RECOGNITION AND VERIFICATION OF ENGLISH BY JAPANESE STUDENTS FOR COMPUTER-ASSISTED LANGUAGE LEARNING SYSTEM

Yasushi Tsubota † *Tatsuya Kawahara* † *Masatake Dantsuji* ‡

† School of Informatics, Kyoto University
‡ Center for Information and Multimedia Studies, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

We address methods for recognizing English spoken by Japanese students as the basis for our Computer-Assisted Language Learning (CALL) system. For automatic phonemic error detection, pronunciation error prediction is executed for a given orthographic text. To improve reliability, speaker adaptation and segment-input pair-wise verification are applied as pre-processing and post-processing, respectively. We also address acoustic modeling as a means for coping with the large acoustic variation seen in non-native speech. First, English acoustic models are trained using a database of English spoken by Japanese students. Japanese phonemes that are regarded as allophones of English phonemes are then incorporated. We present the results of experimental comparison of these models and confirm the effectiveness of speaker adaptation and pair-wise verification.

1. INTRODUCTION

We are developing a Computer-Assisted Language Learning (CALL) system for Japanese students learning the English language. As English has a much larger phonemic inventory than Japanese, Japanese students have to discriminate phonemes that are not used in Japanese, such as /l/ and /r/, /aa/ and /ae/. Since discrimination of these is often critical in mastering English, the primary goal of this study is to address automatic detection of phoneme pronunciation errors. In order to have effective communication, it is vital to pronounce phonemes correctly even though a speaker's English pronunciation may be marked by his/her own native accent. Here "correctly" means that pronounced phonemes are not confused with other English phonemes.

With respect to pronunciation learning systems[1][2], a number of researchers have been using speech recognition techniques. In order to recognize a Japanese student's English and at the same time detect his/her pronunciation errors, we make use of linguistic constraints since the phoneme recognition accuracy even for the native speaker of English is around 60%. Specifically, we implement pronunciation error prediction for a given orthographic text.

Japanese accented pronunciation is acoustically different from native speakers' speech and does not correspond well with a native English acoustic model. Thus, we train acoustic models using a database of English spoken by Japanese students. Moreover, we incorporate Japanese phoneme HMMs by adding these entries as allophones of the corresponding English phonemes. There is still a large variation in the speech of Japanese students according to their varying levels of proficiency. To overcome the limitations that come with a speaker independent model, we introduce a technique for speaker adaptation. Although the correct labels are necessary for adaptation, it is not easy to achieve them in the case of non-native speakers. Therefore we investigated the effectiveness of using baseform (pronunciation dictionary) labels by comparing them with hand-labels that count erroneous pronunciation manually.

With a CALL system, it is necessary to determine pronunciation errors from the view of native English speakers. However, the recognition results achieved using the above method may possibly be different from native speakers' perception. Thus, we verify the error candidates with segment-input pair-wise classifiers which are optimized to discriminate confusing phonemes by using a database comprised of native speakers' speech. Furthermore, we experimentally compared these methods and acoustic modeling of English by Japanese students.

2. OVERVIEW OF CALL SYSTEM

An overview of our CALL system is depicted in Figure 1. The system consists of three parts: (1) speaker adaptation, (2) phoneme alignment and error detection, and (3) generation of instructions.

Speaker adaptation is performed beforehand to cope with the acoustic variety seen in non-native speech and the differences that can be seen when compared with the speech of native speakers.

