A Japanese dialogue-based CALL system with mispronunciation and grammar error detection

Oh Pyo Kweon, Akinori Ito, Motoyuki Suzuki and Shozo Makino

Graduate School of Engineering, Tohoku University
05 Aramaki aza Aoba Aoba-ku, Sendai 980-8579 Japan
(kweop, aito, moto, makino)@makino.ecei.tohoku.ac.jp

Abstract

This paper describes a dialogue-based CALL (Computer Assisted Language Learning) system. One of the major problems in CALL systems is that learners are usually assigned a passive role. Learners have no practices in composing their own utterances. The other major problem is that lots of conventional CALL systems are pronunciation exercise systems. However, pronunciation exercise is only a part of exercise needed to increase a learner’s communication skill. In this paper, we propose a dialogue-based CALL system of new concept that enables exercise of composition, grammar and conversation in addition to pronunciation.

1. Introduction

1.1. CALL system

With the advent of globalization, more and more people are studying foreign languages. Foreigners must learn correct pronunciation, vocabulary and grammar to effectively communicate with native speakers. It is ideal if a foreign language learner can be taught on a one-to-one basis. Unfortunately, this is actually difficult for most students due to economic reasons. Conventional self-study methods such as tape recorder based learning force learners to identify pronunciation and grammar mistakes by themselves. This is a difficult task for a foreign language learner. If a learner could find their mistakes he/she would have no need to learn them. The application of speech recognition technology can solve this problem.

1.2. Conventional CALL system

As development of speech recognition technology in recent years, the expectation of a new concept CALL system has been growing. Various CALL systems are proposed [1, 2]. However, almost all conventional systems are practice systems of pronunciation and listening such as minimal pairs exercise. Only pronunciation and listening exercises are insufficient to obtain the ability of conversation. To obtain communication ability including grammar, pronunciation, listening, and composing sentences and so on, learners need to use a dialogue-based CALL system. A dialogue-based CALL system can roughly be divided into two categories by whether a student creates a response sentence actively or not. The first ones prepare ready-made answers when a learner is asked to answer a question [3]. As a result, learners have no practice in composing utterances on their own. AuraLog [4], for example, has produced an appealing language teaching system that feeds the user’s pronunciation of one of three written sentences to the recognizer. The path of the dialogue depends on the learner’s choice. The second ones let learners create limitedly their own utterances [5, 6]. Those systems give learners an active rather than passive role. For example, when the system asked them “When did you find it?” with a given keyword “Yesterday”, learners construct an answers using a keyword by themselves. However, the path of the dialogue is not changed. Because learners use a keyword, the answer is easily predicted. Both kinds of systems are useful for simple practice. However, using these systems, learners cannot do conversation practices as they do with human teachers.

1.3. Objective

To develop a learner’s communication capability, it is ideal to use a system that allows learners to do exercise “real conversation” as shown in Figure 1. There are two factors to realize “real conversation”. One is that the system should allow learners to practice in composing various utterances actively. For example, when the system asks [May I help you?], they can say [I want to a souvenir] or [What is the most famous?] and so on. The other is that the path of the dialogue should be changed by learners’ response utterances. The object of this work is to develop a CALL system that satisfies these requirements. In this work, the target language is Japanese, and the learners are assumed to be Korean.

2. Overview of the Proposed CALL system

As we mention in section 1.3, our object is to construct the system which allows learners to practice ‘real conver-
System: This is a souvenir store. May I help you?

Foreigner: Gyutan, please. What is the most famous souvenir in Sendai city?

System: The most famous one in Sendai is Gyutan.

Foreigner: How much is it?

System: 1000 for one box.

Foreigner: I would like to buy one box.

System: Do you want to pay by cash or credit card?

Foreigner: Credit card, please.

System: Thanks.

Figure 1: One example of conversation exercise

The proposed system is shown in Figure 2. Using the proposed system, learners can practice conversations and sentence composition. The system checks and returns grammar errors as well as pronunciation errors. Hence, doing conversation exercise, learners can practice pronunciation, grammar, sentence composition.

To construct the proposed system, we have to solve lots of problems. When learners compose and pronounce various utterances, these utterances have many problems shown below.

