» Training English Pronunciation for Japanese Learners of English Online

Wang Shudong & Michael Higgins
Faculty of Engineering, Yamaguchi University, Japan

Yukiko Shima
Tokyo Science University in Yamaguchi, Japan

In this study, a compact but effective Internet-based support system was designed for Japanese English learners to improve their English pronunciation. The system provides several interactive methods for users not only to learn pronunciation from the sample data of native speakers, but also to discover and evaluate their own specific pronunciation problems, and then improve their pronunciation with the help of the system. The system is real-time, Internet based, and the pinpoint feedback from the system after it has analyzed the user's input sounds, will appear in human teacher-like natural language, rather than impersonal scores or abstract bar graphs which appear in other systems. Our present system focuses on pronunciations that are often problematic for Japanese learners of English. During the process of designing the system, we embedded HMM/DMM speech recognition and speech synthesis modules.

Background to build the system

In Japan, teaching English pronunciation is not a compulsory subject in schools. It is up to individual schools and English teachers to decide when, if, or how to teach pronunciation. English teachers with good pronunciation and adequate knowledge about how to teach English pronunciation may teach students IPA (International Phonetic Alphabetic system) or some other pronunciation system, while other teachers just use Katakana (a Japanese syllabary system) to read and write English pronunciation and many junior high school texts and dictionaries only have Katakana pronunciation guides (Sherard, 1986; Japan Science and Technology Agency, 1991). Also English pronunciation is seldom required in the entrance exams to colleges, so that students do not have the motivation to study English pronunciation. In 2004, we performed an experiment while collecting pronunciation data from Japanese university students. Twenty university students, mainly from the Chugoku and Kyushu areas of Japan were asked to read

The JALT CALL Journal

thirty words that contain /p/, /f/, /v/, /l/, /r/ and /z/ sound, and these IPA symbols. Only one student could read IPA symbols; none of the others could. Seventeen made many errors in pronunciation, while only three subjects were able to pronounce previously learned words accurately. Only one of the twenty was able to use his pronunciation knowledge to accurately pronounce a new word.

It is evident that English pronunciation is a big problem for many Japanese learners of English. For example, Japanese people tend to pronounce the sound /r/ as /l/, /v/ as /b/, /s/ as /sh/, /f/ as /h/, /th/ as /s/, and so on. Additionally, they tend to insert extra vowel sounds after final consonants so that /p/ becomes /pu/, /t/ becomes /to/, and /d/ becomes /do/, and so forth (Shima, 1986).

The English of many Japanese learners can not be understood by non-Japanese speakers because of either poor or Katakana pronunciation. Even though a few university English teachers are trying very hard to improve their students' English pronunciation, many times, the effects are limited. There are several reasons for this. First, Japanese students tend to be shy and sensitive; they are afraid of being corrected in front of the other people. If they are corrected a little more frequently than others, they will soon lose the interest in learning pronunciation. Second, private one–to-one pronunciation training, while the best way, is too costly. Third, pronunciation training is currently delayed beyond the age when children's tongues and mouth muscles are flexible. Therefore many easily correctable pronunciation problems become fossilized and difficult to correct by an English teacher who has limited interaction with the student, and who may also have her/his own pronunciation issues.

The above reasons gave us the motivation to build an online pronunciation training system which can allow Japanese users privately, repeatedly and at their own pace to develop their English pronunciation.

Introduction to the system

In the CAPT (Computer Assisted Pronunciation Training) field, many systems have employed speech recognition technologies (e.g., Eskenazi & Hansma, 1998; Delmonte, 2000; Mak et al., 2003). Even though, limited to the current immature technologies of speech recognition, some of the systems indeed were able to interact with users, but just with simple indicative graphs or scores to tell users whether their input sound is good or not, without any details of why their pronunciation or intonation was good or not good. Some other systems have not used ASR (Automatic Speech Recognition) technologies, but rather on speech signal processing techniques (Kawai & Hirose, 1998). These systems do work with some individual sounds and short words, but the feedback offered for the students are abstract waveforms and graphs which are not easy to understand for learners (Neri, Cucchiarina & Strik, 2003). Additionally, most pronunciation self-training systems to date either appear in a format of independent software or have to be used in a space-limited, LAN equipped multi-media classroom, rather than being made available on the Internet.

