Practical Use of Autonomous English Pronunciation Learning System for Japanese Students

Yasushi TSUBOTA **Masatake DANTSUJI** **Tatsuya KAWAHARA**

Academic Center for Computing and Media Studies, Kyoto University,

Yoshida Nihonmatsu-cho, Kyoto, Japan

{tsubota, dantsuji, kawahara}@media.kyoto-u.ac.jp

Abstract

We have developed an original CALL system designed to detect and diagnose English pronunciation errors in Japanese learners' speech. We have begun using the system in an English class at Kyoto University in Kyoto, Japan. The class consists of approximately 40 students, and the students have practiced pronunciation using this system on four separate occasions. Sixty minutes was allocated for each practice session, providing a total of 9600 minutes of practice time per academic calender year. We have examined the recorded speech data and corresponding speech recognition results obtained for the first semester of the preceeding academic year.

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 English speaking practice in (Shimizu and Dantsuji, 02), as well as pronunciation error detection technologies specialized for Japanese students in (Tsubota, Kawahara and Dantsuji, 02) and (Imoto, Tsubota, Raux, Kawahara and Dantsuji, 02).

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 autonomous learning since students tend to spend time on aspects of pronunciation that do not noticeably affect intelligibility (Raux and Kawahara, 02). 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 (Celce-Murcia, Brinton, and Goodwin, 96).

2 Autonomous English Pronunciation Learning System for Japanese Students

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.

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.

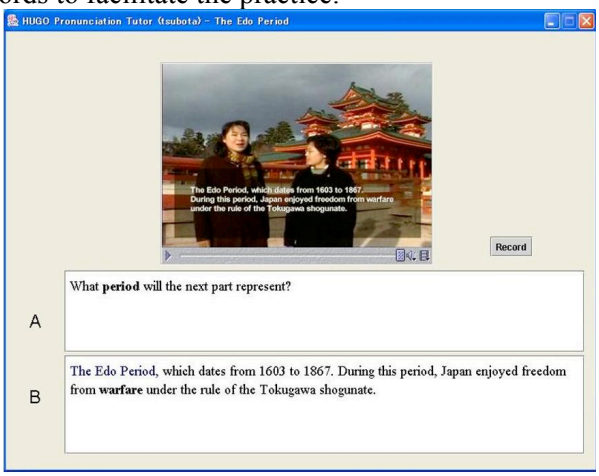

Figure 1 Screenshot of role-play session

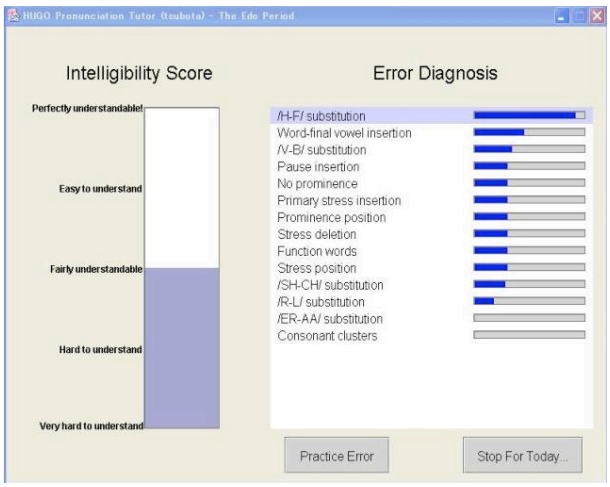

Figure 2 Screenshot of pronunciation error diagnosis

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 (Raux et'al, 02).

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. An example of the practice for correcting the individual pronunciation errors is shown in Figure 3.

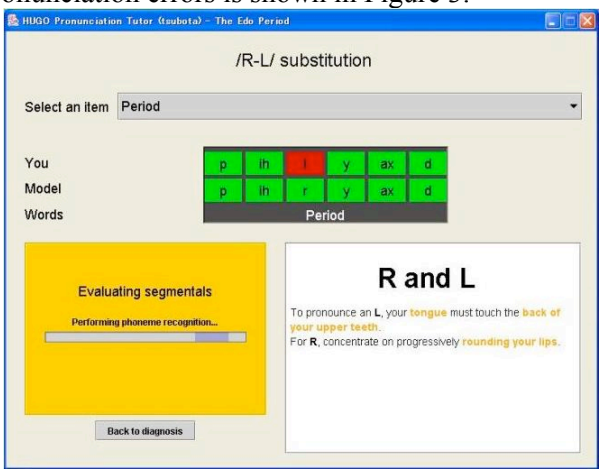

Figure 3 Screenshot of practice for correcting the individual pronunciation errors

The targets used in the detection of pronunciation errors by our system were decided based on (Kohmoto, 65). Examples of the targets are described in Table 1. We also determined error patterns concerning stress based on (Imoto, Raux, Tsubota, Kawahara and Dantsuji, 02).

In order to detect the pronunciation errors, we use a speech recognition system. For phoneme error detection, we use an acoustic model built from the corpus of English spoken by Japanese students based on the results in (Tsubota et'al, 02). The error patterns listed in Table 1 are applied to the orthographic transcriptions in order to obtain the possible error candidates in Japanese students' speech. (Note: speaker adaptation was not applied in the system.)

