Parameters are computed using data rated for intelligibility by a human judge. Combining equations 1 and 2 leads to the following formula for the probability of level i:

$$P(i|O) = K \prod_j \beta_{(a,b)_{i,j}}(r_j) \tag{3}$$

where K is a normalizing constant. We define the intelligibility score as the expected value of the level:

$$I = \sum_i i \cdot P(i|S) \tag{4}$$

Thus, the score can take any value in the range $[1, 5]$.
Diagnosis of Critical Pronunciation Errors To determine which errors should be studied by a given learner, we define the priority $\pi(i, j)$ of error j at the intelligibility level i as the difference between the learner's error rate and the average error rate of level-i students, that is:

$$\pi(j,i) = r_j - \langle r_j \rangle_{level-i\ students} \tag{5}$$

The priority π_j of error j is defined as the expected value of each level's priority:

$$\pi(j) = \sum_i P(i|O) \cdot \pi(j,i) \tag{6}$$

Table 1: Examples of pronunciation errors

No counterpart in L2 syllable	t-ɪ/ʧ-ɪ	t-u/ts-ʊ	s-iː/ʃ-iː
No counterpart in L2 phoneme	l/r	b/v	s/θ
Allophone	m/n/ŋ	ʤ/ʒ	
Vowel substitution	ɔː/oʊ	iː/ɪ	uː/ʊ

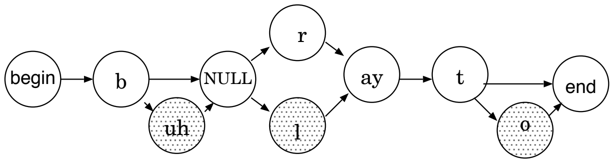


Figure 3: Pronunciation network for the word "bright"

3.2. Phoneme Error Detection

Prediction Method
To predict pronunciation errors, we modeled error patterns of Japanese students according to the linguistic literature [6]. The model includes 79 kinds of error patterns. There are 37 patterns concerning vowel insertion, such as what vowels are inserted between a certain pair of consonants or after the final consonant of words. In addition, there are 35 patterns for substitution errors. For deletion errors, we have 7 patterns, such as /w/, /y/, /hh/ deletion at word beginning and /r/ deletion in some contexts. Examples of the patterns for substitution errors are shown in Table 1. For a given practice text (orthographic transcription), the model is used to automatically generate a network as shown in Figure 3 to predict the possible error patterns. The prediction effectively guides the automatic speech recognition system to align phoneme sequences and identify erroneous phoneme segments.
Speaker Adaptation of Acoustic Model
Accurate segmentation and discrimination is not an easy task since the speech of students using this system is different from that of native speakers. To compensate for acoustic variation, we introduced speaker adaptation using Maximum Likelihood Linear Regression (MLLR). There is a problem in applying supervised adaptation in the case of a CALL system in which the students' pronunciation is not necessarily correct. Thus, we compared two transcription labels for adaptation: lexicon labels (baseform) and hand-labels for counting pronunciation errors manually. Adaptation with the lexicon labels was found to improve accuracy by about 5%, which is comparable to the result obtained using hand-labels [2].

Thus, we concluded that it is acceptable to use lexicon baseform for adaptation. In the following experiment, we used lexicon labels for the adapation.
Comparison of Acoustic Models
In order to determine the best model for automatic recognition of pronunciation errors, we compared recognition results for the native English model and Japanese students' model. We used the TIMIT database for constructing a native English model and used for an English corpus compiled from Japanese students' speech and funded by MEXT[1] for constructing a Japanese students' model.

In order to evaluate the proposed methods, we conducted phoneme recognition experiments with a corpus of English words spoken by Japanese students. The corpus[7] consists

[1] Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research on Priority Areas, No. 12040106

Table 2: Recognition rates for native English model

Type of label for training acoustic model	Baseline	Adaptation
Native English	75.4%	80.6%

Table 3: Recognition rates for Japanese students' English model

Type of label for training acoustic model	Baseline	Adaptation
Baseform	78.0%	81.8%
Automatic labeling Without adaptation	77.1%	81.5%
Automatic labeling With adaptation	78.0%	81.5%

of 5950 speech samples. Seven Japanese speakers (2 male, 5 female) each uttered 850 basic English words. The database contains phonemic hand-labels, which were transcribed faithfully and includes labels for the erroneous phonemes. We confirmed the effect of speaker adaptation in the evaluation phase, as shown in both Table 2 and 3. These techniques were found to significantly improve accuracy in all cases (compare left to right). But they are not as effective in generating training labels (compare top to bottom) in Table 3. As we expected, the best acoustic model is the one trained with the Japanese students' database. This model yields 3% better accuracy than the native English model without speaker adaptation (baseline). The superiority decreased to 2% when speaker adaptation was applied. The results demonstrate that with speaker adaptation, the Japanese student's model can compete with the native English model. Thus, we decided to use the Japanese students' model for the system.

