Ya-Ya Language Box - A Portable Device for English Pronunciation Training with Speech Recognition Technologies

Fu-chiang Chou

Department of Computer and Communication Engineering
Ming Chuan University, Taiwan, R.O.C.

fuchiang@mcu.edu.tw

Abstract

This paper describes the application of speech recognition technologies in computer assisted pronunciation training. A portable pronunciation training device called Ya-Ya language box was completed in this research. The system provides a pronunciation training course and uses the speech recognition technologies to identify the pronunciation errors and provide diagnostic feedbacks for the users to correct their pronunciations. The system had been evaluated by forty students in elementary schools and most of them affirmed the system is helpful for their English pronunciation learning.

1. Introduction

In recent years, especially into the new millennium, with the main trend of the rapid internationalization of the economy, English learning has been promoted in a national scale in Taiwan. The Ministry of Education of Taiwan devoted to the construction of an internationalized environment and decided to teach English in elementary school. However, it’s difficult to have sufficient and qualified English teachers for these new courses in short time, especially for the teaching of pronunciation which needs lots of interaction and practices. To better tackle this challenge in the information age, Computer Assisted Pronunciation Training (CAPT) seems to be a possible solution. In fact, there are lots of CD-title and web site devoted to English learning in Taiwan. But few of them are target at pronunciation training. Because people can only use the keyboard and mouse to interact with computer in these programs, therefore, most of the programs are focused on the learning of vocabulary or grammar instead of pronunciation. To learn pronunciation, the users need to interact with the computer by voice; therefore, speech recognition technologies need to be integrated into the programs to evaluate and diagnosis the pronunciation of the users.

It’s not a new idea to incorporate speech recognition into the software for language learning. There are some CD titles and web sites have provided such kinds of functions, for example the “Tell me more” from Auralog [1] and the “MyET” from L Labs Inc [2]. However, compare to the “electronic dictionary” which seems to be the most popular language learning device in the market, these software programs seem to have larger entry barrier. Most of the students learn English in the school where they don’t have their own computer to run these programs. They use the electronic dictionary to look up for the meaning and listen to the pronunciation of a new word. Even in home where the students have computer and networks, they still need to install the sound card and microphone correctly to use these programs. That's why the electronic dictionary is more popular than these CD-titles and web sites in learning English. Moreover, thanks to the quick growth of microprocessor’s performance, the number of applications in the electronic dictionary also increased quickly. Most of the electronic dictionaries not only provide the “talking” functions to demonstrate the correct pronunciations, it also have a “recording” function to record the pronunciation of the user. The users can compare their own pronunciation with the “normal” pronunciation and try to correct their pronunciation by their own perception. The problem is that sometimes the students can't even tell the differences between the normal pronunciation and their own pronunciations. Therefore, if we can incorporate the speech recognition technologies into the electronic dictionary, and provide adequate feedback functions, the students can follow the instructions to correct their pronunciation. With such kind of function, the electronic dictionary can be upgraded into complete a language learning machine. Therefore, the target of this research is to develop a software application in portable device that incorporates the speech recognition technologies for learning of English pronunciations. It’s difficult to make applications in real electronic dictionaries because all of them use the proprietary operating system and platform. Therefore, the alternative is to use the A WinCE based Personal Digital Assistant (PDA) as the hardware platform because the development tools are free and open to public. At last, an English pronunciation learning program with speech recognition technologies was completed in this research. The system is running on WinCE based PDA and is called “Ya-Ya” which means “tooth tooth” in Chinese.

2. Problems and Research Issues

There are many academic institutions or commercial companies in the world have developed different CAPT systems and few of them could get positive feedbacks in the literature review so far. According to the opinion in [3] · the major problems are:

1. The accuracy of speech recognition doesn’t meet the requirements for non-native speakers.
2. The system can’t identify or provide suitable feedbacks for the pronunciation errors.

For the first problem, most of the commercial products use the existing speech recognition engine for the evaluation of pronunciation accuracy. However, these engines are designed for speech recognition of native speakers; the accuracy for non-native speakers is not in the scope of the original design. Therefore, the accuracy can’t meet the basic requirement in real applications [3] [4]. In practice, if we want to use the speech recognition engine for non-native speaker, it’s necessary to have some adaptation for the acoustic models, some methods can be found in [5][6]. For the second problem,
most of the existing system only provides a score for the accuracy of pronunciation of a specific word or phoneme. Assume that the score is accurate; the users still don’t know how to correct their pronunciation [7]. This problem was addressed in the ISLE project in which the system can provide the adequate feedback according to a database that stores the most common errors of the foreign language learners [8].

