A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training

Goh Kawai and Keikichi Hirose

University of Tokyo, Department of Information and Communication Engineering

email: goh@kawai.com hirose@gavo.t.u–tokyo.ac.jp

http://www.kawai.com/ http://www.gavo.t.u–tokyo.ac.jp/

ABSTRACT

The problem addressed is automatically detecting, measuring and correcting nonnative pronunciation characteristics (so-called "foreign accents") in foreign language speech. Systemic, structural and realizational differences between L1 (native language) and L2 (target language) appear as phone insertions, deletions and substitutions. A bilingual phone recognizer using native-trained acoustic models of the learner's L1 and L2 was developed to identify insertions, deletions and substitutions of L2 phones. Recognition results are combined with knowledge of phonetics, phonology and pedagogy to show learners which phones were mispronounced and to instruct how to modify their articulatory gestures for more native-sounding speech. The degree of the learner's foreign accent is measured based on the number of alternate pronunciations the learner uses; the number decreases as learning progresses. Evaluation experiments using Japanese and American English indicate that the system is an effective component technology for computer-aided pronunciation learning.

1. INTRODUCTION

Acquiring nativelike pronunciation ranks first in desirability among foreign language learners. Unfortunately most adult learners develop fossilized pronunciations that are distinctively nonnative. Research suggests that some adults can attain close-to-native pronunciation through intense training [1]. In order to assist teachers overburdened with large classes, computerized self-learning systems that handle the repetitive tasks of pronunciation teaching are necessary. Using CALL (computer-aided language learning) systems to teach many students individually in parallel may help learners rectify pronunciation errors. Component technologies for such CALL systems include speech recognizers designed for nonnative language pronunciation assessment.

Previous work in this area includes measuring nonnativeness using statistics obtained from speech recognizers using HMMs (hidden Markov models). Some systems use HMMs for L2 trained on native speakers of L2 (e.g. [6]). These systems often use speaker-adaptation to accommodate the learner's speech. However these systems do not distinguish nonnative speakers from nonstandard native speakers, because the phonetic and phonological effects of L1 are ignored. Training L2 HMMs on nonnative speech is a possibility. Such systems have the potential advantage of modeling the learners' pronunciation patterns more accurately. Collecting training speech data from a sizeable number of learners may be a practical problem, especially when the training speech data is to be stratified according to the speaker's pronunciation ability. Regardless of whether HMMs are trained on L2-natives or L2-nonnatives, the primary challenge of using HMM-derived statistics is in translating the statistics into articulatory phonetic terms. Many pronunciation scoring systems measure the reliability of their scores by correlating them with human judgements (e.g. [5]). Problems of this method are that human judgements do not necessarily correlate highly among themselves, and that interhuman correlations set an upper bound on performance beyond which higher reliability cannot be proven.

By contrast, we use a method that automatically measures the pronunciation quality of phones produced by nonnative talkers by using a speech recognizer incorporating native phone models of both L1 and L2. A similar concept was proposed in [3]. HMMs for L1 and L2 are trained separately on language-dependent native-speaker speech data but are bundled together during recognition. As the system is a speaker-independent bilingual phone recognizer, the physiological aspects of the learner's speech are cancelled between L1 and L2 HMMs. Instead of using a single continuous variable to measure pronunciation quality (e.g. phone duration in [2]), this system relies on phone-based categorical recognition results to determine how a phone was articulated. Prior knowledge of L1 and L2 phonetics, phonology and language pedagogy are combined to identify nonnative articulatory gestures that result in pronunciation errors. The system detects errors in the choice of phones, reports the degree of nonnativeness of the learner's pronunciation, and suggests ways to improve speaking ability.

This paper explains details of this method's implementation along with results of feasibility experiments. Our technique can be applied to any language pair by using appropriate knowledge of phonology and acoustic phone models. Some of our experiments use American English as L1 and Japanese as L2; other experiments use these two languages the other way around.

2. SYSTEM STRUCTURE

The system's core is a speech recognizer running in forced-alignment mode (i.e., phone labels and boundaries with respect to the beginning of the utterance are obtained given a correct transcription of the utterance). The speech recognizer used is HTK v2.1 [8]. The learners' speech is recorded via desktop microphone and sampled in 8 bits at 16 kHz. Feature vectors consist of 12th-order melcepstra, their deltas and delta-deltas, delta-power and delta-delta power.

