ReCALL **16** (1): 173–188. © 2004 Cambridge University Press
DOI: 10.1017/S0958344004001314 Printed in the United Kingdom

An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors

YASUSHI TSUBOTA, MASATAKE DANTSUJI and TATSUYA KAWAHARA
Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan
(*e-mail: {tsubota, dantsuji, kawahara}@media.kyoto-u.ac.jp*)

Abstract

We have developed an English pronunciation learning system which estimates the intelligibility of Japanese learners' speech and ranks their errors from the viewpoint of improving their intelligibility to native speakers. Error diagnosis is particularly important in self-study since students tend to spend time on aspects of pronunciation that do not noticeably affect intelligibility. As a preliminary experiment, the speech of seven Japanese students was scored from 1 (hardly intelligible) to 5 (perfectly intelligible) by a linguistic expert. We also computed their error rates for each skill. We found that each intelligibility level is characterized by its distribution of error rates. Thus, we modeled each intelligibility level in accordance with its error rate. Error priority was calculated by comparing students' error rate distributions with that of the corresponding model for each intelligibility level. As non-native speech is acoustically broader than the speech of native speakers, we developed an acoustic model to perform automatic error detection using speech data obtained from Japanese students. As for supra-segmental error detection, we categorized errors frequently made by Japanese students and developed a separate acoustic model for that type of error detection. Pronunciation learning using this system involves two phases. In the first phase, students experience virtual conversation through video clips. They receive an error profile based on pronunciation errors detected during the conversation. Using the profile, students are able to grasp characteristic tendencies in their pronunciation errors which in effect lower their intelligibility. In the second phase, students practise correcting their individual errors using words and short phrases. They then receive information regarding the errors detected during this round of practice and instructions for correcting the errors. 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 forty 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. In order to construct an effective CALL system for speaking practice, we have been working toward the development of original teaching materials designed to improve speaking as well as the development of error detection technology.

We use multimedia English teaching materials developed in our laboratory and available on CD-ROM for the content of the speaking practice (Shimizu & Dantsuji, 2002). Their primary aim is to help Japanese students improve their English skills – particularly in terms of speaking and writing – by fostering the ability to explain a variety of topics on Japanese history and culture in English to visitors from abroad. For example, one of the topics in the CD-ROM series focuses on the Jidai Festival (Festival of Ages), one of the three most famous festivals in Kyoto, the former capital of Japan. In this festival, the people of Kyoto conduct processions around the city in order to recreate images of the old capital. The processions consist of groups representing the different periods in Japanese history and the people in each procession are dressed in costumes from that period.

In order to construct an effective CALL system for speaking practice, one must have the technology for accurate error detection. In developing our system, we analyzed the error tendencies in Japanese students' pronunciation and developed statistical models from a large amount of speech corpus data.

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 (Celce-Murcia & Brinton, 1996).

We have begun using this system in a CALL class at Kyoto University, and are evaluating the system's performance through the use of questionnaires and further analysis of speech data logged in the server.

This paper consists of the following sections: Overview of CALL System, Technology of the System, Actual Use in the Classroom.

2 Overview of CALL system

An overview of our CALL system is depicted in Figure 1. 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 2), 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

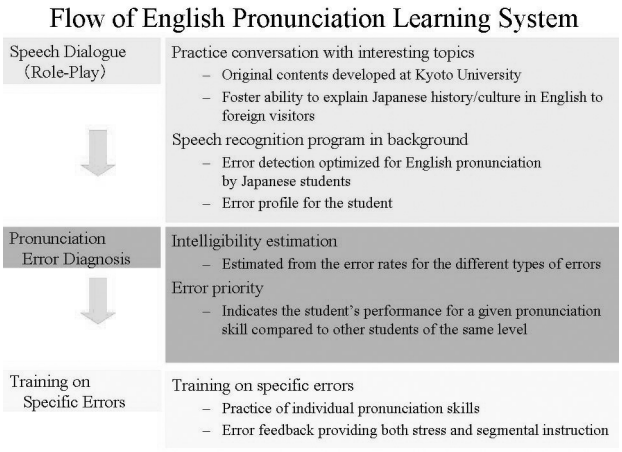

Fig. 1. System overview.

