CALL SYSTEM FOR JAPANESE STUDENTS OF ENGLISH USING PRONUNCIATION ERROR PREDICTION AND FORMANT STRUCTURE ESTIMATION

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

†Graduate School of Informatics, Kyoto University, Kyoto, Japan
‡Center for Information and Multimedia Studies / Graduate School of Informatics,
Kyoto University, Kyoto, Japan
† {tsubota,kawahara}@kuis.kyoto-u.ac.jp ‡ dantsuji@i.kyoto-u.ac.jp

ABSTRACT

We propose an effective pronunciation learning system using speech processing and recognition technologies. For automatic phonemic error detection, we perform pronunciation error prediction with orthographic text. To improve its reliability, we adopt speaker adaptation and segment input pair-wise classifier for effective feedback to generate personalized instruction. We also adopt formant frequency normalization technique (ML-VTLN and MMSE). The effect of adaptation and normalization is addressed in the paper.

1. INTRODUCTION

We present a CALL system for Japanese students learning English language. As English has a much larger phonemic inventory than Japanese language, Japanese students have to discriminate phonemes that are not used in Japanese, such as /l/ and /r/, /aa/ and /ae/. These confusions are often critical in mastering English for Japanese students. Therefore, automatic phonemic error detection and guidance are addressed in this paper.

Our primary goal of this research is to give effective feedback based on reliable phonemic error detection. For this purpose, a number of researchers have used speech recognition techniques for pronunciation learning systems[1][2]. Kawai et al proposed an approach that detects pronunciation errors with both L1 and L2 acoustic models and a pronunciation error lattice[3]. For pronunciation error detection, they use speaker independent model which limits the accuracy of segmentation and detection. To overcome this problem, we introduce speaker adaptation technique. Moreover, we address effective feedback instruction based on articulatory features, which is adopted to the learners.

2. SYSTEM OVERVIEW

Figure 1 shows the overview of this system. The system consists of three parts:(1) speaker adaptation (2) phonemic error detection (3) guidance generation.

In speaker adaptation, we use Maximum Likelihood Linear Regression (MLLR) with speech samples from students' speech to cope with speaker variability and the acoustic difference between native speakers' and students' speech[4].

Phonemic error detection is realized with pronunciation error prediction and pair-wise discrimination with segment input. For a given training text, the system automatically generates a pronunciation error network to cover possible error patterns. This network effectively guides the automatic speech recognition system to align phoneme sequences and identify erroneous phoneme segments. To enhance error detection, we use pair-wise classifiers which are optimized to discriminate the erroneous and correct phoneme segments.

For effective feedback for vowels, we use an articulatory chart and formant frequencies, which represent the articulatory elements. For consonants, we use the figures of articulation corresponding to students' errors.

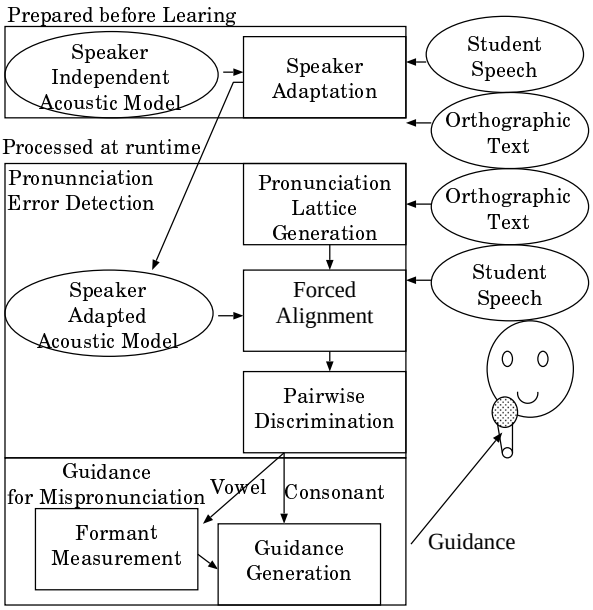

Figure 1: System overview

Table 1: Speaker adaptation with labels (recognition rate)

model	No adaptation	Lexicon label	hand-label
Native English model	72.91%	77.13%	77.27%