The system uses HMMs of L1 and L2 that share the same HMM typology but are trained separately on native speech. The phones in the HMM sets we used are listed in Table 1.

Table 1: Phones used in the combined HMM set.

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
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

The HMMs can be slightly broad monophones to accommodate pronunciation variability by nonnatives. The HMMs of L1 and L2 are combined during recognition (figure 1). There is exactly one recognizer active. The language model consists of monophones of both languages. The learner's speech is recognized phone by phone using a pronunciation lattice of L1 and L2 phones (figure 2). Changing the pronunciation lattice switches the system between a recognizer for teaching Japanese to American English speakers and vice versa.

Two systems for teaching phone quality were implemented: one for detecting phone substitutions (section 3) and another for detecting insertions and deletions (section 4).

3. PHONE SUBSTITUTION

3.1. Interlanguage allophones

Learners sometimes substitute correct L2 phones with incorrect phones from L1 or L2 when the correct L2 phone is absent from the L1 phone inventory. We prototyped a bilingual phone recognizer to detect and correct such mispronunciations. The system displays an L2 sentence on the computer screen and instructs the learner to read the sentence aloud. The sentences are designed to include L2 pronunciation mistakes commonly found among L1 speakers. Figure 3 shows the system's GUI (graphical user interface).

Recognizing the learner's phone allows us to estimate the learner's manner of articulation and to suggest a remedy. Maximizing the use of knowledge of L1 and L2 phonetics and phonology is essential for accurate estimation because markedly different articulatory gestures sometimes yield the same acoustic signal.

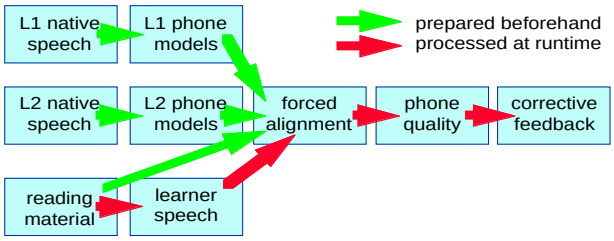

Figure 1: Process flow of bilingual phone recognizer. L1 and L2 HMMs are separately on native speech but combined during recognition. The learner receives categorical articulatory advice on phone quality.

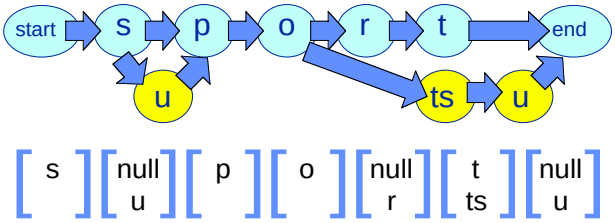

Figure 2: Phone network for the English word "sport" showing possible pronunciation errors by Japanese learners. Above: insertion, substitution and deletion of phones are shown as alternative paths branching off the main path. Below: phone lattice corresponding to the network. Insertions and deletions are represented by null phones.

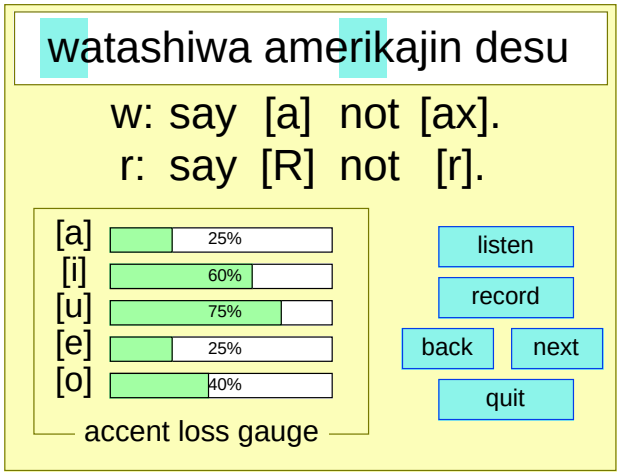

Figure 3: Graphical user interface (GUI) of phone–quality system. The learner receives categorical articulatory advice on phone quality for each mispronounced phone, and an "accent loss gauge" indicating allophone reduction ratios. For explanation purposes, the figure is in English and shows only vowels. The GUI is written in Tcl/Tk.