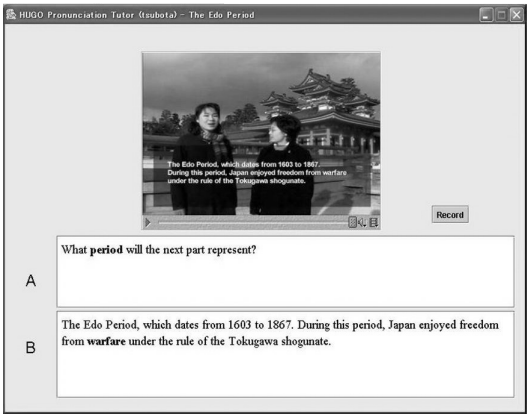

Fig. 2. Example of practising role-play.

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.

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 3.

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 the student made. To determine the order in which the errors

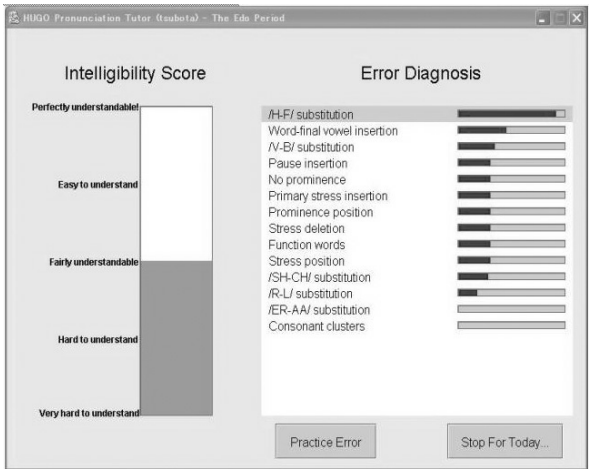

Fig. 3. Example of pronunciation profile.

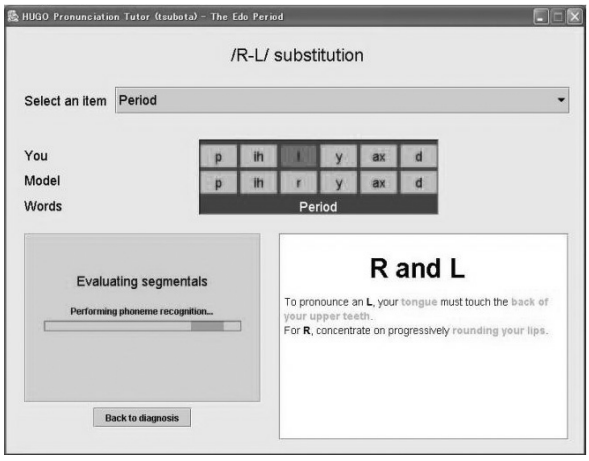

Fig. 4. Example of practising specific pronunciation skills

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. As for the average error rates for each intelligibility level, we adopted the values computed in preliminary experiments (described in section 3.1.1).

In the second phase, the student practises 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 practise 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 is shown in Figure 4.

3 Technology of the system

3.1 Automatic intelligibility assessment

3.1.1 Preliminary experiments

We conducted an experimental study of the relationship between ten selected pronunciation errors common among Japanese speakers of English and a linguistic expert's rating of intelligibility. The list of errors is given in Table 1. There were sixteen subjects, all of whom were students, faculty or staff members from Kyoto University. We recorded their reading of a passage designed for pronunciation evaluation (Prator, 1985). The recordings were sent to a qualified linguist who rated each subject's intelligibility from 1 (hardly intelligible) to 5 (perfectly intelligible). Pronunciation errors in the recordings were then detected using Automatic Speech Recognition System (ASR), and each subject's error rates were computed. The error rate of each pronunciation error is the percentage of that error in relation to all of the possible errors in the entire conversation. We computed the average error rates of subjects for each intelligibility level. Figure 5 shows the error rates for the five levels. As for ASR, we used the HTK Toolkit developed by Cambridge University (Young & Woodland, 1996) and originally-developed acoustic model and language models, which are explained in 3.2.

Figure 5 shows that the manner in which error rates vary across levels differs depending on the error. Students of different levels are grouped according to their performance in each of the different error categories. Three types of errors can be distinguished as follows.