The approach in this research is based on the same idea in the ISLE project. We first find these common pronunciation errors of Taiwanese students in the reports on many interlanguage phonology literatures. Lado had proposed a contrastive analysis hypothesis that implies if we have compared and analyzed the difference between the native language (L1) and the foreign language (L2), we can predict the possible problems that learner will encounter and design the materials for the learner to overcome these problems[9]. In this theory, L1 will interfere the learning of L2 and the acquisition of L2 is the process to overcome the interference of L1 and replace by the features of L2. Therefore, we have investigated the contrastive analysis of Mandarin (L1) and English (L2) by Hide [10]. More than that, considering that Taiwanese is the most common native language in Taiwan; the studies of the common pronunciation errors for Taiwanese people are also included in the analysis [11]. Here are some examples for the common pronunciation errors of Taiwanese English.

For vowels:
- /æ/ and /i/ are often confused.
- /æI/ and /æE/ are confused. Most of the time, /æI/ and /æE/ are replaced by /æ/.
- /æ/ is often realized as /ɑ/.
- /ʊ/ is often replaced by /u/.
- /ɔ/ is often replaced by /o/.
- …

For consonants:
- /θ/ is often replaced by /s/.
- /h/ is often replaced by /x/.
- /r/ and /l/ are confused.
- Initial /ð/ is often replaced by /l/.
- /v/ is often replaced by /b/.
- Final stops are often deleted.
- …

After the studies of these common errors, the speech recognition engine is used to detect these errors in the pronunciation of Taiwanese learners. The method is to add these errors as pronunciation variations into the recognition grammar. Therefore the engine will compare the confidence scores for different pronunciations in the grammar and identify the “real pronunciation” of learners. If the pronunciation error occurred, the system will provide the specific instructions for the type of errors. Then the users can follow the instructions to correct their pronunciations.

3. Speech Recognition and Grammar

It’s not a target to develop a new speech recognition engine for language learning in this research. Therefore, it’s necessary to select an existing recognition engine for the system. In fact, there are not many choices of the recognition engines in PDA platform. Because the author was previously an employee at ScanSoft Taiwan, therefore, the VoCon® 3200 from ScanSoft was licensed for research purpose in this study [12]. The VoCon® 3200 is a complete development suite designed for adding speech recognition functionality to a variety of automotive, mobile, and embedded applications. As a result, it provides a version of engine that can be used for the WinCE based PDA. The VoCon® 3200 comes with a complete lexicon and grammar definition with large flexibility just like the common speech recognition engines on PC platform. The definition of the speech context for recognition could be written in BNF+ format and the system has a built-in grapheme-to-phoneme module and on-line grammar compiler functions to generate the search space for recognition. After recognition, the Top-N candidates with confidence scores will be returned as a result.

To detect the pronunciation errors for a specific word, the possible variations need to be added into the BNF+ grammar. For that reason, a special grapheme-to-phoneme module was developed to generate a “pronunciation error detection grammar” for the words in the learning course. For example, if the vowel /æE/ is the learning target and the word “Apple” is used for practice. Then the confusion set of /æI/, /æE/ and /æ/ will be included in the pronunciation error detection grammar. The pronunciation error detection grammar for “A” in Apple is listed below:

```
# BNF+EM V1.0;
!grammar "00001"; /* A-Apple*/
!language "American English";
!pronounce COR "#@.pl$#";
!pronounce @_E "#E.p$l#";
!pronounce @_ei "#e&.pl$#";
!pronounce @_A "#A.p$l#";
!start <Speech>;
<Speech>: COR | @_E | @_ei | @_A ;
```

The possible results for this grammar are:
- COR: means pronunciation is correct
- @_E: means /æE/ is pronounced as /E/
- @_ei: means /æE/ is pronounced as /ei/
- @_A: means /æE/ is pronounced as /a/
- Unrecognized: other types of errors

According to the recognition result, the type of pronunciation error was detected. For different type of errors, the corresponding directions will be displayed on the screen to instruct the users how to correct their pronunciations. If /æE/ is pronounced as /E/, then the instruction should be “open mouth wider and move your tongue back”.

Unrecognized: other types of errors
4. Contents and System Architecture

The Phonics was selected as the major contents and learning materials in the system because it’s the merely way of teaching pronunciation in the elementary schools on Taiwan. The Phonics includes the teaching of the basic letter-sound relationships and the application of such knowledge to facilitate reading and spelling. Most of the learning materials were selected from the current English textbooks of the elementary schools. We followed the orders and architectures in the textbooks to introduce the Phonics and use the words selected from the 1000 common used words designed by the Ministry of Education as examples. As shown in Figure 1, the contents are classified into three categories, [A~Z], [Vowels] and [Consonants].