A language model is designed by grouping L2 phones and L1 phones that are often erroneously substituted with L2 phones. For instance, to capture a common mistake of English speakers diphthongizing Japanese vowels, we use phone lattices such as:

b_a = { j_a, j_a:, e_aa, e_ae, e_ah, e_aw, e_ay, e_ax, e_axr }

b_e = { j_e, j_e:, e_eh, e_er, e_ey, e_ax, e_axr }

(The prefix "j_" means the phone is a Japanese phone, "e_" means it is an English phone, and "b_" is the label for the bilingual lattice.)

These bilingual phone lattices can be considered as interlanguage allophone sets. Interlanguage allophones are designed for all L2 phones according to pedagogical knowledge. The language model for forced alignment is written using interlanguage phonemes in place of phones; for example, the word "go" is written as "b_g b_o".

As the learner's pronunciation improves over the course of practice, his choice of interlanguage allophones will become closer and closer to correct L2 phones. We hope ultimately they become identical. The process of the learner losing his "foreign accent" might be described as the shrinking process of his interlanguage allophone sets. An American English speaker learning Japanese might cease to substitute [ax] for /a/, for instance, reducing by one the number of elements in his interlanguage /a/ phoneme. We can measure the learner's accent loss by measuring the extent of allophone set reduction:

$$\text{allophone reduction ratio} = \frac{n_{start} - n_{current}}{n_{start} - n_{correct}} \times 100$$

n_{start} : Number of allophones at start of training, including both correct L2 phones and incorrect ("accented") L1 phones.

$n_{current}$: Number of allophones currently in learner's language model. Decreases as pronunciation ability becomes increasingly consistent.

$n_{correct}$: Number of correct target phones. Typically 1, but occasionally several.

As the learner progresses in his training, his allophone reduction ratio will increase from 0 percent (totally nonnative) to 100 percent (totally native), indicating that he has met his goal for that L2 phone. The system's GUI displays allophone reduction ratios as progress gauges labeled "accent loss gauge" (figure 3).

3.2. Evaluation experiments

An experiment designed to measure the accuracy of the system was run under simulated conditions. 340 nonnative phones were studied using American English as L1 and Japanese for L2. Results show that (a) 84 percent of the learner's phones flagged by the system as having an L1 accent influenced by L1 phone *x* were judged by a human native L2 listener as indeed being influenced by phone *x*, and (b) 91 percent of phones judged by the system as sounding perfectly L2 were judged by the human native listener as indeed being free of an L1 accent. Overall, the native listener agreed with the system 88 percent of the time.

This experiment was repeated using Japanese for L1 and American English for L2. Out of a total of 391 nonnative phones, a native L2 listener agreed that (a) for 81 percent of the time, phones flagged by the system as being influenced by L1 phone *x* were in fact influenced by phone *x*, and (b) for 95 percent of the time, phones declared by the system as accent-free did in fact sound perfectly native. Overall, the native listener agreed with the system 89 percent of the time.

These results suggest that our system is a useful component technology for foreign language pronunciation teaching. Future work includes determining the learner's mispronunciation habits by identifying tendencies in his incorrect interlanguage allophones (for instance, if a learner tends to palatalize, his habit will appear as his preference towards palatal sounds). Another task is adding allophones to the language model after they have been removed. In the current system, once an interlanguage allophone is removed, it is never restored. Capability to do so may be necessary for true dynamic modeling of the learner's pronunciation behavior, such as when the learner pronounces phones inconsistently compared to previous practice sessions.

4. PHONE INSERTION AND DELETION

4.1. Phonotactic effects

Structural differences between L1 and L2 (such as L1 phonotactics carrying over to L2 production) can result in the insertion and deletion of phones. For example, epenthesis (specifically apaptyxis) is found in Japanese speakers learning English. Inserting vowels within consonants clusters or after syllable-final consonants is particularly frequent. Since epenthesis mutilates the syllable and stress structure of English, epenthetic speech is incomprehensible to native speakers of English even after considerable exposure to Japanese-accented speech. An example of phone deletion among Japanese speakers learning English is the deletion of syllable-final liquids.

Using Japanese-accented English as an example, we prototyped a system for automatically detecting phone insertion and deletion. The learner's speech is recognized phone by phone using a pronunciation lattice including optional phones where insertion or deletion may occur (figure 2). The system alerts the learner whenever phones are inserted or deleted (figure 4).