I. *Phonemic substitution and deletion*
For these errors (number 1 to 4 in Table 1), only level 5 students have a relatively low (10–40%) error rate. The other students have error rates approximately 10 to 50% greater depending on the error. The error rates among level 1 to level 4 students are similar.

Table 1. *Example of vowel insertion errors (Note: Phonemic descriptions in Table 1 based on TIMIT DATABASE (Garofolo, Lamel, Fisher, Fiscus, Pallett & Dahlgren, 1993)*

Error number (abbv. of error)	Description	Example	Erroneous pronunciation
1 (WY)	Word-initial w/y deletion	***w***ould	uw d
2 (SH)	SH/CH substitution	***ch***oose	***sh*** uw z
3 (ER)	ER/A substitution	pap***er***	p ey p ***ah***
4 (RL)	R/L substitution	***r***oad	l ow d
5 (VR)	Vowel non-reduction	stud***e***nt	s t uw d ***eh*** n t
6 (VB)	V/B substitution	pro***b***lem	p r aa ***v*** l ax m
7 (FI)	Word-final vowel insertion	le***t***	l eh t ***ao***
8 (CCV)	CCV-cluster vowel insertion	s***t***udy	***s uh t*** ah d ih
9 (VCC)	VCC-cluster vowel insertion	act***i***ve	ae ***k uh t*** ih v
10 (HF)	H/F substitution	***F***ire	***hh*** ay er

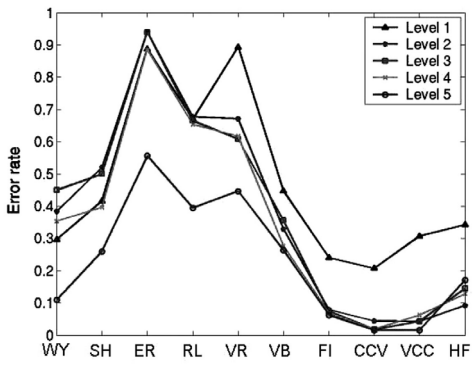

Fig. 5. Average error rates for each intelligibility level.

II. *Vowel non-reduction*
The error rates for error number 5 divide the students into three groups: level 1 students have an error rate of 90%, level 2 to level 4 students have an error rate of 60–70%, and level 5 students have an error rate of 40%.

III. *Vowel insertion, H/F, V/B substitutions*
These errors (number 6 to 10) pertain to syllable structure and two consonant contrasts (/v-b/ and /h-f/). Level 1 students have error rates 20% larger than the other students. Level 2 to level 5 students have similarly low error rates.

These observations show that such aspects of pronunciation as consonant clusters (affected by vowel insertion) or vowel reduction need to be mastered in order to reach even average levels of intelligibility. On the contrary, phonemic substitutions and deletions, except for the two pairs H/F and V/B, do not prevent speakers from being largely intelligible, since even largely intelligible speakers (level 4) have high error rates. These results are consistent with the position of most recent linguists regarding the teaching of pronunciation (Celce-Murcia *et al*, 1996). 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.

3.1.2 Intelligibility assessment

Based on the findings of the preliminary study, we propose a probabilistic approach to intelligibility assessment. 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, which is noted $P(i \mid O)$, can be computed using Bayes formula:

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

where probability $P(i)$ is the ratio of level-i students in the considered population and $P(O \mid i)$ 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 $P(O \mid i) = \Pi_j P(r_j \mid i)$, where $P(r_j \mid i)$ is the probability distribution of the jth error rate among students of level i. We model each $P(r_j \mid i)$ 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 linguistic expert. Combining equations (1) and (2) leads to the following formula for the probability of level i:

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

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

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

Thus, the score can take any value in the range [1, 5].

3.1.2 Diagnosis of critical pronunciation errors

To determine which errors should be studied by a given learner, we define the priority $\pi(j, i)$ 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:

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

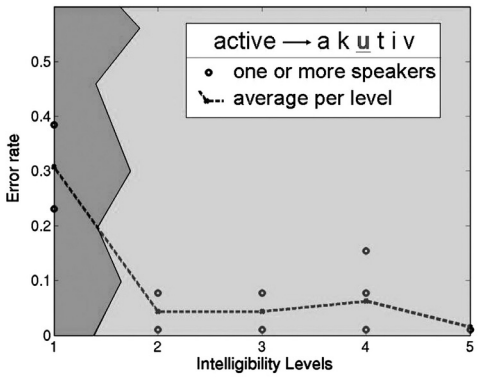

Fig. 6. Average error rates and priority for VCC vowel insertion.

