ON THE USE OF SPEECH RECOGNITION TECHNOLOGY FOR FOREIGN LANGUAGE PRONUNCIATION TEACHING

Keikichi Hirose*, Carlos T. Ishi** and Goh Kawai***

(* Dept. of Frontier Informatics, University of Tokyo

** Dept. of Information and Communication Eng., University of Tokyo

*** Department of Psychology, University of California Santa Cruz)

hirose@gavo.t.u-tokyo.ac.jp

Abstract

Recently speech technologies have shown notable advancements and they now play major roles in computer-aided language learning systems. In the current paper, use of speech recognition technologies is viewed with our system for teaching English pronunciation to Japanese speakers.

1. Introduction

Recent development of internationalization largely increased the situation where a person should speak in his/her non-mother tongues. For many foreigners, it is not an easy task to reach an affordable level of making conversations with native speakers. When a person is looking for a job in foreign countries, the situation is serious; his/her mental ability is sometimes ranked in lower level because of his/her accented pronunciation. In Japan, situation of language learning may be worse as compared to other countries; even all Japanese learn English at high schools, their speaking ability is still poor. One reason will be the lack of good teachers at their early stage of training. The situation may be also hard for foreigners living in Japan, and trying to learn Japanese. Although better language training is possible if we have enough "good" teachers, which will not be the case. If a device enabling the self-training of pronunciation skill for learners of foreign languages is available, especially for their early stage of training, the situation will be largely eased.

Technologies on spoken language processing, such as speech recognition and synthesis, have shown remarkable progresses in these several years. These technologies reached a level

enough to be utilized in language learning systems. Speech recognition can be utilized to assess learner's pronunciation, while speech synthesis may be utilized to generate speech with correct pronunciation in learner's voice quality. Therefore, a rather large number of computer-aided language learning (CALL) systems have already been developed. Some systems succeeded to keep learner's interests by well-planned scenarios and attractive visual displays. When they evaluate learner's speech, spectral distance measure or speech recognition is mostly utilized. Although these methods can score the learner's speech depending on how far it is from reference speech, the score does not necessarily correspond to the learner's pronunciation ability. These systems may score a native speaker's utterance as an accented pronunciation if his/her voice quality is largely different from that of reference speech. A scheme is necessary to avoid a learner to be forced to mimic teacher's voice quality. Furthermore, corrective feedback from the system is rather poor in almost all the systems. Usually, learners are only noticed that their pronunciation is wrong, with no information on how it is wrong. If they correct their pronunciation only with this notice, the pronunciation training will not be necessary for him from the first place.

From these considerations, the following three will arise as necessary functions, which a CALL system should have:

1) Clarify where and how a learner makes a mispronunciation,

2) Indicate whether the mispronunciation is acceptable to natives or not,

3) Show clearly how the learner can correct his/her pronunciation.

We are developing a CALL system for English and Japanese pronunciation training, which has these functions.

2. System Overview

Configuration of the CALL system we are developing is schematically shown in Fig. 1. Stored speech data by teachers were first segmented into phones through recognition using phone hidden Markov models (HMMs). Different from the case of normal speech recognition, transcription of input speech is usually given beforehand for CALL systems and is utilized in the matching process with HMMs. This process is usually called "forced alignment." Fundamental frequencies (F0s) are also extracted and are utilized for training accent and intonation. These process may be done online or offline. For the latter case, the analysis

results are stored and are utilized when they are necessary. Manual correction is possible for the analysis results before storing. The segmentation and F0 extraction are also conducted in the same way for learners' speech. Learners' speech should be processed online and manual correction will not be allowed. Analysis results for learners' speech are compared with those for teachers' speech to evaluate learners' speech. Corrective feedback is then generated based on the evaluation. Evaluation results are also utilized to correct learners' speech through the analysis-synthesis process.