For stress error detection, we use an acoustic model built from the TIMIT database (Garofolo, Lamel, Fisher, Fiscus, Pallett and Dahlgren, 93) and based on the results in (Imoto et'al, 02). The acoustic features for the model are as follows: MFCC(4), log(F0), log(Power) and its delta and acceleration. To cope with the changes in the number of syllables due to vowel insertions, a common error in Japanese students' speech, we utilize information on vowel insertions based on the results of the phoneme error detection described above.

Table 1. Examples of typical pronunciation errors by Japanese students

Abbv. of error	Description	Example	Erroneous pronunciation
WY	Word-initial w/y deletion	***w***ould	uw d
SH	SH/CH substitution	***ch***oose	***sh*** uw z
ER	ER/A substitution	pap***er***	p ey p ***ah***
RL	R/L substitution	***r***oad	***l*** ow d
VR	Vowel non-reduction	stud***e***nt	s t uw d ***eh*** n t
VB	V/B substitution	pro***b***lem	p r aa ***v*** l ax m
FI	Word-final vowel insertion	le***t***	l eh t ***ao***
CCV	CCV-cluster vowel insertion	***st***udy	***s uh t*** ah d ih
VCC	VCC-cluster vowel insertion	a***ct***ive	ae ***k uh t*** ih v
HF	H/F substitution	***f***ire	***hh*** ay er

Note: Phonemic descriptions in Table 1 based on TIMIT DATABASE (Garofolo et'al, 93).

3 Actual Use in the Classroom

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.

3.1 Analysis of Logged Data

After listening to the speech data collected in the classroom, we confirmed that some of the recorded speech has noise or hesitation. As these kinds of speech data degrade the accuracy of pronunciation error detection, a feature that prompts users to re-record is desired. To make this possible, we are conducting a thorough analysis of the speech data, which entails a comparison of the degree of alignment based on speech recognition with sound spectrograms of the students' speech. We have successfully defined the alignment errors and categorized them as follows (percentage and number of errors/total number of utterances given in brackets). Note some of the speech data are categorized into multiple categories.

1) Errors in automatic detection of the end of a recording session [6.0%, 116 /1929]

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, which inevitably causes misalignment. We found this error is often caused by improper configuration of recording levels.

2) Addition of noise [13.1%, 252/1929]

Some of the speech data has white noise throughout the utterance due to improper microphone settings or broken microphone devices. Moreover, while some of the speech data contains noise, other data do not contain noise even though they were recorded with the same machine. A possible solution for these types of errors is to provide a feature that automatically checks the configuration of the microphone settings or verify whether or not the speech data has the characteristics of white noise.

3) Hesitation [4.2%, 81/1929], Speech errors [1.8%, 34/1929]

Our 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 (hesitation) and/or reading the

sentence incorrectly (speech errors). 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.

4) Misalignment by the speech recognition system [12.8%, 246/1929]

Even without the errors above, some of the speech data are misaligned by the speech recognition system. A typical example is the noise model. The noise model is optionally inserted at the beginning and end of a sentence and between words to deal with possible noises, which include all sounds that are not in the target speech, such as the sound produced when touching the microphone, coughing or repetition of words. While this noise model works well in most cases, it is sometimes matched with the words or phonemes at the beginning of a sentence, such as "uh yes," the word "the," or phonemes at the end of a sentence such as "t" or "z".

5) Recognition errors [1.5%, 29/1929]

To enable interruption during the role-play session, the system is designed to recognize the sentence "I'd like to stop now", at any time during the session. When this sentence is recognized, the role-play session ends and the pronunciation error diagnosis appears. In some cases, however, an utterance is recognized incorrectly as this sentence, and the role play session is unwillingly ended. For instance, the utterance "Sure" is often recognized as this sentence.

3.2 Improvement in the 2nd Trial

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. Based on the analysis above, we instructed students to set their recording levels prior to recording during the second trial of the system, and as a result reduced the number of errors by 75%.

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 2, 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.

Table 2. Comparison of recording and recognition errors for 1st and 2nd trials

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

4 Summary

We have begun using our CALL system for speaking practice in an actual CALL classroom. Speech data in the first trial were analyzed using spectrograms, and the errors were categorized into five categories. Improper configuration of the headset microphone, a cause of three-quarters of the errors, was solved by instructing students in advance to properly configure their settings in the second trial. As a result, the number of recording and recognition errors dramatically decreased.

References

Shimizu, M. and Dantsuji, M. (2002) A Model of Multimedia-Based English CALL Contents for Japanese Students. In Proc. World Multiconference of Systemics, Cybernetics and Informatics.

Tsubota, Y., Kawahara, T. and Dantsuji M., (2002) Recognition and verification of English by Japanese students for Computer-Assisted Language Learning System. In Proc. ICSLP, pages 749-752.

Imoto, K., Raux, A., Tsubota, Y., Kawahara, T. and Dantsuji, M. (2002) Modelling and automatic detection of English sentence stress for computer-assisted English prosody learning system. In Proc. ICSLP, pages 737-740.

Raux, A. and Kawahara, T. (2002) Automatic Intelligibility Assessment and Diagnosis of Critical Pronunciation Errors for Computer-Assisted Pronunciation Learning. In Proc. ICSLP, pages 737-740.

Celce-Murcia, M., Brinton, D. M. and Goodwin, J. M. (1996) A Reference for Teachers of English to Speakers of Other Languages, CUP.

Kohmoto, S. (1965) Applied English Phonology: Teaching of English Pronunciation to the Native Japanese Speaker. Tanaka Press.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L. (1986) The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. Technical Report NISTIR 4930, National Institute of Standards and Technology.