Figure 6 gives an illustration of how error priority is affected by the overall intelligibility of the student. In this example, we consider a student whose error rate for error 9 (vowel insertion in VCC clusters) is 15%. Circles represent students from the training samples and the dashed line connects the average value for each intelligibility level. Priority is negative for level 1 (the student's rate is below the average) and positive for all other levels. Thus, this error is likely to be proposed for study if the learner's intelligibility level is 2 or more, because speakers of these levels usually master this error. On the other hand, if the learner's level is 1, other errors are more likely to be proposed to improve intelligibility. We computed error priorities for the sixteen speakers of the preliminary experiment and found that they agreed with subjective judgements of the strength and weaknesses of each speaker.

3.2 Phoneme error detection

3-2-1 Prediction method

To predict pronunciation errors, we modeled error patterns of Japanese students according to the linguistic literature (Kohmoto, 1965). 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, /w/, /y/, /hh/ deletion at word beginning and /r/ deletion in some contexts. Examples of the patterns for insertion and substitution errors are shown in Tables 2 and 3, respectively. The Parentheses [] in Table 2 indicate the positions of the vowel insertion for the different errors, and the phonemes following the equal sign indicate what

Table 2. *Example of vowel insertion errors initial cluster (Note: Phonemic descriptions in Table 2 based on TIMIT DATABASE (Garofolo et al., 1993))*

Initial Cluster (CCV)	In the case of p [] (l/r) [] = uw/uh	In the case of t [] r [] = ow/ao	In the case of b [] (l/r) [] = uw/uh
Final cluster (CCV)	In the case of p [] (t/th/s) [] = uw/uh	In the case of k [] (t/th/s) [] = uw/uh	In the case of b [] (d/z) [] = uw/uh
Final consonant	In the case of s [] [] = ow/ao	In the case of d [] [] = uw/uh	In the case of k [] [] = uw/uh

Table 3. *Examples of substitution errors (Note: Phonemic descriptions in Table 3 based on TIMIT DATABASE (Garofolo et al., 1993)*

No counterpart in L2 syllable	t-ih/ts-ih	t-uh/ ts-uh	s-ih/sh-iy
No counterpart in L2 phoneme	l/r	b/v	s/th
Allophone	m/n/ŋ	dz/zh	
Vowel substitution	ao/ow	iy/ih	uw/uh

vowels are inserted.

For a given practice text (orthographic transcription), the model is used to automatically generate a network as shown in Figure 7 to predict the possible error patterns.

The prediction effectively guides the automatic speech recognition system to align phoneme sequences and identify erroneous phoneme segments.

3.2.2 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) (Leggetter & Woodland, 1995). 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. From a corpus of 850 basic English words, 100 word samples were used for adaptation and the remaining 750 samples for evaluation. The baseline acoustic model was trained with the TIMIT database (Garofolo *et al.*, 1993). We set up monophone hidden markov models (HMMs) for 41 English phonemes. Each HMM has three states and 16 mixture components per state. Phoneme recognition rates based on adaptation type are listed in Table 4. Adaptation with the lexicon labels was found to improve accuracy by about 5%, which is comparable to the result obtained using hand-labels. Thus, we judged it acceptable to use lexicon baseform for adaptation in the following experiments.

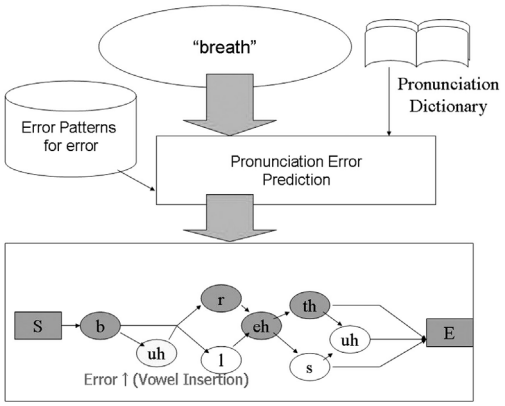

Fig. 7. Pronunciation network for the word "breath".