3.3. Stress Detection

Modeling of Sentence Stress by Japanese Students

In English, stressed syllables are characterized by not only power level, but also pitch, duration and vowel quality [8]. However, pitch in natural conversation rises rapidly at the beginning of each phrase unit and falls gradually, resulting in complex influences on sentence stress.

We analyzed the causes of pronunciation errors by Japanese students, and based on the results [3],we present three classes of stressed syllables. Their combinations yield different models.

I. Classification by stress (base)

We divide the syllables in a given sentence into three categories. Primary-stressed syllables (PS) are syllables that carry the major pitch change in a tonal group (phrase). Hence, there is only one PS in each phrase, usually placed on the word containing the most important piece of information. Secondary-stressed syllables (SS) are all other stressed syllables. Finally, non-stressed syllables (NS) are syllables that do not bear any mark of stress.

II. Classification by syllable structure (syl)

Syllable structure and stress are correlated such that complex structures have a larger probability of being stressed [9]. We classify syllables into four categories: V, CV, VC, CVC. We also classify vowels into four categories: schwa (Vx), short vowel (Vs), long vowel (Vl), and diphthong (Vd). Thus, combinations of these two factors give rise to 16 possible categories of syllables.

III. Classification by position in phrase (pos)

Since pitch movement behaves differently at the beginning and end of a phrase, the resulting prosody pattern also differs depending on the position of the syllable in the phrase. Thus, we also classify syllables into three types according to their position in a phrase: head (H), middle (M), and tail (T).

Based on the above classification, we set up models for three stress categories, sixteen syllable structures and three phrasal positions. Thus, in the most complicated case, there are 144 (=3*16*3) stress models.

Automatic Detection of Sentence Stress

In order to reliably align the syllable sequence which includes the phoneme insertions and substitutions by non-native speakers, we make use of a speech recognition system with error prediction for a given sentence as described in 3.2. Based on this alignment, the syllable units and their structures and positions within phrase units are determined. For each syllable, NS, PS and SS models are applied to determine the stress level. Linguistic studies suggest that all syllables but one in a word tend to be un-stressed in continuously spoken sentences [10]. Hence, we constrain the number of PS to one per phrase unit. The most probable stress (syllable) sequence is obtained by matching the aligned phrase segment. Syllables whose detected stress level differs from the correct level are labeled as pronunciation errors. If the syllable structure and/or the position in the phrase are incorrect, such information is presented to the student as possible causes of the stress error.

Since PS, SS and NS have different acoustic characteristics, the primary features for discrimination will differ according to the stress level. For example, PS is characterized by a tonal change; thus F0 should be the most important feature for discrimination. We propose a two-stage recognition method that applies different weights. During the first stage, the presence of stress is detected. Here, a stress model (ST) that merges PS and SS syllables is compared against NS using weights optimized for the two-class discrimination. For syllables detected as stressed, the stress level (PS or SS) is recognized during the second stage using different weights.

4. Actual Use in the Classroom

4.1. CALL Class at Kyoto University

We have begun using this system in an English class for second-year students of Kyoto University. The syllabus for the class is as follows.

I. Comprehension of the first topic (covered in 3 classes)

In the classroom, we use original multimedia CD-ROM teaching materials to provide training on grammar and vocabulary. The skits and lessons are based on the Jidai Festival (Festival of Ages), one of the three most famous festivals in Kyoto.

II. Role-Play (Using CALL pronunciation learning system)

After 15 minutes of instruction on how to use the system, students use the system freely for 60 minutes.

III. Comprehension of the second topic (covered in 3 classes)

The Jidai Festival consists of several processions representing different periods in Japanese history. The second topic covers a procession of the Jidai Festival featuring people dressed in costumes of the Edo period. Thus, students are given the opportunity for a more in-depth look at the Jidai Festival.

IV. Role-Play (Using CALL pronunciation learning system

Students practice pronunciation through role-play in the same manner as described in II above, focusing on the Edo Period.

4.2. Analysis of logged data

When we first used this system in the classroom, unexpected problems arose. We divided these problems into 4 categories.

I. Errors during recording

To cope with mistakes at the start of recording, we designed the system to deliver a pop-up dialogue message to indicate a recording error when there is a long period of silence. This error occurred 16 times on average during the first classroom trial of the system. We determined this was caused by improper configuration of recording levels. Thus, during the second trial of the system, we instructed students to set their recording levels prior to recording, and as a result reduced the number of errors by 75%.