Table 2: Phoneme recognition rate(%) with several types of acoustic modeling

model	label (adaptation,PW[1])	baseline	adaptation	PW	adaptation+PW
Native English		72.91%	77.67%	77.13%	81.93%
Japanese students' English	baseform	75.41%	79.81%	80.52%	82.40%
	Automatic labeling1(x,x)	73.36%	77.20%	80.09%	81.92%
	Automatic labeling2(o,x)	74.60%	77.54%	80.60%	82.34%
	Automatic labeling3(x,o)	75.42%	79.04%	80.18%	82.25%
	Automatic labeling4(o,o)	74.60%	77.54%	80.60%	82.34%

3. PRONUNCIATION ERROR DETECTION

3.1. Method of Pronunciation Error Detection

Prediction Method

To predict pronunciation errors, we modeled error patterns of Japanese students according to linguistic literature[5].

For a given training text (orthographic transcription), the model is used to automatically generate a network as shown in Figure 2 to cover possible error patterns. The model includes the typical Japanese substitution errors listed in Table 3 and vowel insertion between consonants which affects the alignment of phonemes. The prediction effectively guides the automatic speech recognition system to align phoneme sequences and identify erroneous phoneme segments.

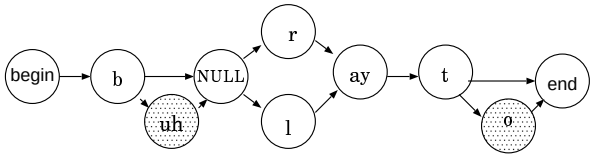

Figure 2: Example of pronunciation network

Table 3: Examples of substitution errors

No counterpart in L2 syllable	tɪ/ʧɪ	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/ɪ	uː/U

Speaker Adaptation

Accurate segmentation and discrimination is not an easy task since the students' speech is different from native speakers'. To compensate for acoustic variation, we introduce speaker adaptation using Maximum Likelihood Linear Regression (MLLR). There is a problem in applying supervised adaptation in CALL systems since the students' pronunciation is not necessarily correct. Thus, we compare two adaptation labels: lexicon labels (baseform) and handlabels that counts erroneous pronunciation. The speech corpus contains about 6000 words speech utterance by 7 Japanese students[6]. Among them, 100 word samples are used for adaptation and other samples for evaluation. The result is shown in Table 1. Adaptation even with the orthographic transcriptions improves recognition accuracy by about 10%.

Pair-Wise Discrimination

To enhance error detection, we also introduce pair-wise segment classifiers that are specifically designed to discriminate the confusing phonemes. When a pronunciation error is detected with automatic speech recognition, three frames in the middle of the phoneme segment are extracted and verified that the phoneme is actually pronunciation error.

We have confirmed with native speech data that pairwise classification achieves accuracy of more than 90% while the HMM-based phone recognition accuracy is around 60%. Thus, together with error prediction mechanism, the error detection with high accuracy is realized.

3.2. Comparison of Acoustic Models

For detection of pronunciation errors of students, we examine two kinds of acoustic models: native English acoustic model and students' English model.

Native English Model

We construct native English model from TIMIT database. The speech data were sampled at 16kHz and 16 bit. Twelfth-order mel-frequency cepstral coefficients(MFCC) are computed every 10 ms. Temporal difference of the coefficients (ΔMFCC) and power (ΔLogPow) are also incorporated. Cepstral mean normalization (CMN) is performed on every utterance.

Japanese Students' English Model

We use Japanese students' English corpus compiled by MEXT funded project[1]. But it does not have phonetic labels. We make two kinds of phonetic labels for them: phonetic labels from lexicon baseforms and labels from automatic labeling with native English model. We also compare four kinds of automatic labeling: No adaptation, adaptation, pair-wise discrimination.

[1]PW indicates pair-wise discrimination

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

Experimental Results

Table 2 shows the phonemic recognition results of comparison among various acoustic models. Four kinds of ecognition results for each model are calculated: baseline, adaptation, pair-wise discrimination and both adaptation and pair-wise discrimination. We confirmed the synergistic effect of adaptation and pair-wise discrimination. The best acoustic model is Japanese students' model with lexicon labels. Without adaptation and pair-wise discrimination, this model yields 3% better accuarcy than native English model. This superiority decreases to 2% after adaptation. After pair-wise discrimination, it almost disappear. This result indicates that with adaptation technique and pair-wise discrimination, native English model can achieve the same performance as Japanese students' model.