In designing our system, we have attempted to adopt the advantages of the other systems. In building acoustic modules, we only focused on the recent technological development of speech recognition, speech signal processing, speech synthesis and English corpora database. The biggest differences in our pronunciation training system from similar systems are: 1) Feedback from the system is real-time, detailed and human teacher-like natural language. 2) Pronunciation training is conducted online. Most of the programs will be run

on the server side, which lowers the hardware requirement of the users' computers. 3) The system is designed for Japanese English learners and so the focus is on pronunciation errors that commonly occur among Japanese English learners. 4) The database is compact. It just includes sample modules of all consonants (27 in British English, 24 in American English), vowels (20 in British English and 16 in American English), and about 2000 mini-pair words related to these basic consonant and vowels sounds (Neri, Cucchiarina & Strik, 2003). Mini-pair word resources were selected from the Sound Approach Workbook (Higgins & Higgins, 2001), whose authors have been teaching English in Japan from children to adult levels for more than 25 years. The reason we gave up the idea of auto-detecting the prosody errors of sentences input from users is that prosody detection is still beyond the current reach of speech recognition technology (Kirriemuir, 2003). 6) We employed not only ASR HMM/GMM modules, but also speech synthesis technology and some other speech processing technologies.

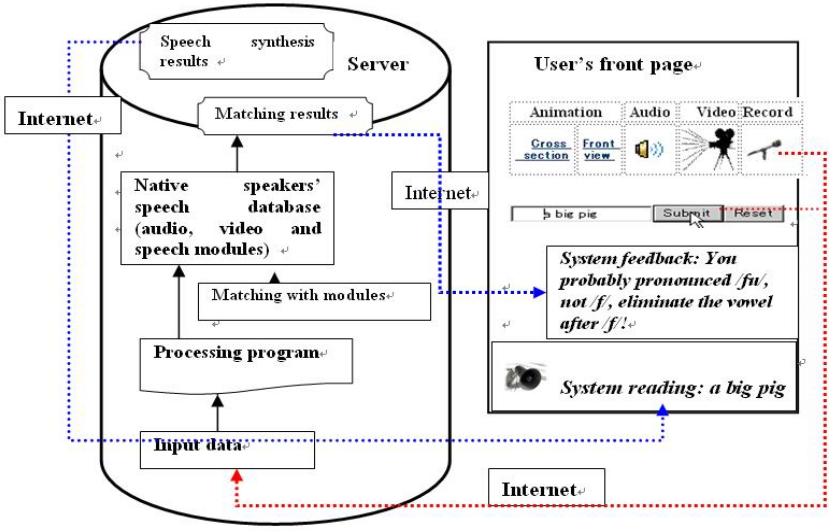

Figure 1: The system structure

Process of building up the system

1) Build-up native speakers' pronunciation database

We invited 10 (5 males and 5 females) native speakers from America (Canada) and Britain (Australia) who speak standard American/British English to read all of the vowels and consonants, as well as some other carefully selected representative words, including mini- pair words and phrases/ short sentences. We also made use of the online BNC (British National Corpus) and standard American English TIMIT corpus (CD-ROM). Please remember, this system is mainly designed for Japanese English learners, so the error detection will be

focused on the sound errors which Japanese often make. Hence, a set of Japanese-accented corpora data had to be collected. We made use of The NICT JLE Corpus published in 2004 with a CD-ROM distribution (National Institute of Information and Communications Technology, 2004).

For mini-pair words, we invited 20 university students to record 2000 mini-pair words. And the same 20 university students were asked to record the phrases and short expressions. The video images of native speaker's mouth movement and pronunciation audio data was also saved in a database. For some particular pronunciations like [r] and [ch], which are not easily expressed in either video or audio data, we used software applications like Flash MX and some other computer graphics techniques to show the articulating process. See Figure 2.