1. Pronunciation errors
2. Grammar problem
   2.1. grammatical and lexical errors
   2.1. Out-of-grammar utterances
3. Poor recognition performance for foreigners

Language learners have pronunciation problems. For example, a learner may pronounce a Japanese sentence [hitotsu kudasai: one please] as [hitotsyu kudasai]. They also make grammatical and lexical errors. For example, they pronounce [ ni kudasai: second please] instead of [hutatsu kudasai: two please]. They also tend to compose utterances which are out of system grammar (Out-of-grammar utterances). The other problem is poor recognition performance. The system can not recognize utterances correctly, especially utterances of foreigners because foreigners’ speaking speed is slow and pronunciation is ambiguous. We discuss about each problem in section 3-5.

3. Evaluation of pronunciation errors

First, mispronunciation rules were prepared according to previous pronunciation analysis described in [7]. When a learner pronounces the sentence, sentences with mispronunciation are automatically generated according to the mispronunciation rules. Then, HMMs correspond to each sentence are matched against Japanese utterance by Korean speaker in order to calculate the likelihood. Using a standard utterance of a native speaker, a standard likelihood is calculated as well using above method. The system calculates the normalized likelihood difference $D_l$ and $D_n$ for the learner and native speaker,

$$
D_l = \frac{L_{l1}}{F_{l1}} - \frac{L_{l2}}{F_{l2}} \quad \text{and} \quad D_n = \frac{L_{n1}}{F_{n1}} - \frac{L_{n2}}{F_{n2}}
$$

where $l_{(1 \text{ or } 2)}$ and $n_{(1 \text{ or } 2)}$ are the sentences which are automatically generated by correct pronunciation rules and the mispronunciation rules (1 or 2) for the learner and the native speaker. $L$ and $F$ are the likelihood and the duration in frames correspond to $l_{(1 \text{ or } 2)}$ and $n_{(1 \text{ or } 2)}$. Then, the system calculates the scores for the learner’s speech. Comparing the difference of a normalized likelihood of learner with that of a standard native speaker using the weighting factor $W (-1 \leq W \leq 1)$, pronunciation level of the utterance is degraded and returned such as ‘Very Poor’, ‘Poor’, ‘Ambiguous’, ‘Good’ and ‘Very Good’. The advantage of the proposed method is that the training data can be easily extended by adding new rules of mispronunciation into the system [2].

![Figure 2: Overview of the proposed CALL system](image-url)

![Figure 3: The evaluation result by system and native speakers](image-url)
3. The correlation coefficient is 0.72. It is verified at a significance level of 0.05. We thus consider intelligibility scores by the system to be similar to pronunciation evaluation scores by native speakers.

4. Grammar problem

4.1. Detection and recovery of grammatical and lexical errors

To treat grammatical and lexical errors, we defined the dialogue situation as a particular task such as a souvenir task and hotel reservation task. Hence, a system can easily predict utterances which learners compose. Although a learner’s utterances are restricted by limitation of the task, learners can compose their own utterances in conversation practices. However, learners may compose utterances with grammatical and lexical errors such as ‘I want to buying’ instead of ‘I want to buy’ as we mentioned in section 2. To solve this problem, we also extended vocabulary and grammar in the dictionary using conventional work [10]. Hence, even though learners make grammatical and lexical errors in sentence composition, a system can recognize errors, and point out their mistakes.

4.2. Suppression of out-of-grammar utterances

To decrease out-of-grammar utterances, we introduced pre-exercise before doing conversation exercise with the system. Firstly, learners do pre-exercise such as vocabulary, grammar, and typical conversation examples. Then, they do real conversation exercises with the system. Because of the influence of pre-exercise, learner naturally compose sentences which are inside of system grammar. However, we do not exactly know influence and effectiveness of pre-exercise. We investigated the influence and effectiveness of pre-exercise.

4.3. Experiment method

We did interview experiment without/with pre-exercise [8][9]. First, situation is explained to learners. For example, native Japanese speaker explains the situation like [Here is a souvenir store. I am a staff of the store. You are a customer]. Then, learners are directly interviewed with native speaker without any pre-exercise (Interview 1). The other way is that learners are interviewed with native speaker after they do pre-exercise such as typical conversation examples and vocabulary (Interview 2). Because our purpose is not to investigate listening ability, native speaker writes the sentences if foreigners have listening problem. Two Korean speakers were interviewed about a souvenir task and hotel reservation task. One Korean has studied Japanese about 1 year and the other has done about 1.6 year.