Mispronunciations the system currently can deal with are:

1) Phone substitutions, insertions and deletions for both English and Japanese [1, 2],

2) Improper duration control in Japanese double-mora phonemes and their normal phoneme counterparts [3],

3) Accent type errors in Japanese and erroneous stress assignment in English [4].

In the current paper, discussions are mainly on the first type errors, whose detection is tightly related to speech recognition.

3. Detection of Mispronunciation Using Speech Recognition Technologies

In early experiments, nonnative pronunciation was evaluated as its distance from native pronunciation in acoustic feature space, such as Cepstrum coefficients. Although the distances can be used directly as indices of learner's pronunciation quality, two problems are included;

1) Larger distances do not necessarily indicate poorer pronunciation quality. They may only indicate it is rare as native pronunciation.

2) Usually, learners and teachers have no knowledge on how they should control articulators to correct acoustic features.

Systemic, structural and realizational differences between L1 (learner's native language) and L2 (target language, the learner is learning) appear as phone substitutions, insertions and deletions. In several CALL systems, speech recognizers are designed to detect these mispronunciations and to assess learner's speech: mispronunciations when the recognizer mis-recognizes. Recently, HMM-based speech recognition became a standard technology and was often utilized in CALL systems. In many systems, HMMs for L2 trained on native speakers of L2 are utilized [5]. However, these systems still cannot distinguish nonnative speakers to

nonstandard native speakers, though they use speaker adaptation to accommodate the learner's speech. Phonetic and phonological effects of L1 are not taken into account. Training L2 HMMs on nonnative speech is a possibility to solve this problem. By doing so, learners' pronunciation patterns can be modeled more accurately. The problem will be collecting training speech data from a sizeable number of learners, especially when the data are to be stratified according to the speaker's pronunciation ability. Furthermore, we have another problem. When HMM-based recognition is used, regardless of whether the HMMs are trained on L2 native speech or L2 nonnative speech, we need a scale to translate statistics into scores of pronunciation quality. Although it is obtainable by relating statistics with human judgements, no reliable one is possible because of rather large variations in human judgements.

By contrast, we developed a method to measure the pronunciation quality by using both L1 and L2 phone HMMs, which are trained separately for each language on native speakers' speech data [1, 2]. Collecting speech data from natives usually is not a hard task; they or even trained phone HMMs already exist for major languages. Learners may substitute correct L2 phones with incorrect L1 phones, especially when they have no correct L2 phone in their L1 phone inventory. Furthermore, learners carry over L1 phonotactics to L2 production. When phonotactics of L1 and L2 are different, phone insertions and deletions occur. These substitution, insertion and deletion errors can be detected through speech recognition using L1 and L2 phone models. Since HMMs for L1 and L2 are speaker independent, effects of learner's individuality on speech recognition are cancelled between L1 and L2 HMMs. As for the scoring of pronunciation quality, percentage of actual to possible errors of phone substitution, insertion and deletion is introduced, making the process of relating statistics with human judgements not necessary.

4. Developed System of Phone Quality Detection

Two types of language learning are possible; one with reading material and the other without. Phone transcriptions of learners' speech are available for the first type, which is usually the case of CALL systems. Therefore, the core of our system is the bilingual phone recognizer, which align learner's speech to given phone sequences using L1 and L2 HMMs. More accurate results are possible than normal speech recognition process. The speech

recognizer used in the current system is HTK v2.1 [6]. The learner's speech is recorded through desktop microphone and sampled in 8 bits at 16 kHz. Feature parameters for recognition consist of 12th-order melcepstra, their deltas and delta-deltas, delta power and delta-delta power. The following Japanese and English phone HMMs trained elsewhere [7, 8] were used:

1) English phones (45 phones):

aa ae ah ao aw ax axr ay b ch d dh eh el em en er ey f g hh ih ix iy jh k l m n ng ow oy p r s sh t th uh uw v w y z zh

2) Japanese phones (40 phones):

sp N a a: b by ch d e e: f g gy h hy i i: j k ky m my n ny o o: p py q R Ry s sh t ts u u: w y z

In order to accommodate pronunciation variability by non-natives, mono-phone models were adopted instead of tri-phone models, which might yield better results in normal speech recognition process. As shown in Fig. 2, L1 phone HMMs and L2 phone HMMs are combined during recognition (forced alignment) of learner's speech. In order to make forced alignment possible, a phone network (lattice) covering all the possible mispronunciation by learners was constructed manually for each item of the reading material. Figure 3 shows an example of network for English word "sports" when learners are Japanese.