In pronunciation training, students select a sentence

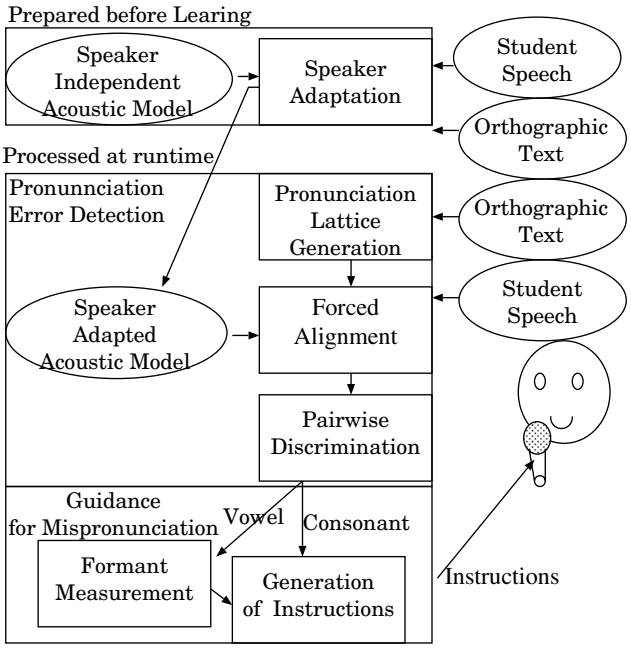

Fig. 1. System overview

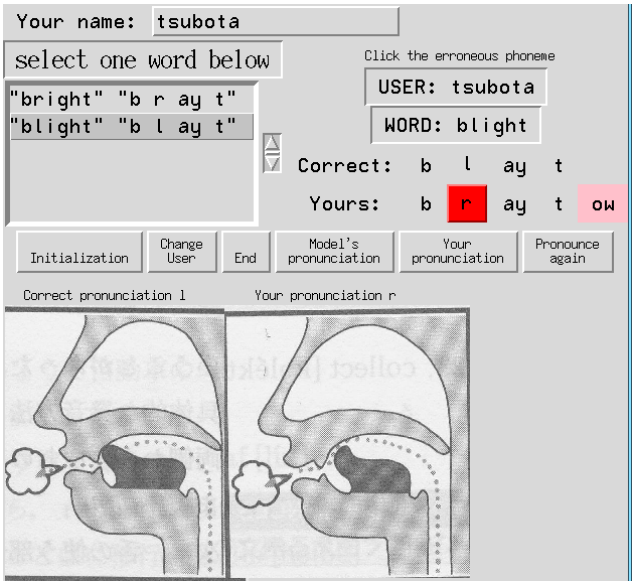

Fig. 2. Example of guidance

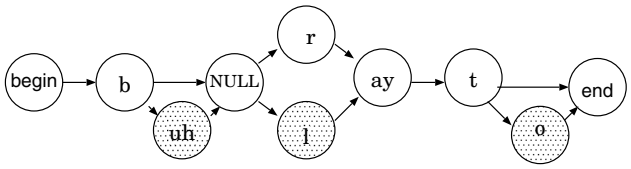

Fig. 3. Pronunciation network for the word "bright"

or phrase for practice. A pronunciation network is automatically generated to predict pronunciation variants by Japanese students. With this network and adapted acoustic model, the automatic recognition system effectively aligns the phoneme sequence and identifies erroneous phonemes. For more accurate error detection, we verify the erroneous phoneme segment with a pair-wise discriminant classifier which tests the erroneous phoneme against the correct phoneme using discriminative segment features.

For effective instruction of vowel pronunciation, we use an articulatory chart and formant frequencies which represent the articulatory elements[3]. As for consonants, phoneme recognition results are presented to the user. When the user selects a highlighted erroneous phoneme, the figure of articulation corresponding to the error is displayed. An example guidance is illustrated in Figure 2.

3. PRONUNCIATION ERROR PREDICTION

To predict pronunciation errors, we modeled error patterns of Japanese students according to the linguistic literature[4]. The model includes 79 kinds of error patterns. There are 37 patterns concerning vowel insertion, such as that pertaining to what vowels are inserted between a certain pair of consonants or after the final consonant of a word. In addition, 35 patterns for substitution errors were prepared. For deletion errors, there are 7 patterns, /w/, /y/, /hh/ deletion at word beginning and /r/ deletion in certain contexts. Examples of the patterns for insertion and substitution errors are shown in Table 1 and Table 2, respectively. Parentheses [] in Table 1 indicate the position where the vowel is inserted, and the phonemes after an arrow indicate what vowel is inserted.

For a given practice text (orthographic transcription), the model is used to automatically generate a network as shown in Figure 3 to cover possible error patterns. The prediction effectively guides the automatic speech recognition system to align phoneme sequences and identify erroneous phoneme segments.

4. VERIFICATION WITH PAIR-WISE CLASSIFIERS

To enhance error detection, we introduced pair-wise verifiers that are specifically designed to discriminate between a confusing pair of phonemes. Using segments from each pair of phonemes, discriminative features are derived by linear discriminant analysis (LDA) so that the classifiers are optimized to discriminate between the two classes[5]. When a pronunciation error is detected using HMM-based automatic speech recognition, three frames in the middle of the phoneme segment are extracted for verification of the erroneous phoneme against the correct one.

