ENGLISH CALL SYSTEM WITH FUNCTIONS OF SPEECH SEGMENTATION AND PRONUNCIATION EVALUATION USING SPEECH RECOGNITION TECHNOLOGY

Yasuo Ariki and Jun Ogata

Ryukoku University
Department of Electronics and Informatics
Seta, Otsu-shi, Shiga, 520-2194, Japan
ariki@rins.ryukoku.ac.jp

ABSTRACT
In communication learning by second language, three abilities have to be improved; listening, speaking and writing abilities. In this sense, it is important to evaluate user’s pronunciation ability and to detect mispronunciations in English CALL (Computer-Assisted Language Learning) systems. In this paper, we propose three functions (segmentation, phrasing and pronunciation evaluation) in the CALL system using speech recognition technologies. The system was evaluated by the answer to questionnaires for ten learners.

1. INTRODUCTION
Computer Assisted Language Learning (CALL) systems have been studied owing to language globalization and e-learning requirement[1]-[2]. In the CALL system, it is important for users not to be tired in language learning. Cinema film is a good material for CALL systems because it includes vivid and casual conversation in a daily life. In this viewpoint, we propose a call system which employs cinema film as the video materials.

In a CALL system, a user is sometimes forced to memorize and repeat one sentence. However, this sentence learning is meaningless because one sentence is composed of several phrases so that the sentence learning prevents the generation ability of variety of sentences by combining phrases. From this viewpoint, we propose phrase learning instead of sentence learning in the English CALL system.

In our proposing CALL system, important and interesting phrases to be learned are selected from a list of phrases by a user at first. Then the video clip, extracted from 100 movies and including the phrase, is selected from a list of video clips by a user. After watching the movie clip, the user can improve his/her language ability by the following functions provided by the CALL system.

(1) Pronunciation evaluation: Pronunciation of user speech is evaluated at the word level as well as sentence level based on phoneme recognition results. This function is utilized when a user tries to improve his/her pronunciation. The CALL system advises a user what is wrong in his/her pronunciation and how to pronounce each word correctly.

(2) Word segmentation: Speech from user and actor/actress is segmented into individual word by using a forced alignment technique between the speech and the closed caption. This function is utilized when a user wants to compare his/her own speech with the actor/actress speech at the word level.

(3) Phrase segmentation: Speech is segmented into phrases based on the results of word segmentation. This function is utilized when a user can not understand the sentence meaning after listening to the sentence.

2. OVERVIEW OF THE CALL SYSTEM

2.1. Process flow
Fig.1 shows an overview of our CALL system. Language learning proceeds as follows;

(1) A user selects one phrase from a list of phrases. Then the system extracts video clips including the selected phrase from 100 movies.

(2) The titles of movies where the movie clips are extracted are presented to the user. When the user selects one movie clip from a movie list, then the movie clip is played back.

(3) After watching the movie clip several times and understanding the phrases and sentence composition, the user speaks the sentence. This learning method is similar to the sentence learning but different in that the user is learning phrases and by combining the learned phrases, the user can compose any kind of sentences.
The CALL system evaluates the user pronunciation at the word level and sentence level.

Two kinds of evaluated scores are displayed to the user. One is sentence pronunciation score. The other is individual word score. When the user wants to know what is wrong in his/her pronunciation at the word level, the user can click the displayed word and get the detailed information.

When the user wants to compare his/her own speech with the actor/actress speech at the word level, the system can segment the speech into individual word by using a forced alignment technique between the speech and the closed caption. The sentence is segmented into phrases by putting pause at the phrase boundaries.

The segmented words or phrases are played back to the user.

The segmented words or phrases are played back to the user.

### 2.2. User interface

Fig. 2 shows the main user interface after selecting the phrases in the CALL system. In the figure, “select” indicates the selection of the movie clips listed in the above. “Movie” indicates the movie clip starting. The buttons at the below indicate the recording and record stopping. “Score” indicates the evaluation of user speech and score display. The captions in the upper part and lower part correspond to the actor speech and user speech. When clicking the word, the word is played back.

### 2.3. Phrase list

Table 1 shows the phrases to be selected by the user. The phrases are collected from 100 movies and occurrence frequencies are computed. In the table, one word and two words phrases are listed with the frequencies (%) in the descending order. Based on this table, the system can make the several groups of phrases. For example, the most frequently occurred phrases are provided for beginners and the least frequent words are provided for skilled person.