4. GUIDANCE GENERATION

4.1. Guidance for Vowel

Formant Structure Estimation

For effective feedback for vowels, it is necessary not only to point out incorrect articulatory categories, but also to precisely specify how to correct articulation. We use an articulatory chart and formant frequencies, which represent the articulatory elements. However, absolute values of the formant frequencies, which we call formant structure, are different for each person and must be normalized. In second language learning, unlike automatic speech recognition systems, students' pronunciation samples may not be reliable for estimating normalization parameters. So, we select correct vowel segments through the error detection method mentioned above.

Normalization of Formant Frequency

1. ML-VTLN

Maximum Likelihood Vocal Tract Length Normalization (ML-VTLN) is used to deal with speaker variability in speech recognition task. ML-VTLN performs spectral warping so that warped feature X gives maximum matching probability for given acoustic model A and transcription W.

$$\hat{\alpha} = \arg\max_{\alpha} P(X^{\alpha}|W, A) \tag{1}$$

ML-VTLN is efficiently integrated with filter bank Warping parameter α is chosen between 0.70 and 1.3 with 0.2 step. We use the α as normalization parameter for formant frequency.

2. MMSE

We also test Minimum Mean Square Error (MMSE) criterion. The parameter α is chosen to minimize the following equation, where M is mel-scale formant frequency of the speaker and $\overline{M_j}$ is the mean of the formant frequency among the native speakers. For each vowel, we compute the standard deviation of the formant frequency among the native speakers. Then, we minimize the mean of the standard deviations among vowels.

$$E = \frac{\sum_{i}^{\#vowels} \sqrt{\sum_{j}^{2} (M_{i,j} - \overline{M}_{i,j})^2}}{\#vowels} \tag{2}$$

Table 4: Evaluation of Normalization(Hz)

	i	I	e	æ	ʌ
before	91.3	77.6	75.5	85.9	82.5
ML-VTLN	104.4	79.4	74.6	84.1	75.0
MMSE	72.5	66.4	58.3	63.9	80.9
	a	ɔ	U	u	**total**
before	94.8	86.9	111.8	164.3	**96.7**
ML-VTLN	86.1	91.3	92.5	150.2	**93.1**
MMSE	73.4	82.2	85.3	160.9	**82.3**

Evaluation of Formant Frequency Normalization

To verify the normalization methods, we collect speech samples from native speakers of English. They utter words in context /b-V-t/[2]. Evaluation is done with equation (2). The results are shown in Figure 3 and 4 and Table 4. It is confirmed that standard deviation decreases by 14Hz. Using this normalization method, we make guidance as shown in Figure 5.