Table 1. Example of vowel insertion errors

initial clusters (CCV)	p [] (l|r) → uː|U	t [] r →oU|ɔː	b [] (l|r) → uː|U
final clusters (CCV)	p [] (t|θ|s) → uː|U	k [] (t|θ|s) →uː|U	b [] (d|z) → uː|U
final consonants	s [] → uː|U	d [] →oU|ɔː	k [] → uː|U

Table 2. Example of substitution errors

No counterpart in L2 syllable	tI/ʧI	tu/tsU	siː/ʃiː
No counterpart in L2 phoneme	l/r	b/v	s/θ
allophone	m/n/ŋ	ʤ/ʒ	
vowel substitution	ɔː/oU	iː/I	uː/U

Table 3. Performance of Japanese students in pair-wise classification for typically confusing phonemes (% correct)

phoneme pairs	l/r	b/v	s/θ	s/ʃ	f/h
classification rate	97.1	96.6	97.7	94.2	95.6
phoneme pairs	oU/ɔː	iː/I	uː/U	æ/ɑ	æ/ʌ
classification rate	89.4	91.1	89.1	95.3	92.2

In training with the use of pair-wise classifiers, we use a TIMIT database consisting of 6300 sentences spoken by 630 speakers uttering 10 sentences each. Performance of pair-wise classifiers is shown in Table 3. Here, accuracy is measured by cross-validation in the database. It has been confirmed that pair-wise classification generally achieves an accuracy level of over 90%, while the HMM-based phoneme recognition accuracy level is around 60%. Hence, we see an improvement in the reliability of error detection using pair-wise classification.

5. ACOUSTIC MODELING

Next, we describe acoustic modeling for automatic recognition. For evaluation of the proposed methods, we conducted phoneme recognition experiments with a corpus of English words spoken by Japanese students. The corpus[6] consists of 5950 speech samples. Seven Japanese speakers (2 male, 5 female) uttered 850 basic English words respectively. The database includes phonemic hand-labels, including erroneous phonemes, which are transcribed faithfully in order to meet the primary goal.

Speech data were sampled at 16kHz and 16 bit. Twelfth-order mel-frequency cepstral coefficients (MFCC) were computed every 10ms. Temporal difference of the coefficients (ΔMFCC) and power (ΔLogPow) were also incorporated.

5.1. Speaker Adaptation of Acoustic Model

Accurate segmentation and discrimination are not easy tasks since Japanese students' speech differs from that of native speakers. To compensate for acoustic variation, we introduce speaker adaptation using Maximum Likelihood Linear Regression (MLLR)[7]. There is a problem in applying supervised adaptation in a CALL system, in which students' pronunciation is not necessarily correct. Thus, we compared two transcription labels for adaptation: lexicon labels (baseform) and hand-labels that count erroneous pronunciation manually. Among the corpus of basic English words, 100 word samples were used for adaptation and other samples were used for evaluation. The baseline acoustic model was trained using the TIMIT database. We set up monophone HMMs for 41 English phonemes. Each HMM has three states and 16 mixture components per state. Phoneme recognition rates by adaptation are listed in Table 4. Adaptation even with the lexicon labels was shown to improve accuracy by about 5%, which is comparable to the result seen using hand-labels. Thus, we determined to use lexicon baseform for adaptation in the following experiments.

Table 4. Phoneme recognition rate by speaker adaptation

model	no adaptation	lexicon label	hand-label
native English	75.42%	80.55%	80.98%

5.2. Training with the use of Japanese Students' Speech

To improve the baseline model, we explored the use of data samples spoken by Japanese students. By using a model based on speech samples taken from Japanese students, we predicted better recognition of pronunciation errors.

We used a corpus of English compiled from Japanese students in a MEXT-funded project[1]. The corpus contains a total of 13129 sentences spoken by 178 Japanese speakers (85 male, 93 female). It does not include phonemic labels although there are a lot of pronunciation errors. Thus, we set up two kinds of phonemic labels for comparison: labels from lexicon baseform and automatic labeling using speech recognition. In automatic labeling of the data, we were also able to make use of speaker adaptation and pair-wise verification. Specifically, we applied four kinds of automatic labeling: (1) labeling with the native English model and error prediction; (2) labeling with the speaker adapted model from the native English model; (3) and (4)verified labeling with pair-wise classifiers on the first and second labels, respectively.

Table 5 lists phoneme recognition results obtained by comparing various acoustic models. In the phoneme recognition, we applied speaker adaptation and pair-wise verification. Thus we calculated four results for each model: baseline, adaptation, pair-wise verification and both adaptation and pair-wise verification. We confirmed the synergistic effect of adaptation and pair-wise verification in the recognition phase. These techniques significantly improve

[1]Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research on Priority Areas, No.12040106.

