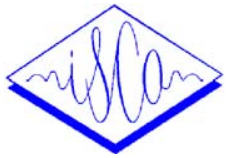

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

AN ENGLISH CONVERSATION AND PRONUNCIATION CAI SYSTEM USING SPEECH RECOGNITION TECHNOLOGY

Yasuhiro Taniguchi, Allan A. Reyes, Hideyuki Suzuki and Seiichi Nakagawa

Toyohashi University of Technology
Department of Information and Computer Sciences
Tenpaku-cho, Toyohashi, 441, Japan
Tel. +81 532 44 6777, FAX: +81 532 44 6777
E-mail: {taniguc1,nakagawa}@slp.tutics.tut.ac.jp

ABSTRACT

This paper describes an English conversation and pronunciation CAI using speech recognition techniques. This system was intended to recognize user's utterances and to respond to him properly according to the recognized results. In the case of a learner with unskilled pronunciation, because of differences in the phonemic system between his mother tongue and the second language, the speech recognition system cannot run normally. After this improvement, evaluation experiments were conducted. The results indicate that learners' ability in speaking and in listening to English is improved by using the system.

1 INTRODUCTION

Recently, a number of CAI (Computer Assisted Instruction) systems for second language learners have been constructed. It is also called CALL (Computer Assisted Language Learning) and there has been much interest in this especially for foreign language educators. Bernstein et.al have evaluated Japanese English pronunciation by speaker-independent HMM trained using native speaker utterances [1, 2, 3]. Also, Hamada et.al have proposed the various evaluation measures to mark Japanese English word pronunciation [4]. Both of these systems are reliable with the high correlation between the marks given by the system and those given by human experts.

In our laboratory, an English conversation CAI system has been developed using speech recognition techniques [5, 6]. This system was intended to recognize user's utterances and to respond to him properly according to the recognized results. This system can also evaluate user's pronunciation.

In the speech recognition part of these systems, if the second language learner's utterance spoken by a learner with unskilled pronunciation is recognized by phoneme-based HMMs trained using native speaker utterances, the speech recognition system cannot run normally, because of differences in the phonemic system between his mother tongue and the second language

In this study, as preliminaries of the construction of a CAI system for the acquisition of pronunciation and conversation skills, we describe an implementation of a speech recognition system using phoneme-based HMMs for the second language learner. Their phoneme models can be adapted using utterances from the target country's speakers (foreigners) and their phoneme label sequences.

After the improvements, evaluation experiments for speaking and hearing were conducted.

Table 1: Speech data

For Native	326 natives× 8 sentences = 2680 sentences 10 natives × 30 sentences = 300 sentences
Adaptation for Japanese	20 Japanese × 30 sentences= 600 sentences
test	10 Japanese × 50 sentences = 500sentences 5 American natives× 50 sentences = 250 sentences
Adaptation for speaker	20 sentences per 1 Japanese (× 10) 20 sentences per 1 American native (× 5)

2 ACOUSTIC MODEL

The system uses the 52 or 39 English phoneme-based HMM set which were trained from the utterances of American English speakers of the TIMIT database. And for beginners, we prepare two methods:

- One uses an HMM set which were adapted from the utterances of Japanese English speakers.
- The other uses a word dictionary which admits an inserted vowel between consonants expected by Japanese English speakers.

2.1 Adaptation

Table 1 shows speech data. In Table 1, English phoneme models were trained by 326 male speakers of TIMIT database, in addition they were retrained by 10 male native speakers for environmental adaptation of microphone, etc. These were initial models for adaptation for Japanese.

Figure 1 shows the correct recognition rate by using phonetic models adapted for Japanese or American native. "Adap" denotes the speaker adaptation mode by using their own voices. The more the

number of Japanese sentences for adaptation is, the higher the correct rate by phonetic models adapted for Japanese is.

We consider the evaluation measure of pronunciation based on the maximum likelihood of phoneme recognition or Rating (which is similar to the reference [2]) :

$$\text{Rating} = 10 - \left\{ \begin{pmatrix} maximum \\ likelihood \\ of \\ any \\ phoneme \\ sequence \end{pmatrix} - \begin{pmatrix} likelihood \\ of \\ phoneme \\ sequence \\ corresponding \\ to \\ utterance \end{pmatrix} \right\}$$

Figure 2 shows Rating of Japanese or American native using native acoustic models. The Rating values for native speakers are higher than those of Japanese. The rate of native speakers judged as a native speaker was 84.8%, while 80.4% by using only the likelihood of phoneme sequence corresponding to utterance. For Japanese, it was the same as native speakers (on the condition of Equal Error Rate). We found Rating is useful.