<table>
<thead>
<tr>
<th>Interview 1</th>
<th>Interview 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. of out-of-grammar utterances</td>
<td>23.5 %</td>
</tr>
</tbody>
</table>

4.4. Result of interview

In interview 1, learners used a lot of filler words. Because they had no pre-exercise, learners have problems to construct sentences immediately to make a conversation. Comparing the result of interview 2, there are lots of sentences which are outside of system grammar. For example, one of foreigners asked [For my friend who lives in Korea now, I would like to buy a souvenir?] corresponding to a question of native speaker [May I help you?] in a souvenir task. However, words and sentences used by learners in interview 2 are similar to that in pre-exercise. The result is shown in Table 1. As comparing results of interview 1 to 2, in case of using pre-exercise, the number of out-of-grammar is decreased. Using pre-exercise, learners naturally compose utterances which are inside of system grammar.

<table>
<thead>
<tr>
<th>Interview 1</th>
<th>Interview 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. of grammatical errors per one utterance</td>
<td>0.26</td>
</tr>
</tbody>
</table>

As shown in Table 2, grammatical and lexical errors of foreigners are also decreased in case of using pre-exercise. However, tendency of grammatical and lexical errors with/without pre-exercise were happened in Japanese particles such as [soreni simasu → soredesimasu]. It is possible to deal with grammatical and lexical errors, using extended vocabulary and grammar as we mentioned in section 4.1 (Interview data is small to cover and analyze all kind of grammatical and lexical errors. It is future work).

5. Improvement of recognition performance for utterances

As we mentioned in section 2, recognition performance is poor because speaking speed of foreigners is slow and pronunciations are ambiguous. For example, a Korean can hardly distinguish the voiced and unvoiced sound. They only have a distinction of aspiration and un-aspiration. If those ambiguous pronunciations are trained, the recognition rate could be increased.
5.1. Experiment method

The details of experiment are shown in Table 3. We checked phoneme recognition performance. We constructed Japanese phone models (baseline models) using Japanese speech DB. Then, baseline models are trained using Foreigners’ speech DB\[11\](trained models).

<table>
<thead>
<tr>
<th>HMM Features</th>
<th>Monophone HMM Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC+ΔMFCC+Pow+ΔPow, Total 34features</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baseline Model</th>
<th>Japanese Phoneme Model from ATRB, C sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training data</td>
<td>8 Korean male, 4 Korean female</td>
</tr>
<tr>
<td>Testing data</td>
<td>4 Korean male, 4 Korean female (about 100 sentences per each Korean speakers)</td>
</tr>
</tbody>
</table>

5.2. Result of Recognition rate

The result of recognition rate is shown in Figure 4. M-Base and F-Base indicate the recognition rate using baseline models of male and female. M-training and F-training indicate the recognition rate after baseline models are trained by Foreigners’ speech DB. After training the baseline models, the correct/accuracy rate of male are increased from 57.97/10.23% to 66/13.37%. The rate of female is also increased from 61.78/15.10% to 66/19.96%.

We also check the recognition tendency using about 20 sentences which are derived from interview experiment in section 3. In all case except one sentence, recognition results are better when we used trained models instead of baseline models. Word correct rate is increased from 79.7% to 81.1% after training the baseline models.

6. Conclusions

In this paper, we discussed about a dialogue-based CALL system. We proposed the CALL system to allow learners to exercise ‘real conversation’. Learners compose various sentences and the path of dialogue is depended on the composed sentences of learners. Learners can also check their grammatical and pronunciation errors. We also discussed about the effect of pre-exercise. The number of out-of-grammar sentences is decreased if learners do pre-exercise. In order to increase the recognition performance of foreigner, ambiguous pronunciations were trained. We checked that the recognition rate was increased.

7. Future work

There still remain lots of problems to be discussed. First, interview data is too small to analyze grammatical and lexical errors. For future work, we need to collect more interview data and analyze grammar and lexical errors more correctly. Grammatical error and lexical errors are also influenced by language abilities of learners. We also need to analyze the grammar error depending on language abilities of learners.

8. References