5. System for Teaching English Pronunciation to Japanese

5.1 Vowel insertions

Although the developed system can teach Japanese speakers English pronunciation and vise versa, the first case is explained in the current paper. When learning English, Japanese speakers frequently insert vowels within consonant clusters or after syllable-final consonants. Typical examples are observable in loan words, where Japanese phonotactics are carried over to English. When they try not to insert vowels, they even delete syllable-final consonants. Since anaptyxis multilates the syllable and stress structure of English, anaptyctic speech is incomprehensible to native speakers of English even after considerable exposure to Japanese accented speech. However, Japanese teachers of English overlook anaptyxis because they understand anaptyctic speech completely. This misunderstanding causes undesirable situation in English training in Japan. Rules of vowel insertion by Japanese speakers can be summarized as follows:

1) Syllable final stops and affricates are converted to mora obstruent and a vowel is attached. Example: up.

2) Vowel [o] is inserted after [t] and [d] sounds. Examples: kit, advantage.

3) Vowel [i] is inserted after palatal affricates. Examples: match, edge.

4) Vowel [u] is inserted after consonants other than the above. Syllable final [n] sound is converted to mora nasal [N] without vowel insertion. Examples: spice, dam, spoon.

5) Vowel [i] is inserted instead of [u], when loan was done in ancient periods, say before 19th century. Examples: stick, excite.

Words, short sentences, and sentences including loan words were utilized as the reading material. Some examples are listed below:

1) Words: touch, but, class, drama, extra, etc.

2) Short sentences: Please pay promptly, The trains are filmed in the Alps, etc.

3) I have an [atlas] and [album] at home, etc.

Phone networks for these items are constructed based on the vowel insertion rules. Inserted vowels are those of Japanese. For original sounds, substitutions by Japanese sounds are included in the networks. Score of learner's speech is defined as $(n\text{-}m)/n^{*}100$ [%], where n and m respectively indicate maximum number of possible vowel insertions and number of actual vowel insertions. Figure 4 shows an example of feedback display.

An evaluation experiment of the system was conducted for 16 Japanese university-students (14 males and 2 females). A native speaker of English with ample knowledge on Japanese pronunciation of English also scored their utterances manually. Scores from the system and from the native speaker were compared. The maximum correlation 0.81 was obtained when the pruning threshold was set to 70. Figure 5 is the scattering chart with the regression line.

5.2 Phone substitutions, insertions and deletions

Although major mispronunciations of Japanese speakers are vowel insertions, deletions and substitutions frequently occur in the pronunciation training. For instance, syllable final [r] sound is deleted (far-> fa) and [θ] sound is substituted to [sh] or [s] sound (think->shink). We should note that substitutions, insertions and deletions occur differently when Japanese

speakers learn English and when English speakers learn Japanese. For example, Japanese word "shingu (bedclothes) [sh][i][N][g][u]" will be mispronounced as [sh][i][ng] by English speakers, and English word "thing [θ][i][ng]" will be mispronounced as [sh or s][i][N][g][u] by Japanese speakers.

Based on the above considerations, phone networks were constructed for sentences in the reading material, such as "Good morning," "School starts the first week of April," "The library is next to the nice Japanese garden in the park," and so on. The scoring was done in the similar way as the case of vowel insertions. Figure 6 shows an example of feedback display.