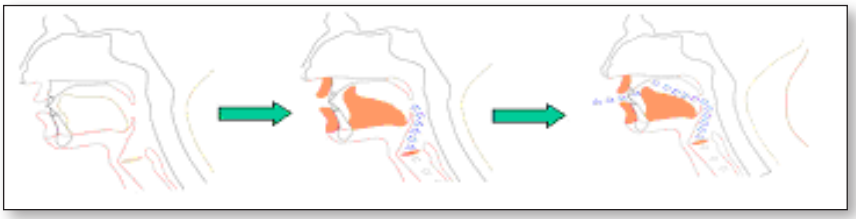

Figure 2: Flash animation to indicate how to pronounce sound /r/

2) Build-up acoustic modules

We used many kinds of speech signal processing technologies to analyze the native speakers' audio sample pronunciation data and extracted the most important and representative feature parameters such as power, amplitude, wavelength, frequency, time, velocity, and intensity (Kita, 1996), for each target sound/word/expression in order to build sound templates. The British (Australian) native speaker's data we collected, the online BNC, and the Japanese-accented English corpora were used together to develop Japanese-accented British English phoneme HMMs. The American (Canadian) native speakers' data and Standard American English Corpus materials, together with the Japanese people's English corpora were used to create Japanese-accented American English HMMs. Three kinds of modeling techniques were deployed: for individual pronunciations, Context–independent Modeling; Position-Dependant HMMs were used for the word which the target pronunciation/phoneme appears at the beginning or middle/or the end of the word; Discriminative Training Model was designed for recognizing word list phrases and short expressions. For each vowel, consonant and representative word, we have individual templates (HMM/GMM module) for both American and British English.

The reason we generated two types of template-based recognizers (American and British English) is so that the users can freely choose what kind of English pronunciation he/she wants to practice and compare their own pronunciation to.

3) Build up natural language feedback database

We wrote all of the possible feedback for each consonant and vowel and selected word, mini-pair words and short expressions in American/British English. This was a very labor in-

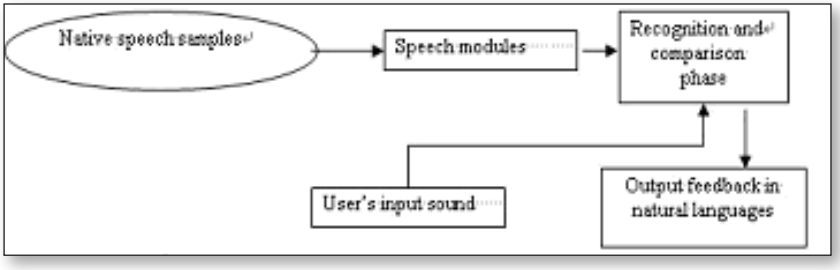

Figure 3: The users' input data is processed by the server

tensive process. For example, input from the user for the consonant [f] in American English, the feedback from the system would include:

a: "You seemed to have pronounced /hu/. Bite firmly against the lower lip with the upper teeth and breathe out from the lateral gaps. If you try to say the sound with the breath rising only from the chest, it will merely produce the weak Japanese sound, 'hu'". (This would happen when the system finds that the frequency and intensity are far from the average parameter of the model.)

b: "The system detected that your vocal cords vibrated. The breath which is required here does not rise from the chest but comes from the abdomen, controlled by the abdominal muscles." (This would be displayed when the system detects certain intensive power after /f/ sound.)

c: "You may have pronounced a completely different sound. Listen to or watch the native speaker's samples, or train yourself with the help of the flash animation." (This is the response when the system detects that all of the parameters of user's input sound are too far from the baseline of target model.)

d: "Either you pronounced /f/ too soft, or your lower lip is pulled too far into your mouth. Say it again." (This is the response the system gives when it detects that the frequency of users' input sound is too low in comparison to the model parameter.)

e: "Your pronunciation is quite good! Go ahead and practice the words that start with /f/ and with /f/ in the middle and at the end of the words." (This message pops-up when overall parameters of user's input sound are very well matched the model.)

In order to write the feedback, we have done many experiments and a lot of Japanese university students at Yamaguchi University of Japan were invited to pronounce sounds, selected words and short expressions. The "typical" errors we defined in this system were extracted not only from certain reference books, but also confirmed from our own experiment using a video camera. So they are confirmed common pronunciations errors of Japanese English learners.

All of the HMM/GMM speech recognition modules were set on our server side. The test for some limited words shows that the system can catch the most typical errors among Japanese people's English pronunciation.

4) Decide confidence base-line

For each HMM /GMM module, we have set up a different baseline. Confidence baseline decides which feedback in the database should be launched. We started the recognition accuracy from 40%~75% for the input Japanese-accented English pronunciation. For example, the recognitions baseline for /p/ will be set very low (40%), because, in any case, /p/ can be easily understood by a large margin from the native speaker's module by frequencies and intensity, as long as after /p/, there is no inserted vowel sound.