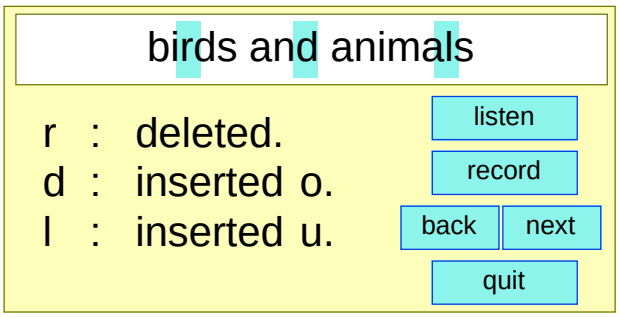

Figure 4: GUI of system for detecting phone insertion and deletion. The GUI shows where phones were inserted or deleted by highlighting their locations. The learner's task is to avoid insertions and deletions. The GUI is written in Tcl/Tk.

4.2. Evaluation experiment

7 native speakers of Japanese read English text consisting of 12 isolated words and 11 sentences for a total of 85 words, 105 syllables and 324 phones per speaker. The subjects' pronunciations were checked by the system and two native English speakers for phone insertion and deletion.

To evaluate the system's performance, error rates for (a) phone-level speech recognition and (b) insertion and deletion detection were calculated by comparing the system's recognition results with hand-labeled phone sequences. Phone recognition error rate is the number of recognition errors in the recognizer output (as compared with hand-labeled transcriptions) divided by the total number of underlying correct phones. Detection error rate is the sum of insertions and deletions the system misidentified (including false alarms) divided by the number of possible locations of insertions and deletions in the reading material.

For isolated words, the overall phone recognition error rate was 2 percent, and the detection error rate was 1 percent. For sentences, phone recognition error rate was 1 percent, and detection error rate was 2 percent.

Experimental results suggest that our system is a useful component technology for foreign language pronunciation teaching. This system can be improved by measuring the duration of fricatives such as [sh] [s] that become geminates in epenthetic contexts. For example, in the word "push" [p u sh: u], gemination on the word-final obstruent [sh] adds a mora; this geminate mora plus the vowel added to the end of the word results in a 3-mora pronunciation of a 1-mora word.

5. CONCLUSION

A bilingual phone recognizer was used 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 realizational differences (how similar phones in L1 and L2 can be uttered with different phonetic realizations). Identifying L2 phones being substituted by L1 phones shows how the L2 phone was articulated, and knowing the articulatory differences between the L1 and L2 phones allows us to provide the learner with feedback similar to that given by language teachers. Experiments indicate the method is a useful component technology for computer-aided pronunciation learning.

Implementing our method is straightforward because it uses only native speech of L1 and L2 to train acoustic models. HMMs used in this method can be by-products of regular speech recognizers for L1 and L2 speakers. Our system can expect strong support from language teachers because it involves their pedagogical expertise from the beginning of system development (as opposed to delivering the final product to teachers who are seeing it for the first time and are not predisposed to using it). Collaboration between speech engineers and language teachers is crucial to successfully deploying speech-enabled CALL systems in the field.

6. ACKNOWLEDGMENTS

We thank Kazuya Takeda for providing us with Japanese HMMs [4], and Steve Young for providing us with American English HMMs [7].

7. REFERENCES

[1] Bongaerts, T. et al "Age and ultimate attainment in the pronunciation of a foreign language." Studies in Second Language Acquisition, 19(4)447-465, 1997

[2] Kawai, G. et al "A CALL system using speech recognition to teach the pronunciation of Japanese tokushuhaku." Proc. STiLL (Marholmen, Sweden), pp 73-76, 1998

[3] Rooney, E. et al "Training consonants in a computer-aided system for pronunciation teaching." Proc. Eurospeech (Berlin, Germany), pp 561-564, 1993

[4] Takeda, K. et al "Common platform of Japanese large vocabulary continuous speech recognition research: construction of acoustic model." IPSJ SIG Notes, 97-SLP-18-3, 1997

[5] Townshend, B. et al "Estimation of spoken language proficiency." Proc. STiLL (Marholmen, Sweden), pp 179-182, 1998

[6] Witt, S. et al "Performance measures for phone-level pronunciation teaching in CALL." Proc. STiLL (Marholmen, Sweden), pp 99-102, 1998

[7] Woodland, P. et al "The HTK Large Vocabulary Recognition System for the 1995 ARPA H3 Task." Proc. ARPA CSR Workshop (Arden House), 1996

[8] Young, S. et al "The HTK Book", v2.1, Cambridge University, 1997