Table 4. *Phoneme recognition rate based on adaptation type*

Model	No adaptation	Lexicon label	Hand-label
Native English	75.4%	80.6%	81.0%

3.2.3 Comparison of acoustic models

Next we describe acoustic modeling for automatic recognition of pronunciation errors. For evaluation of the proposed methods, we conducted phoneme recognition experiments with a corpus of English words spoken by Japanese students. The corpus (Tanaka, Kojima, Tomiyama & Dantsuji, 2001) consists of 5950 speech samples. Seven Japanese speakers (two male, five female) each uttered 850 basic English words. The database contains phonemic hand-labels, which were transcribed faithfully and include labels for the erroneous phonemes.

Speech data were sampled at 16 kHz and 16 bit. Twelfth-order mel-frequency cepstral coefficients were computed every 10 ms. Temporal differences of the coefficients and power were also incorporated.

Native English model

We constructed a native English model from the TIMIT database, as our baseline acoustic model. The database was collected from 8 major dialect regions of the United States. It contains a total of 6,300 sentences (ten sentences spoken by 630 speakers). Acoustic analysis conditions are the same as described above.

Training with Japanese students' speech

To improve the baseline model, we explored the use of speech data spoken by Japanese students. We used an English corpus compiled from Japanese students' speech and funded by MEXT[1]. The corpus contains a total of 13,129 sentences spoken by 178 Japanese speakers (85 male, 93 female). Although the corpus contained a large number of pronunciation errors, it did not contain phonemic labels. Thus, we set up two kinds of phonemic labels for comparison: labels from baseform and automatic labeling with speech recognition. With automatic labeling of the training data, we can also make use of speaker adaptation. Specifically, we applied two kinds of automatic labeling, one with speaker adaptation and one without.

Table 5 lists the phoneme recognition results based on the various acoustic models. In

Table 5. *Phoneme recognition rates for several types of acoustic modeling*

Model		Baseline	Adaptation
Native English		75.4%	80.6%
Japanese students' English			
	Baseform	78.0%	81.8%
	Automatic labeling without adaptation	77.1%	81.5%
	Automatic labeling with adaptation	78.0%	81.5%

1 Ministry of Education, Culture, Sports, Science and Tecnhology, Grant-in-aid for scientific research on priority areas, No. 12040106

the evaluation of test data, we applied speaker adaptation. Thus, two kinds of results for each model were computed: baseline and adaptation.

We confirmed the effect of speaker adaptation in the evaluation phase. 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).

As we expected, the best acoustic model is the one trained with the Japanese students' database. This model yielded 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.

3.3 Stress detection

3.3.1 Analysis and modeling of sentence stress by Japanese students

In isolated words, the position of stressed syllables is fixed. In the case of sentence utterances, stress patterns are affected by the context of component words. Typically, content words such as verbs and nouns are stressed, whereas function words such as articles and prepositions are not (Watanabe, 1994). But many other factors affect sentence stress and several patterns are often possible. In a CALL system, the use of a predetermined skit allows for the prediction of a correct stress pattern.

In English, stressed syllables are characterized by not only power level, but also pitch, duration and vowel quality (Sugito, 1996). 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 as follows:

I. *Difference between stress and pitch accent*
   In Japanese, important words are characterized by a change in pitch. Thus, Japanese students tend to mark stressed syllables using only pitch instead of the entire set of acoustic features that characterize stress in English. This results in perceived errors.
II. *Incorrect syllable structure*
   English has a large number of possible syllable structures, ranging from a single vowel (V) to syllables including as many as seven consonants. By contrast, Japanese syllables are essentially limited to V and CV. As a consequence, Japanese students often insert unnecessary vowels when pronouncing English syllables whose structures do not exist in Japanese. This deformation of syllable structure can also lead to stress errors. For example, Japanese students tend to pronounce the word "strike" as /s-u-t-o-r-ay-k-u/ and place stress on the added vowels instead of the main vowel /ay/.
III. *Improper phrase units*
   When pronouncing complex sentences, non-native speakers may divide them into phrase units that do not match the sentences' syntactic structure. Changes in pitch as a result of improper division lead to stressing of syllables at inappropriate positions.