5) Write server-side interaction programs

Server-side interaction programs (Java +PHP) were written to enable users to record and send their pronunciation with the least manual work. In our system, when a user clicks on the record icon on the system home page, the recording device on his computer will be automatically started, and transfer the data when the recording is over.

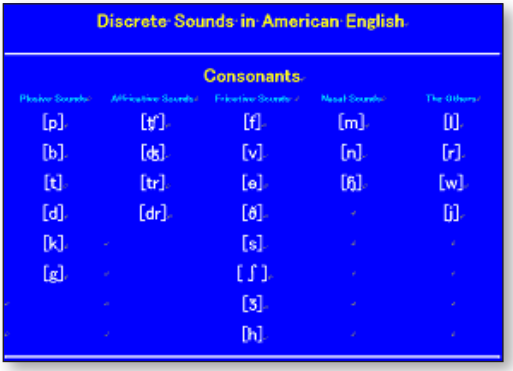

Figure 4: >From system top page to American pronunciation consonant sub-page

The system work flow

1) A user who wants to practice his/her pronunciation signs in to the system front page, where he will find he has to choose which kind of pronunciation he wants to learn or improve. If s/he clicks on the button of American English, the American sound index page will appear. Then s/he has to decide which sound he wants to practice, let us say, /p/. S/he then clicks on the /p/, and the lesson for /p/ pronunciation will appear. See Figures 4 and 5.

2) On the "p-lesson" page, a user can choose any the following interactive actions.

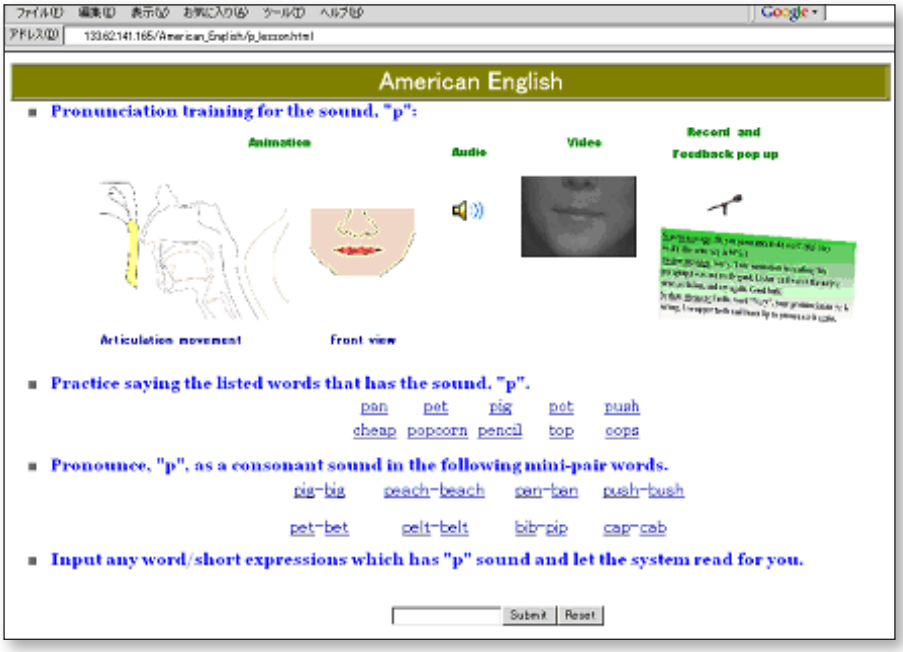

Figure 5: System user front page-"p- lesson" example

a) Click on the speaker, an audio icon, to listen to the native speaker's pronunciation.

b) Click on the image of video camera to download the video clip of native speaker reading the sound /p/ and the related words.

c) After listening to the sound and watching the video clip, if the user still does not know how to pronounce the sound, he can click on any of the flash animations which dynamically reflect the tongue's, jaw's, lip's, pharynx's…movement. In the case of /p/, they will see the process of pursing the lips firmly and then letting the air stream pop out as the lips open suddenly. On either the animation image of the articulation movement or the front view, the written pronunciation instructions in Japanese and English are provided. See Figure 6.

d) When a user wants to verify how well he can pronounce the sound /p/, he should click on the icon of the microphone on the front page. Then, automatically, his computer will start the record function (in the case of Windows 9x, 2000, Windows Me and XP). He will need a microphone to record the sound /p/, or the example words that have /p/, or any of the mini-pair words. The current system restricts their recordings to just the sound and the words listed on the page. After he finishes recording a discrete sound or word, again, automatically, his input data will be transferred to the server side and in a few seconds, a human-like feedback/evaluation will appear on the user's computer screen. With the help of the feedback, the user can decide the next learning step.