![Fig. 1. Overview of the CALL system](image)

**Table 1. Phrase list with frequencies**

<table>
<thead>
<tr>
<th>rank</th>
<th>1 word</th>
<th>2 words</th>
<th>Frequencies(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>do</td>
<td>come on</td>
<td>6.06</td>
</tr>
<tr>
<td>2</td>
<td>know</td>
<td>thank you</td>
<td>4.00</td>
</tr>
<tr>
<td>3</td>
<td>come</td>
<td>see</td>
<td>2.96</td>
</tr>
<tr>
<td>4</td>
<td>see</td>
<td>tell me</td>
<td>2.47</td>
</tr>
<tr>
<td>5</td>
<td>use</td>
<td>tell you</td>
<td>2.38</td>
</tr>
<tr>
<td>6</td>
<td>look</td>
<td>use me</td>
<td>2.35</td>
</tr>
<tr>
<td>7</td>
<td>think</td>
<td>excuse me</td>
<td>1.96</td>
</tr>
<tr>
<td>8</td>
<td>say</td>
<td>do it</td>
<td>1.84</td>
</tr>
<tr>
<td>9</td>
<td>tell</td>
<td>come here</td>
<td>1.46</td>
</tr>
<tr>
<td>10</td>
<td>take</td>
<td>know you</td>
<td>1.38</td>
</tr>
</tbody>
</table>

### 3. PRONUNCIATION EVALUATION

#### 3.1. Evaluation interface

Fig. 3 shows a window display of the evaluation score to the user speech. In the upper of the figure, the sentence score is displayed. The first line and the second line are the phoneme recognition sequences of the actor speech and the user speech respectively. Just below the phoneme recognition sequence, the evaluation score 81 and the number of deleted(0), substituted(1) and inserted(2) phonemes are presented.

In the lower of the figure, the word score of the user speech is shown. In the left, the word names and their scores are shown. If the word “stopping” is clicked, the explanation is shown in the right part. In this example, shown are the correct phoneme sequences (staapixng) and the recognized phoneme sequence of the user speech (stojpixngu-j). Finally the explanation and advice are shown that the phoneme [aː] is substituted by [oː]-j and the [g] is followed by [u]-j in the user speech. Here, phonemes followed by “-j” indicate the Japanese pronunciation of the phonemes. For example, [oː]-j means the Japanese pronunciation of phoneme [o].
The user can understand what is wrong in his/her pronunciation and try again to improve the sentence and word scores.

3.2. Evaluation method

In order to evaluate the word pronunciation, the user speech has to be segmented into each word. This function can be carried out by a forced alignment between the user speech and the corresponding closed caption. The forced alignment is a method to compute the optimal path on the two dimensional matrix; the horizontal axis is input speech and the vertical axis is sentence (closed caption) HMM produced by concatenating the word HMMs which are composed of phoneme HMMs.

To detect the deletion, substitution and insertion phonemes, we prepared the error pronunciation (phoneme sequence) for each word including deletion, substitution and insertion which are often observed in the Japanese pronunciation. The example is the phoneme sequence (sto-jpixngu-j) for word “stopping” whose correct sequence is (staapixng). Therefore, several phoneme sequences are prepared for each word. The error pronunciation is produced based on the following rules:

- Vowel insertion: Consonant concatenation is natural in English. However in Japanese pronunciation, a vowel is sometimes inserted between consonants. For example, [u] is sometimes inserted after [p] of “up” and makes (_pdu) instead of (p).
- Substitution: Phonemes not observed in Japanese are often substituted by the most similar Japanese ones. This is called substitution. For example, English [l] and [r] are substituted by Japanese [r]. The other example is that [th] is substituted by [s].
- Deletion: Some phonemes are deleted in Japanese. For example, “far” is caused “fa” after deletion of ending [r].

The forced alignment can compute the optimal path of the phoneme sequence to the user speech. The word score is computed by conventional phoneme accuracy; the ratio of the number of phonemes subtracted by deletion, substitution and insertion phonemes to the number of correct phonemes in the word. The sentence score is computed by the phoneme accuracy at the sentence level.

4. SPEECH SEGMENTATION

Speech segmentation is required in a case to compare user speech with the actor speech at the word level. Furthermore, this word segmentation can be used to display the closed caption word by word synchronously to the speech being spoken. Here we describe the word level segmentation at first and then phrase level segmentation.

4.1. Segmentation method

Word level segmentation is carried out using the forced alignment between the user or actor/actress speech and the sentence HMM as described in section 3. The word dictionary is relatively simple compared with that used in the pronunciation evaluation. Each word is simply presented as a phoneme sequence. If English phoneme is only employed, the Japanese user speech can not be well segmented because English spoken by the Japanese user includes Japanese pronunciation. Consequently we employed mixture of English 46 phonemes and Japanese 41 phonemes in the word level segmentation of the speech.