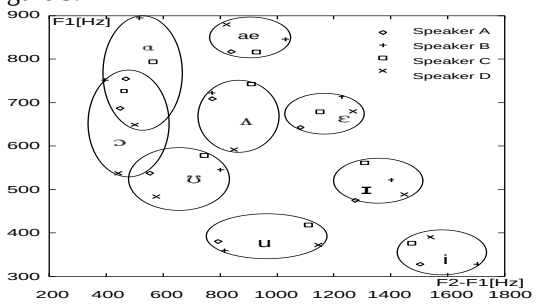

Figure 3: Formant distribution of vowel before normalization

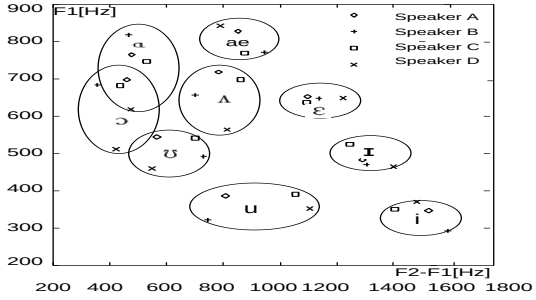

Figure 4: Formant distribution of vowel after MMSE normalization

[2]V means one of English vowels

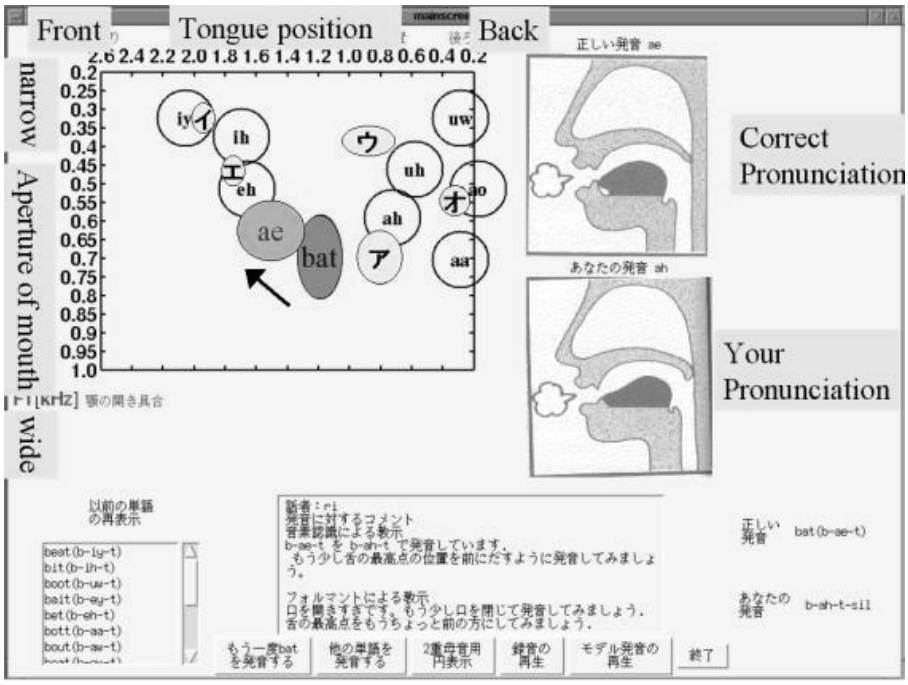

Figure 5: Example of Guidance for Vowels

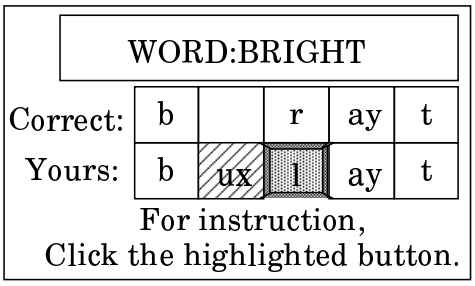

Figure 6: Example of Guidance for Consonants

4.2. Guidance for Consonant

For consonants, phonemic errors are indicated with a highlighted button as shown in Figure 6. Students get articulatory instruction for erroneous phonemes such as shown in the right part of vowel guidance by clicking the corresponding button.

5. CONCLUSION

In this work, we have examined the feasibility of CALL system for Japanese students of English using pronunciation error prediction and formant structure estimation. In the experiment, we have confirmed that these techniques effectively detect pronunciation errors.

References

[1] Chul Ho Jo, Tatsuya Kawahara, Shuji Doshita, and Masatake Dantsuji. Japanese pronunciation instruction system using speech recognition methods. In *IEICE Transactions*, Vol. E83-D, 2000.

[2] Horacio Franco, Leonardo Neumeyer, Yoon Kim, and Orith Ronen. Automatic Pronunciation Scoring for Language Instruction. In *Proc. ICASSP*, Vol. 2, pp. 1471–1474, 1997.

[3] Goh Kawai, Akira Ishida, and Keikichi Hirose. Detecting and correcting mispronunciation in non-native pronunciation learning using a speech recognizer incorporating bilingual phone models. In *Journal of the Acoustical Society of Japan*, Vol. 57, 2001.

[4] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, Vol. 9, No. 2, pp. 171–185, 1995.

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

[6] Tanaka Kazuyo, Kojima Hiroaki, Tomiyama Yoshihiro, and Dantsuji Masatake. Acoustic Models of Language-Independent Phonetic Code Systems for Speech Processing. In *Proc. of Spring meeting of the Acoustical Society of Japan*, 2001.

[7] Thomas Hain, Philip C Woodland, Thomas R Niesler, and Edward W.D Whittaker. The 1998 HTK System for Transcription of Conversational Telephone Speech. In *Proc. ICASSP*, Vol. 1, pp. 57–60, 1999.