![Figure 1: The architecture of contents in Ya-Ya](image)

Before learning the rules for Phonics, the students must familiar with the sounds corresponding to different alphabets. Thus, [A–Z] is the first category. For the procedures in Phonics, the first step is to identify the vowels then combined with the constants. As a result, [Vowel] is the second category and [Consonants] is the third category. The Vowels are further classified into three groups: short vowels, long vowels and vowel teams. The consonants are also divided into three groups according to the similarity in articulation.

For each page of a specific sound to learn, the system will show the words for practice on the left side. After the users select the word they want to practice, they can push the “Teach” button to hear the normal pronunciation or push the “Speak” button to practice their pronunciation. This page is shown in Figure 2.

![Figure 2: The learning page in Ya-Ya](image)

If the users choose to practice their pronunciation, the recognition page will be displayed on the screen as shown in Figure 3. After pushing the start button, the users need to speak out the word in five seconds. The system will display the instructions to correct the pronunciation after identify the error type as described in the previous section.

![Figure 3: The recognition page in Ya-Ya](image)

5. Evaluation

To evaluate the value of the system to pronunciation training, an initial survey by the target users was carried out. There are 20 students of grade one and 20 students of grade three in elementary school involved in this evaluation. The procedures for this evaluation are:

1. Introduce the basic operations of PDA and demonstrate the operations for the learning page, especially focus on how the system would feedback with different errors and how to correct pronunciation according to the feedback.
2. Let the students use the system, including browsing different pages and practice for different examples.
3. Make a questionnaire survey for each student.

The result of this survey is shown in Table 1.

<table>
<thead>
<tr>
<th>Grade 1</th>
<th>Grade 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>The system is attractive to you or not?</td>
<td>20</td>
</tr>
<tr>
<td>The system is helpful for pronunciation learning or not?</td>
<td>18</td>
</tr>
<tr>
<td>Do you think it’s easy for you to operate the PDA?</td>
<td>1</td>
</tr>
<tr>
<td>The instruction of recognition result is reasonable or not?</td>
<td>4</td>
</tr>
<tr>
<td>Do you want to buy this system or not if affordable?</td>
<td>19</td>
</tr>
</tbody>
</table>

![Table 1: The result of survey for Ya-Ya](image)
In the same time of questionnaire survey, we also ask the students the reasons for their answers. Most of the students like the graphical user interfaces because the pictures are cute and attractive. However, the students of grade 1 think the screen is too small and the stylus is not easy to handled in there hands. But this seems not to be a problem for the third grade students. The common negative opinions for the most of the students are the instructions of the recognition result seem to be not so reasonable. Too many utterances are classified as unrecognized. The situation is extremely severely for the students of grade 1. This problem seems to be caused by the mismatch of the acoustic models with the target users. The acoustic model of VoCon 3200 was trained by the corpus of adult speakers, but the target users for this system are the students in the elementary school. This phenomenon was not occurred when the system was tested by college students. However, most of the students involved in this evaluation think the system is useful for their learning of pronunciation. They like to learn the pronunciation by themselves, anytime and anywhere, instead of learning in the classroom. They also feel it’s more comfortable to talk to the computer than their teachers. Therefore, most of the students would like to buy the system if it’s affordable.

6. Conclusions

According to the results of the initial evaluation, the confidence score of the students in elementary schools are extremely lower than the score of adults. Therefore, most of the students think that the criterion of the system is too strict for them. The mismatch in training corpus and target users makes the degradation in confidence score. The only solution is to have a recognition engine trained or adapted for children’s speech. It’s impossible to do this with the VoCon 3200 engine if the Children’s market is not the priority of ScanSoft. An alternative is to use the kid’s speech corpus of CSLU to train an acoustic model for Children. However, it also needs time to develop a recognition engine that can be run on PDA platform in real time. Therefore, it’s difficult for us to correct this problem in the near future.

It’s also a problem to use only pure text as the feedback to correct pronunciation. In real world, the teacher will give the instructions and demonstrate the manner of articulation to the students. Without the demonstration by real person, the students can’t completely understand how to adjust their articulators by a pure text instruction. A possible solution is to provide a 3D facial animation to show the difference of the position and movement of articulators between the normal pronunciation and the error one. There are some literatures in the study of using the talking head as a tutor to teach the pronunciation for normal persons or disabled persons [13]-[14]. These studies have shown encourage feedback in using talking head for pronunciation learning. Combined with the diagnostic results of speech recognition technologies in this study, the users can understand their pronunciation with more details. This is the target in the further development for Ya-Ya.

Acknowledgements

This research was supported by the National Science Council of Taiwan under operating grants NSC93-2213-E-130-002.

7. References