Table 2 shows word examples in the dictionary presented by English, Japanese and mixture phoneme sequence. In the table, sh-j or i:-j indicates the Japanese phoneme.

<table>
<thead>
<tr>
<th>Table 2. Example of word representation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>English representation</td>
</tr>
<tr>
<td>Japanese representation</td>
</tr>
<tr>
<td>Mixture representation</td>
</tr>
</tbody>
</table>

4.2. Experimental condition

In the forced alignment, context independent monophone HMM is employed. In the training of English phoneme HMMs, TIMIT data is used for initial HMM training and WSJ data of 147 males with 20614 spoken sentences is used for concatenation training.

In the training of Japanese phoneme HMMs, ATR data set (a-j) is used for initial HMM training and JNAS data of 137 males with 21782 spoken sentences is used for concatenation training. The acoustic features are 39 dimensional vectors (12 dimensional MFCC and power, their Δ...
and ΔΔ) and are same for both English and Japanese phoneme HMM.

4.3. Experimental result

Word segmentation was carried out for 3 sentences (112 words) spoken by four Japanese males using English phoneme HMMs and mixture phoneme HMMs. Segmentation result is evaluated by the average of the difference (ms) at the word boundaries between true ones and segmented ones.

The result is shown in Table 3. In the table, the correct rate indicates the ratio of the correct words that are defined as the words whose difference at the word boundary is less than 3 frames (30ms). The correct rate showed that the word segmentation is 10% higher by mixture HMMs than that by English HMMs. From this result, the mixture HMMs are employed in the word level segmentation.

<table>
<thead>
<tr>
<th></th>
<th>English HMMs</th>
<th>Mixture HMMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaged difference</td>
<td>20.8 ms</td>
<td>12.5 ms</td>
</tr>
<tr>
<td>Correct rate</td>
<td>83.0 %</td>
<td>92.9 %</td>
</tr>
</tbody>
</table>

4.4. Phrasing function

When the user can not get the meaning of the spoken sentence, the user can hear the sentence by pausing the phrase boundaries. The user also clicks the phrases and listens to them independently. In the bottom of Fig.2, the symbols S, V and O are shown and they indicate the subject, verb and object phrases respectively.

In the language learning, the user listens to the whole sentence at first. When the user can not understand the meaning, then the user listens to the sentence by pausing the phrase boundaries. Nevertheless, if the user can not get the meaning, the user clicks the each phrase. If the phrase can not be understood, the user clicks each word and listens to the word.

5. SYSTEM EVALUATION

The system was evaluated by 10 person answers to questionnaires after using the system. The questionnaires are as follows;

(1) Could you operate the system easily?
(2) Is the phrasing function is useful?
(3) Did you enjoy this system?
(4) Do you want to use this system again?
(5) Is this system useful for English learning?
(6) Is the system working well?
(7) What is good in this system?
(8) What is poor in this system?
(9) What do you think of other important functions?

The questions 1-6 are evaluated at 5 levels (max is 5 and min is 1) and the averaged score is computed. The result is shown in Table 4.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Could you operate the system easily?</td>
<td>4.0</td>
</tr>
<tr>
<td>(2) Is the phrasing function is useful?</td>
<td>3.0</td>
</tr>
<tr>
<td>(3) Did you enjoy this system?</td>
<td>3.9</td>
</tr>
<tr>
<td>(4) Do you want to use this system again?</td>
<td>4.2</td>
</tr>
<tr>
<td>(5) Is this system useful for English learning?</td>
<td>4.2</td>
</tr>
<tr>
<td>(6) Is the system working well?</td>
<td>3.7</td>
</tr>
</tbody>
</table>

The user answered the following points as the goodness of the system.

- Cinema video clip is interesting.
- It is useful that the user speech is compared with the actor speech at the word level.
- It is useful that the speech is segmented into words in real time.
- The processing time is almost real time so that the users have almost no stress.

6. CONCLUSION

In this paper, we proposed thee functions (pronunciation evaluation, word segmentation and phrase segmentation) using speech technology in the CALL system. In the word segmentation, the forced alignment with mixture HMMs is employed and showed the good result compared with English HMMs. We also proposed the pronunciation evaluation function which can evaluate the English sentence spoken by Japanese at the word level and sentence level. At the word level evaluation, the system can give the explanation and advice to the user how to speak the word. The system was evaluated by the questionnaires to 10 users. Future works will be to give the more informative explanation and advice for user speech improvement.

7. REFERENCES