1. Look carefully at the computer screen showing how to pronounce the sound, "p".
2. If you let your lips open vertically, the vowel sound, /uh/, follows.
3. Like the image on the screen, hold a thin sheet of paper in front of your mouth. As you exhale it should flutter every time you pronounce the sound, "p".
4. However, the sheet will not flutter if you make only the vowel sound, "uh".

Figure 6: Language explanation and animation for instructing user how to pronounce /p/

e) If a user is not happy with just practicing the words listed by the system, then he can input any words or short expressions which include the sound he is practicing (in the case of this example, /p/) into the dialogue box, then click on the submit button, and in a few seconds, he will see another dialogue box. If he clicks on the play button on that image, he can hear the word read by the system in the clear Standard English owning to the speech synthetic program on the server side.

We understand that some users are visually oriented and they learn English by seeing it spoken, while some others are more aural, and they learn English by listening. And there are certain people who can learn English by reading instructions. Every type of English learner will find he or she has a way to connect with and learn from this system.

Conclusion

In sections 1 and 2, we introduced the background and purpose of designing this pronunciation training system, and what we believe to be some unique features of the system. In section 3, we presented the structure of the system. In section 4, we showed the process of building-up the system. In section 5, we indicated how to use the system from the point of view of the users.

This system provides a dynamic, real-time and interactive way for Japanese English learners to learn and improve their English pronunciation online. People who use this system can learn from samples of native English speaker's pronunciation by listening and watching. They can also find out the strengths and weakness of their own pronunciation in more detail with the help of this system than any other system. Even though a long-term evaluation needs to be done, our initial test with university students in Japan showed that the system will work very effectively with pronunciation training for general Japanese English learners.

References

Delmonte, R. (2000). SLIM prosodic automatic tools for self-learning instruction. *Speech Communication, 30,* 145-166.

Eskenazi, M., & Hansma, S. (1998). The fluency pronunciation trainer, Proc. *STiLL Workshop on Speech Technology in Language Learning, Marhollmen.*

Higgins, Marilyn, & Higgins, Michael. (2001). *The new sound approach workbook.* International Education Initiatives.

Japan Science and Technology Agency. (1991). *White Paper on the globalization of Japanese scientific efforts.* Reported in The Asahi Evening News, Tokyo: Asahi Publishing Company, October 11, 1991. Translated by Shima, Y. (1986). *The phonovisual method* (in Japanese). New Japan Books.

Kawai, G., & Hirose, K. (1998). A CALL system for teaching the duration and phone quality of Japanese Tokushuhaku. *Proceedings of the Joint Conference of the ICA (International Conference on Acoustics) and ASA (Acoustical Society of America),* 2981-2982.

Kirriemuir, J. (2003). *Speech recognition technologies.* The Joint Information Systems Committee. [Available http://www.jisc.ac.uk/uploaded_documents/tsw_03-03.pdf].

Kita, G. (1996). *Speech processing* (in Japanese), Morikita Publishing Co. Ltd.

Mak, B., Siu, M., Ng, M., Tam, Y.-C., Chan, Y.-C., Chan, K.-W., Leung, K.-Y., Ho, S., Chong, F.-H., Wong, J., & Lo, J. (2003). PLASER: Pronunciation Learning via Automatic Speech Recognition, *Proceedings of HLT-NAACL, Edmonton, Canada.* [Available http://acl.ldc.upenn.edu/W/W03/W03-0204.pdf].

National Institute of Information and Communications Technology. (2004). The NICT JLE Corpus (with CD-ROM).

Neri, A., Cucchiarina, C., & Strik, W. (2003). Automatic Speech Recognition for second language learning: How and why it actually works. *Proceedings of the 15th ICPhS, Barcelona,* 1157-1160.

Sherard, M. (1986). The effect of the Japanese kana filter on English phonotactics: Pedagogical considerations. *Doshisha Daigaku Eigo Eibungaku Kenkyu, 40.*

Shima, Y. (2004). *Look, listen and learn English pronunciations CD.* Tokyo: Tokyo Science University in Yamaguchi.