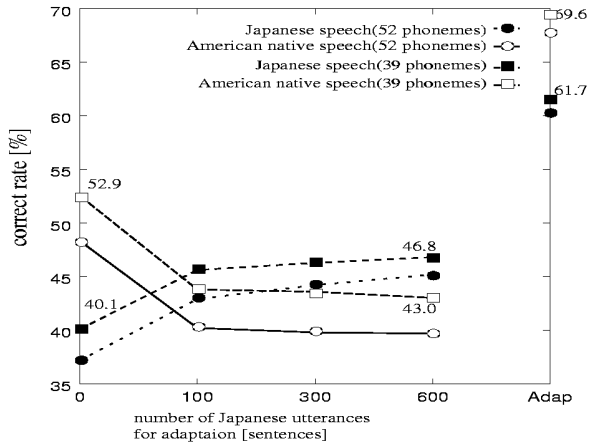

Figure 1: Correct recognition rate of phonemes by phonetic models adapted for Japanese or American native

2.2 Inserted Vowel Model

In the same situation, a English word dictionary for Japanese was changed, in which an inserted vowel between consonants expected by Japanese English speakers was admitted. Four short vowels and their HMMs were prepared as candidates for insertions.

Table 2 shows correct the sentence recognition rate (vocabulary size: 250 words) by phonetic dictionary adapted for using inserted vowels or not. The inserted vowel was effective for Japanese English speaker.

3 OVERVIEW OF CAI SYSTEM

Figure 3 shows a block diagram of English conversation CAI system. This system consists of speech input part, speech recognition part, conversation

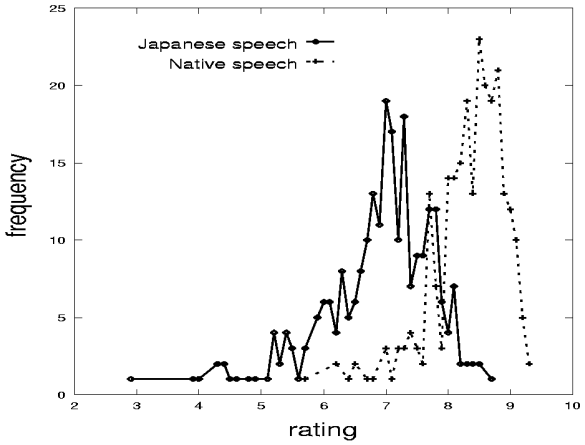

Figure 2: Rating of Japanese or American native by using native acoustic models

Table 2: Correct recognition rate of sentence by phonetic model adapted using inserted vowels or not

	with inserted vowel [%]	without inserted vowel [%]
correct recognition rate	55.0	54.2

control part and system output part. This system processes speech input, speech analysis and speech recognition in parallel. The speech recognition part refers to a grammar given by a context free grammar (CFG), chooses the best sentence from about ten sentences, that is matched an input speech data, regards the recognition result as an input speech sentence, and transfers the result to conversation control part. The conversation control part changes the grammar corresponding to the next conversation situation, and transfers this information to the system output part. The system output part presents the response using recorded speech and displays the situation image on display and a set of sentences that user can speak at the next conversation.

The system can treat dialogue on three topics :

(1) Immigration and Customs

(2) Hotel Check-In

(3) In the Street

The topic (1) has conversations on the immigration purpose, the place of his staying temporally, the length of his stay, the tax exemption (duty free) limit, and so on, (2) has conversations on the fare, the facilities, the reservation, the kind of room, and so on, and (3) has conversations on the way to some place, the way to take a bus, the emergency, and so on.

Figure 4 shows an example of conversation. This is the different strategy from the reference [7].

The user practices English conversation through a role-playing manner, wherein he chooses an appropriate utterance from among several choices which