Table 5. Phoneme recognition rate with several types of acoustic modeling

model	label (adaptation, PW)	baseline	adaptation	PW	adaptation+PW
native English		75.42%	80.55%	84.28%	85.51%
Japanese students' English	baseform	77.95%	81.75%	85.28%	85.97%
	automatic labeling1 (x,x)	77.09%	81.48%	83.92%	85.28%
	automatic labeling2 (o,x)	77.95%	81.53%	84.18%	85.38%
	automatic labeling3 (x,o)	78.74%	82.37%	84.63%	85.80%
	automatic labeling4 (o,o)	77.95%	81.53%	84.18%	85.38%

Table 6. Phoneme recognition rate by adding Japanese phoneme models in parallel

model	label (adaptation, PW)	baseline	adaptation	PW	adaptation+PW
native English		78.94%	81.29%	85.22%	86.04%
Japanese students' English	baseform	78.70%	81.46%	85.12%	85.74%
	automatic labeling1 (x,x)	78.01%	81.54%	84.76%	85.76%
	automatic labeling2 (o,x)	78.24%	81.88%	84.79%	85.88%
	automatic labeling3 (x,o)	78.55%	82.16%	84.89%	86.21%
	automatic labeling4 (o,o)	78.24%	81.88%	84.79%	85.88%

PW: Pair-Wise verification

accuracy in all cases (compare left to right). However, they are not as effective in generating training labels (compare top to bottom).

As expected, the best acoustic model is the one trained using the Japanese students' speech database. On the baseline, this model yields a 3% improvement in accuracy over the native English model. The superiority decreased to 2% after speaker adaptation was applied. With pair-wise verification, the improvement is negligible. This result demonstrates that by using the adaptation and verification techniques, the native English model can compete with the Japanese students' model.

5.3. Combination of Japanese Phoneme Model

For better coverage of acoustic variety within one phoneme, we also incorporated fourteen entries from Japanese phoneme models, which are acoustically different but phonemically identical to (allophones of) English phonemes. They cover accented but acceptable pronunciation. For these entries, the Japanese phoneme HMM is used in parallel with the English phoneme HMM in recognition. Thus, the pronunciation dictionary has two entries for /b/, 'b_j' and 'b_e', where the suffixes 'e' and 'j' indicate the English and Japanese phoneme HMMs, respectively. Specifically, the following phonemes were given consideration: /b,d,f,g,h,k,m,n,p,s,t,w,y,z/.

The Japanese phoneme HMMs are trained using ASJ speech databases comprised of phonetically balanced sentences (ASJ-PB) and newspaper article text (ASJ-JNAS). We had access to approximately 20K sentences uttered by 132 speakers[8]. Specification of HMM was the same as for English models.

Table 6 lists phoneme recognition results. Compared with Table 5, the method improves accuracy by 1 to 3%. After speaker adaptation and pair-wise verification, improvement was maintained with some models. This result shows that Japanese phoneme models are effective in recognizing English spoken by Japanese students.

6. CONCLUSIONS

We have addressed methods for accurate recognition of English speech by Japanese students for use in a pronunciation learning system. To cope with acoustic variation and the differences between native and non-native speech, we introduced (1) training with the use of a database of English spoken by Japanese students, (2) bypass entries of Japanese phoneme models, and (3) speaker adaptation. In model training and speaker adaptation, accurate transcription was not available for non-native speakers. However, we were able to demonstrate that baseform label is sufficient. Together with the pronunciation error prediction and pair-wise verification methods proposed in this paper, we achieved a phoneme recognition accuracy level of 86%, which is 8 to 10% higher than the baseline result.

7. REFERENCES

[1] Chul Ho Jo et. al, Japanese pronunciation instruction system using speech recognition methods. In *IEICE Transactions*, vol. E83-D, 2000.

[2] Horacio Franco et.al, Automatic Pronunciation Scoring for Language Instruction. In *Proc. ICASSP*, vol. 2, pages 1471–1474, 1997.

[3] Y. Tsubota et.al, Computer-assisted english vowel learning system for japanese speakers using cross language formant structures. In *Proc. ICSLP*, vol. 3, 2000.

[4] Sutesaburo Kohmoto. *Applied English Phonology –TEACHING OF ENGLISH PRONUNCIATION TO THE NATIVE JAPANESE SPEAKER–*. Tanaka Press, 1965.

[5] Tatsuya Kawahara et. al, HMM based on Pair-Wise Bayes Classifiers. In *Proc. ICASSP*, vol. 1, pages 365–368, 1992.

[6] Tanaka Kazuyo et.al, Acoustic Models of Language-Independent Phonetic Code Systems for Speech Processing. In *Proc. of Spring meeting of the Acoustical Society of Japan*, 2001.

[7] C. J. Leggetter et.al, Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9(2):171–185, 1995.

[8] Tatsuya Kawahara et.al, Free Software Toolkit For Japanese Large Vocabulary Continuous Speech Recognition. In *Proc. ICSLP*, 2000.