Based on these observations, 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 (Dauer, 1983). 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 sixteen 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 the three stress categories, sixteen syllable structures and three phrasal positions. Thus, in the most complicated case, there are 144 (= 3*16*3) possible stress models.

3.3.2 Automatic detection of sentence stress

We used the following acoustic features used in the detection of sentence stress: pitch (log(F0)), power (log(power)) and spectral (MFCC: Mel-Frequency Cepstral Coefficients) parameters. Preliminary experiments showed that addition of derivatives and accelerations of these features improves performance. Thus, we used an 18-dimen-

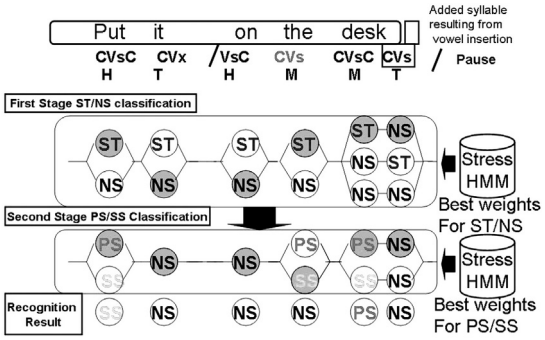

Fig. 8. Two-stage sentence stress recognition.

sional acoustic feature vector. Pitch, power and spectral features can be regarded as independent, and are thus processed as three different streams in the model. Preliminary experiments also showed that modeling the distribution with a mixture of eight Gaussians brought about the best result.

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 (Watanabe, 1994). 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 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 as shown in Figure 8. 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 used Java for windows OS to implement the system, and have installed it in a CALL classroom at Kyoto University. In this particular CALL classroom, there are 48 PCs, each equipped with a headset and microphone. Each PC has a 1.5 GHz Pentium IV CPU and 256 mega-byte memory.

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 three classes)*
   We use multimedia CD-ROM teaching materials, described in the introduction, which 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 three 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 practise 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 four 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 sixteen times on average during the first classroom trial of the system. We determined that 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 students actually pronounce 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, some students uttered "sixteen three" for "1603 (sixteen-o-three)". These errors occurred because the students were not familiar with

Table 6. *Analysis of logged data*

	#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)

Table 7. *Evaluation by class participants*

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

these words. A possible solution to this problem 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 of the 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 practise 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 6, 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 7 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.

5 Conclusion

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 intelligibility 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, and confirmed that the results agree with the position of most recent linguists. For phoneme error detection and stress error detection, we

examined the error tendencies of Japanese students and developed an acoustic model and a method for error prediction to detect these errors accurately.

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 largely 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.

Acknowledgements

This system was put into practical use through the great effort and help of many people. I would like to express my thanks to the following people: Masaaki Shimizu, Shikiko Kawakami and Takeshi Sengiku for helping to run this system in the classroom, Raux Antoine and Kazunori Imoto, co-developers of this system, and all of my fellow laboratory members who offered me much advice for improvements. And I would like to say thanks to the students who used this troublesome system.

References

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

Dauer, R. M. (1983) Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* **11**: 51–62.

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.

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

Leggetter, C. J. and Woodland, P. C. (1995) Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language* **9**(2): 171–185.

Prator, C. and Robinett, B. (1985) *Manual of American English Pronunciation, HRW International Editions*. Orland: Harcourt College Publishers.

Shimizu, M. and Dantsuji, M. (2002) A model of multimedia-based English CALL contents for Japanese students. *World Multiconference of Systemics, Cybernetics and Informatics: Proceedings*. Orland: IIIS Publishers.

Sugito, M. (1996) *English Spoken by Japanese*. Osaka: Izumishoin.

Tanaka, K., Kojima, H., Tomiyama, Y. and Dantsuji, M. (2001) Acoustic models of language-independent phonetic code systems for speech processing. *Spring Meeting of the Acoustical Society of Japan: Proceedings. Tokyo: Acoustical Society of Japan*, 1: 191–192.

Watanabe, K. (1994) *Instruction of English Rhythm and Intonation*. Tokyo: Taishukanshoin.

Young, S. and Woodland, P. (1996) *HTK Hidden Markov Model Toolkit*. Washington DC: Entropic Research Laboratories.