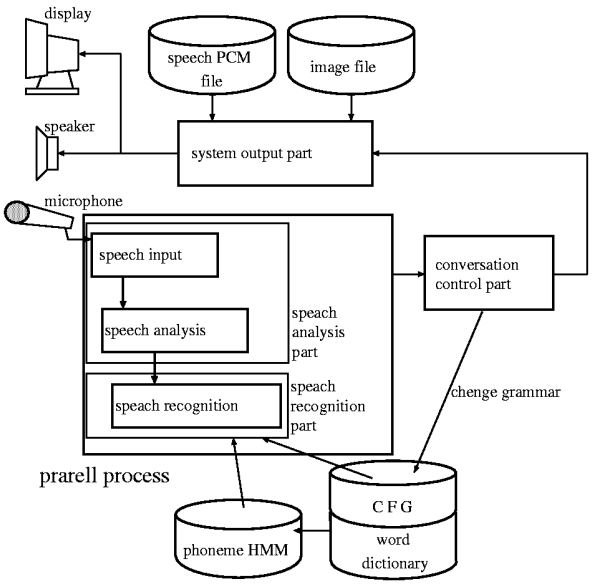

Figure 3: Block-diagram of English conversation CAI system

are displayed to him. In this example, user choices "Sightseeing". But if the user choices "I'm visiting my friend.", a sentence set of next choice is different (e.g. "What does your friend do here?" and so on). There are inappropriate sentences among the choices displayed and the system points the reason of wrong choice out if they are selected.

There is a function that enables the user to make the system repeat its utterance and another function that enables the system to reject indistinct utterances. The system has two modes: a free-talking mode and a test mode. In the free-talking mode, the user can set the conditions of his conversation by himself. On the other hand, in the test mode, the user has to conduct the conversation in such a way that he satisfies the condition set by the system.

And for beginners, system responses are also modified in speech rate using an analysis-synthesis technique[8].

In Figure 4, "( )" means optional words, "/ /" means learner must choose one from the same column's words. The recognition process was activated by native models and Japanese models in parallel. The system chooses better result of Rating, and its corresponding sentence is the system's output. If the recognition results by two sets of acoustic models are different each other, the rating for the system's output is recalculated by using opposite models. The system displays two Rating scores and corresponding models (Japanese/native) as shown in Figure 4. Thus, leaner knows whether his pronunciation is near native or not.

4 PRELIMINARY EXPERIMENTAL RESULT

Evaluation experiments were conducted. First, 11 learners were divided 3 groups :

```
System:  What is the purpose of your visit?
User:
   1.  Sightseeing.
   2.  (For) pleasure.
   3.  I'm here on vacation.
   4.  (I'm here on) business.
   5.  I'm visiting  /a /   /friend  /.
                     /my/   /relative/
                            /brother /
                            /sister  /
   6.  I'm here for two weeks.
   7.  I'm from Japan.
   8.  Pardon (me).
   9.  Excuse (me).
  10.  Please say it again.
  11.  I beg your pardon.
  12.  I don't understand.
User Response:  Sightseeing.
Rating:  Japanese 8.5 (native 6.3)
System:  Where do you plan to go?
User:
   1.  I don't know (yet).
   2.  ((I'm going) to)  /San Francisco/.
                         /Disneyland   /
                         /Florida      /
   3.  (I'm here for) one week.
   4.  I'm staying with my friend.
   5.  Pardon (me).
   6.  Excuse (me).
   7.  Please say it again.
   8.  I beg your pardon.
   9.  I don't understand.
```

Figure 4: An example of conversation of our CAI system

Table 3: Sentence recognition accuracy

	model	native		Japanese	
speaker	input sentence	correct sentence	recognition rate[%]	correct sentence	recognition rate[%]
1	41	24	59	37	90
2	39	30	77	36	92
total	80	54	68	73	91

(a) using the CAI through speech input (4),

(b) using the CAI through keyboard input (2),

(c) not using the CAI, but using a text book(5).

Group (a) practiced the CAI for about 30 minutes at every day during 5 days, and had hearing test and speaking test before/after CAI training at every day, and at 10th day and 30th day after final day (5th day). Groups (b) and (c) do in the same way as group (a).

4.1 Evaluation of speech recognition in this CAI system

Table 3 summarizes this English conversation CAI system's sentence recognition accuracy in a computer room environment. Results of sentence recognition accuracy of speech recognition by 2 Japanese male speakers were 68% (using native speaker models) and 91 % (using Japanese speaker models).

4.2 Hearing test

Hearing test was executed before/after CAI training at every day, and at 10th and 30th day after final day. Figure 5 illustrates the comparison result on 3 groups. The measure is a hit rate for words which are dictated. The most rising rate was performed by group (a). (We should notice that the hearing rate depends on contents of tested sentences, which are different from the three topics.)