II. Errors related to automatic stop of recording

In order for students to maintain concentration during role-play, we designed the system to automatically stop recording when there is a long period of silence after an utterance. However, in the initial trial, the system sometimes stopped recording in the middle of an utterance or did not stop recording even after a period of silence. We found this error is also caused by improper configuration of recording levels. After making the appropriate adjustments, this error also decreased in the second trial.

III. Unpredicted pronunciation errors

This system is designed to predict the possible pronunciation errors for a given sentence before a student actually pronounces the sentence. However, students make a lot of unexpected pronunciation errors. Most of them involve repetition of words and/or reading the sentence incorrectly. For example, some students uttered 1607 (sixteen-o-seven) although the correct phrase is “1603 (sixteen-o-three)”. In other cases, students uttered “sixteen three” for “1603 (sixteen-o-three)”. These errors occurred because the students were not familiar with these words. A possible solution is to add error candidates. However, this method inevitably degrades the accuracy of error detection. A better option would be to simply add an explanation for the reading of the phrase in question and a function for re-recording.

IV. Recognition Errors

The system delivers a message indicating recognition error when the utterance differs greatly from the corresponding model. While 755 errors of this type were observed during the first trial, the number of errors decreased to 176 in the second trial after students were instructed to properly set their recording levels.

4.3. Evaluation by class participants

There are several pronunciation learning systems using speech recognition systems on the market. But there are few systems which provide diagnostic evaluations and instructions on pronunciation interactively. We have received numerous positive opinions regarding this brand-new approach. For example, a student who used this system remarked, “It’s very interesting as I haven’t experienced this kind of English practice. I want to practice more with this system.” Another one commented, “Other classes don’t offer the opportunity to use interesting systems like this one.” On the other hand, some students voiced complaints regarding improper configuration of the microphone settings.

We also counted the number of utterances students made and the number of errors made using the logged data, and compared the results for the two trials. As shown in Table 4, the number of utterances more than doubled on average from 52.1 to 111 and the number of errors dramatically decreased from 20.4 to 4.9 for recording errors and from 1.24 to 0 for recognition errors. We also asked students to evaluate and score the system on a scale of 0-100. As Table 5 shows, more than half the class gave the system a score of more than 70; the mode value was between 81-90. Thus, we can conclude that it is likely that the students were highly satisfied with the system.

Table 4: Analysis of logged data

	1st trial	2nd trial
#Utterances	52.1 (Avg.) 1929 (Total)	111 (Avg.) 3982 (Total)
Error Rate (Recording)	20.4 (Avg.) 755 (Total)	4.9 (Avg.) 176 (Total)
Error Rate (Recognition)	1.24 (Avg.) 46 (Total)	0 (Avg.) 0 (Total)

Table 5: Evaluation by class participants

Score	<50	51-60	61-70	71-80	81-90	91-100
#Students	2	2	8	11	13	4

5. Summary

We have addressed a CALL system which estimates the intelligibility of Japanese students’ speech and ranks their errors in terms of improving their intelligibility to native English speakers. To construct this system, we introduced (1) automatic intelligiblity assessment, (2) phoneme error detection and (3) stress error detection. To estimate intelligibility, we modeled the relationship between error rates for each type of pronunciation error and intelligibility.

We have begun using our CALL system for speaking practice in an actual CALL classroom. The number of recording and recognition errors during the first trial of the system was large due to improper configuration of the headset microphones. After checking the microphone settings, performance dramatically improved. Evaluation of the system by the class was quite positive.

6. References

[1] M. Shimizu et’al A Model of Multimedia-Based English CALL Contents for Japanese Students. In *Proc. World Multiconference of Systemics, Cybernetics and Informatics*, 2002.

[2] Y. Tsubota et’al. Recognition and verification of english by japanese students for computer-assisted language learning system. In *Proc. ICSLP*, pages 1205–1208, 2002.

[3] K Imoto et’al. Modeling and automatic detection of english sentence stress for computer-assisted english prosody learning system. In *Proc. ICSLP*, pages 749–752, 2002.

[4] Celce-Murcia M, D. et’al. *A Reference for Teachers of English to Speakers of Other Languages*. CUP, 1996.

[5] A. Raux et’al. Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning. In *Proc. ICSLP*, pages 737–740, 2002.

[6] S Kohmoto. *Applied English Phonology: Teaching of English Pronunciation To the Native Japanese Speaker*. Tanaka Press, 1965.

[7] K. Tanaka et’al. Acoustic Models of Language-Independent Phonetic Code Systems for Speech Processing. In *Proc. of Spring meeting of the Acoustical Society of Japan*, 2001.

[8] M. Sugito. *English spoken by Japanese*. Izumishoin, 1996.

[9] Dauer. R.M. Stress-Timing and Syllable Timing Reanalyzed. In *Journal of Phonetics*, pages 51–62, 2001.

[10] K. Watanabe. *Instruction of English Rhythm and Intonation*. Taishukanshoin, 1994.