6. Conclusion

After discussions on how speech technology can be applied to CALL systems, our CALL system for training English pronunciation by Japanese speakers is explained. It uses a bilingual phone recognizer to capture systemic differences (how L1 or L2 phones are substituted for novel L2 phones), structural differences (L1 phonotactics carrying over to L2 production), and realization differences (how similar phones in L1 and L2 can be uttered with different phonetic realization). Implementing our method is straightforward, because it uses only native speech of L1 and L2 to train phone models, which can be by-products of regular L1 and L2 speech recognizers. Although potential validity of our system for pronunciation training was shown, collaboration with language teachers is crucial for the further development.

Acknowledgements

We thank Kazuya Takeda for providing us with Japanese HMMs [7], and Steve Young for providing American English HMMs [8].

<References>

[1] Kawai, G. and Hirose, K., "A method of measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training," Proc. ICSLP, Sydney, pp.1823-1826 (1998).

[2] Kawai, G. et. al., "A Call system for correcting vowel insertions in English spoken by native speakers of Japanese," Proc. ICPhS, San Francisco, pp.767-770 (1999).

[3] Kawai, G. and Hirose, K., "Teaching the pronunciation of Japanese double-mora

phonemes using speech recognition technology," Speech Communication, Vol.30, Nos.2-3, pp.131-143 (2000).

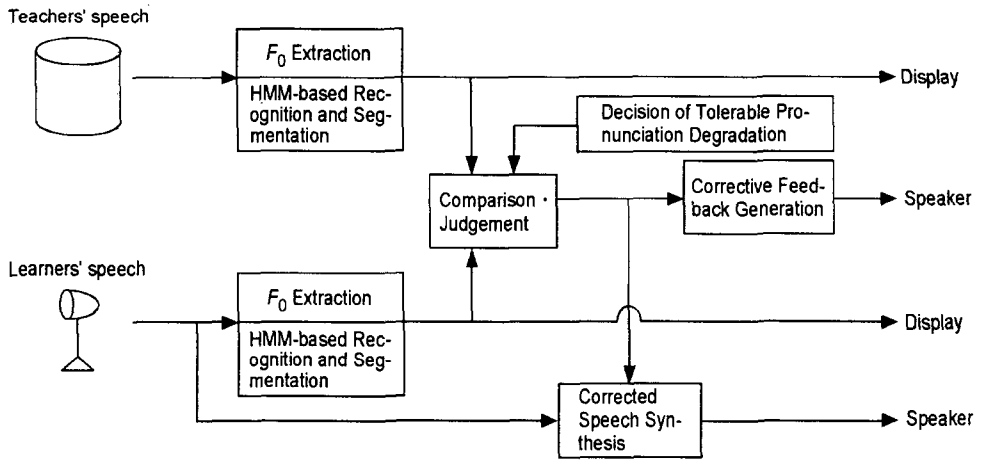
[4] Goh Kawai and Carlos T. Ishi, "A system for learning the pronunciation of Japanese pitch accent," Proc. EUROSPECH, Budapest, Vol.1, S1.PO3.2, pp.177-180 (1999-9).

[5] Witt, S. et. al., "Language learning based on non-native speech recognition," Proc EUROSPEECH, Rhodes, pp.633-636 (1997).

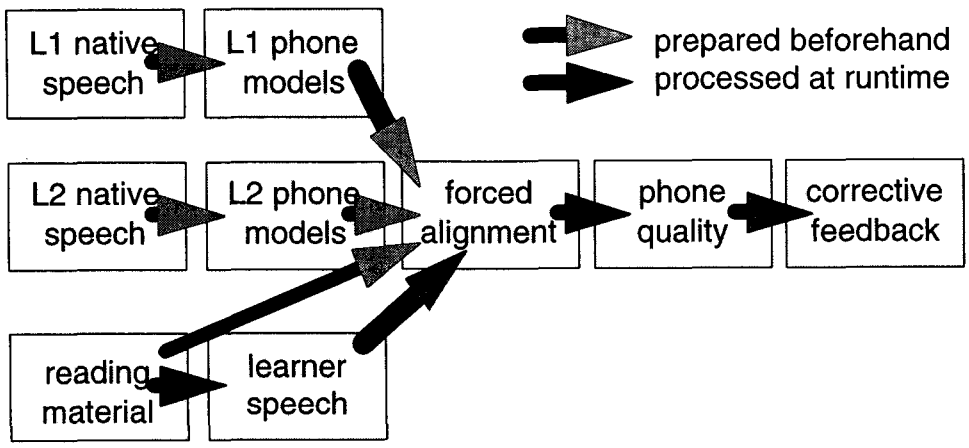
[6] Young, S. et. al., "The HTK book for version 2.1," Cambridge University (1997).