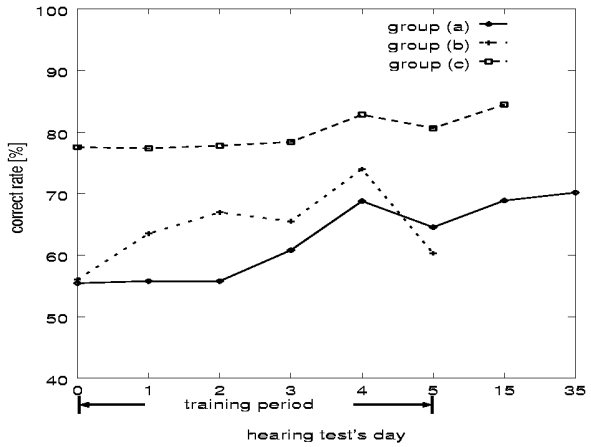

Figure 5: Results of hearing test (correct perception percentage of words)

4.3 Speaking test

Speaking test was also executed in the same time as hearing test. The test was based on Rating measure. Figure 6 illustrates the comparison result on 3 groups. And group (a) was also the best.

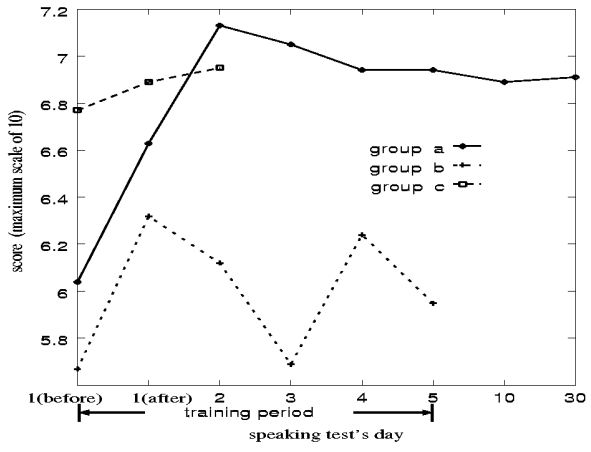

Figure 6: Results of speaking test (score using Rating at Section 2)

The results indicate that learners' ability in speaking and in hearing to English is improved by using the CAI system.

From results of questionnaires, learners answered they were satisfied with system's response time, and wanted to use the CAI system continuously.

5 CONCLUSION

In this paper, we describe an English conversation and pronunciation CAI system, and the evaluation result. As a result, we find this system works well for English study.

We expect such a CAI system stimulates learners' motivation for second language training. But we have some problems as follows:

Like this system, although limiting acceptable sentences of learner's utterance leads to the improvement of recognition accuracy, conversation flows are fixed. In the case of unlimiting, the set of expected sentences becomes large and leads to a low speech recognition rate. This brings two new problems, namely, 1) continue the conversation including mistaken utterances. (System doesn't point out it.) 2) not care of interjection ("eh", "oh") that appeared in real conversation.

Above functions are future work. Further, the judgment function for pronunciation, learning function for grammar, extension this CAI system for other topics are also future works.

References

[1] J.Bernstein, M.Cohen, H.Murveit, D.Rtischev and M.Weintraub : "Automatic Evaluation and Training in English Pronunciation", *Proc. ICSLP*, pp. 1185–1188 (1990).

[2] Horacio Franco, Leonardo Neumeyer, Yoon Kim, Orith Ronen : "Automatic Pronunciation Scoring for Language Instruction", *ICASSP*, pp. 1471–1474 (1997).

[3] Ahuping Ran, Bruce Millar, Phil Rose : "Automatic Vowel Quality Description Using A Variable mapping to An Eight Cardinal Vowel Reference Set", *Proc. ICASSP*, pp. 102–105 (1996).

[4] H.Hamada, S.Miki, R.Nakatsu : "Automatic Evaluation of English Pronunciation Based on Speech Recognition Techniques", *IEICE Trans. INF. &S YST.*, Vol. E76-D, pp. 352–359 (1993).

[5] Allan A. Reyes, Seiichi Nakagawa : "An English Conversation CAI System through Speech", *Proc. IPSJ*, No. 7R-07 (1993, in Japanese).

[6] Seiichi Nakagawa, Allan A. Reyes, Hideyuki Suzuki : "Spoken English recognition for Japanese utterances and a CAI system for English conversation" , *Proc. JSAI*, No. 22-13 (1995, in Japanese).

[7] Amir Najmi, Jared Bernstein : "Speech recognition in a system for teaching Japanese", *Journal of the Acoustical Society of America*, Vol.100, No.4, Pt.2, 3pSC13 (1996).

[8] Nobuaki Minematsu, Seiichi Nakagawa, Keikichi Hirose : "Prosodic Manipulation System of Speech Material for Perceptual Experiments", *ICSLP*, pp. 2056-2059 (1996).