[7] Takeda, K. et. al., "Common platform of Japanese large vocabulary continuous speech recognition research: construction of acoustic model," Note of Spoken Language Information Processing Group, Information Processing Society, 97-SLP-18-3 (1997).

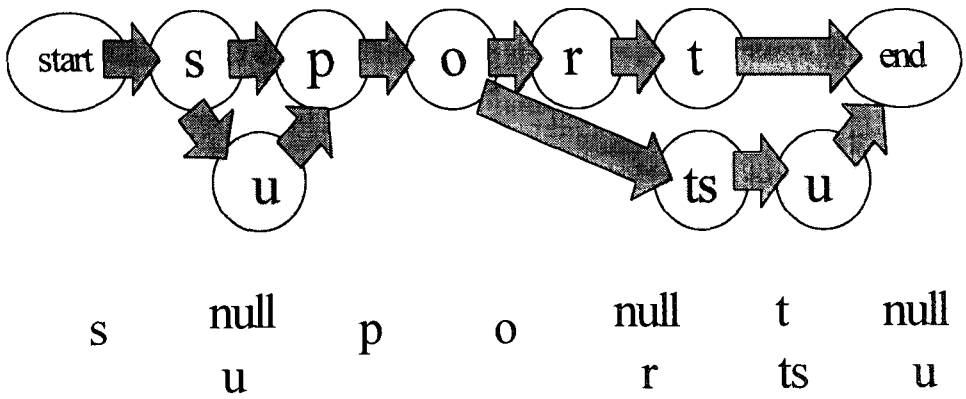
[8] Woodland, P. et. al., "The HTK large vocabulary recognition system for the 1995 ARPA H3 task," Proc. ARPA CSR Workshop, Arden House (1996).



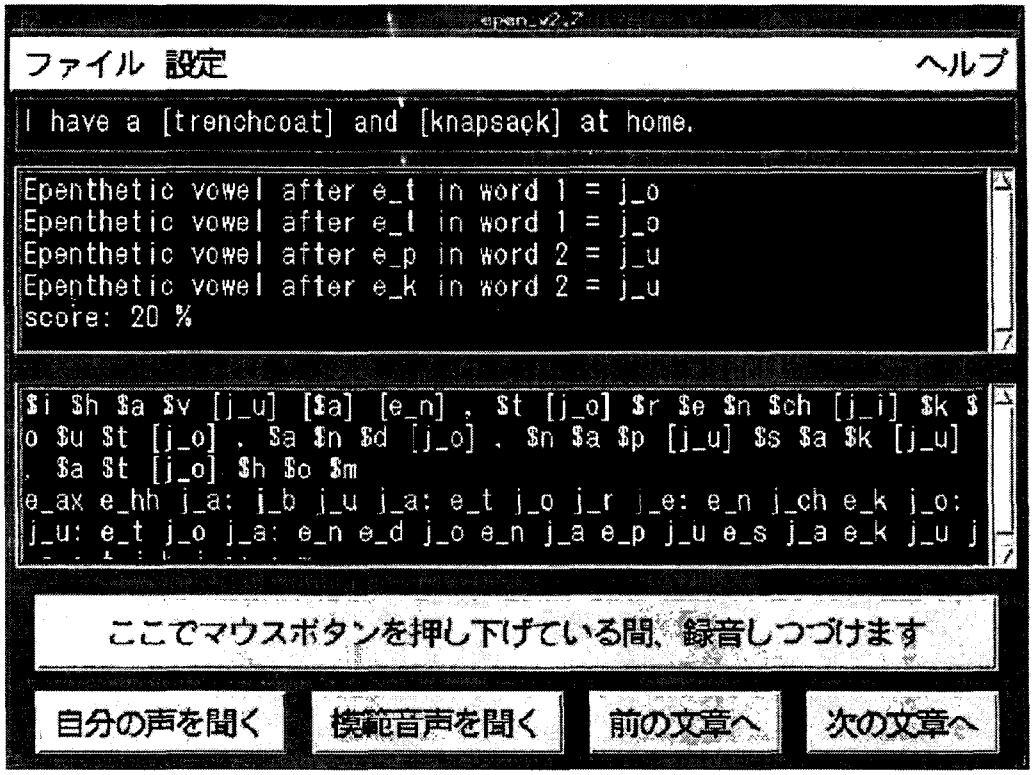
<Figure 1> Configuration of the CALL system under development.



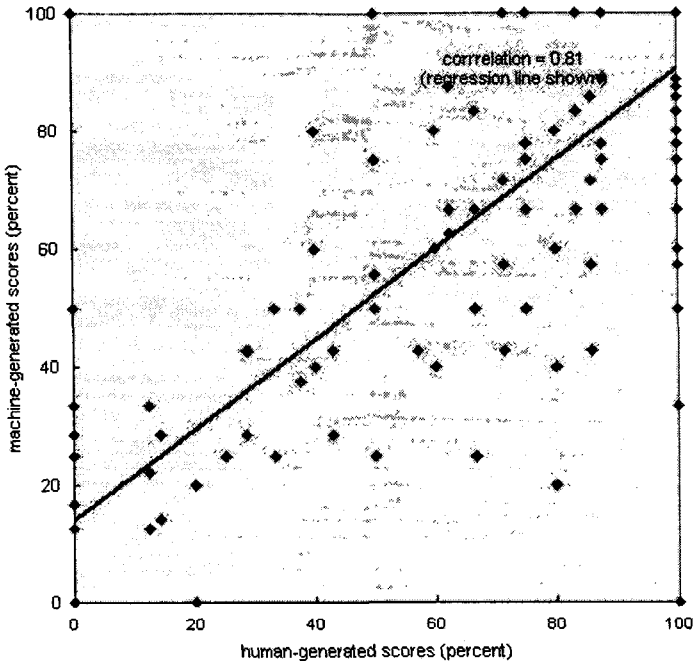
<Figure.2> Process flow of the system based on the bilingual phone recognizer. L1 and L2 HMMs are separately trained on native speech and combined during forced alignment. The learner receives categorical articulatory advice on phone quality.



<Figure 3> Phone network for English word "sport" showing possible pronunciation errors by Japanese speakers. The lower panel indicates phone lattice, where insertions and deletions are represented by null phones.



<Figure 4> An example of feedback display for mispronunciation of vowel insertions. The learner said "trenchcoat" with a Japanese-accented [o] after the [t] sounds at the beginning and end of the word, and "knapsack" with a Japanese-accented [u] after [p] and [k] sounds. As "trenchcoat" and "knapsack" have five possible locations of vowel insertions, the system returns a score of 20 % correct. The lower-half feedback window shows a phone pronunciation network (with anaptyctic vowels shown in [brackets] and English/Japanese phone combinations symbolized with $ signs) plus the actual recognized phones (English phones are shown prefixed "e_" and Japanese phones "j_").



<Figure 5> Best-case match of human and system-generated scores obtained by adjusting the speech recognizer's pruning threshold to 70. Correlation between human and system is 0.81.

<Figure.6> An example of feedback display for substitutions, insertions and deletions. The system asked the learner to say "There are many birds and animals in Australia." No insertion errors were found in this utterance.