COMPUTER-ASSISTED ENGLISH VOWEL LEARNING SYSTEM FOR JAPANESE SPEAKERS USING CROSS LANGUAGE FORMANT STRUCTURES

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

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

ABSTRACT

We present a novel Computer-Assisted Language Learning (CALL) system for Japanese students who learn English as a second language. We regard formant structure of Japanese vowels pronounced by Japanese learners of English as their own formant structure. This structure is transformed to learner's ideal English formant structure based on the relationship between English and Japanese articulation charts both of which are corresponded with formant structure. When the learners' English pronunciation is input, it is compared with the estimated ideal one, and articulatory instructions are given. We verified that the mapping and estimation of English vowel parameters are correct with bilingual speakers' speech and observed the learning effect with five students who tried the system.

1. INTRODUCTION

Our goal is to construct a system to instruct English vowel pronunciation for Japanese learners. The number of English vowels is twelve, more than twice as much as that of the five vowels in Japanese, and the phonetic values of two or more English vowels are recognized as same as a single Japanese vowel by the ordinary Japanese learners. To acquire skill in English pronunciation, students need to learn the phonological structure of English vowels. Japanese students hear English sounds based on their knowledge of the phonological system. Even if they feel their pronunciation is improper, they do not know how to correct it in proper way. Therefore, objective evaluation and an appropriate feedback mechanism is necessary.

To satisfy these demands, our system makes use of the findings of articulatory phonetics. Students' articulation is evaluated with acoustic features and the feedback of articulation will be given. We use formant frequency which are said to represent three articulatory elements. To realize this articulatory instruction, we need to normalize formant frequency for each student. The distribution of formant frequencies of vowels are relatively constant, but the absolute values, which we call formant structure, are different for each person [1]. In second language learning, students' pronunciation may not be correct and it is difficult to get reliable estimation of the formant structure for ideal pronunciation.

We propose a new method of normalization using a cross language formant structure. First, we normalize the formant structure for the Japanese vowels and estimate English vowels using the correspondence knowledge of articulation in both languages.

2. ERROR TENDENCY IN ENGLISH PRONUNCIATION BY JAPANESE SPEAKERS

The phonological structures of Japanese and English are totally different. While the Japanese language has an open syllable structure and mora timing rhythm, English has a closed syllable structure and stress timing rhythm. Learners must be aware of these differences while practicing pronunciation [2]. Otherwise their effort will bring out only a little improvement. In this section, the typical errors in English vowel pronunciation by Japanese speakers are presented [3].

2.1. Vowel Insertion after Consonants

Japanese language has an open syllable structure and almost all words end with vowels. Japanese learners of English tend to mispronounce for "beat"/biːto/ instead of /biːt/. Learners who make this insertion, seem not to pay attention to the different syllable structure of both languages. If the learners notice these insertions, they will correct the error. Our system checks for vowel insertions and alerts the learners of their presence.

2.2. Confusion between Monophthongs

Japanese learners hear English sounds based on their knowledge of phonological system. They hear the sound /ɑ/ as Japanese /a/, and pronounce /a/ as /ɑ/. In fact, the formant frequency space of Japanese /a/ overlaps those of English

Table 1: Minimal pairs

1	bott[bɑt]	bat[bæt]	but[bʌt]
2	beat[biːt]	bit[bɪt]	
3	boot[buːt]	book[bʊk]	
4	bit[bɪt]	bet[bet]	
5	bought[bɔːt]	bott[bɑt]	
6	bait[beɪt]	bet[bet]	
7	bought[bɔːt]	boat[boʊt]	

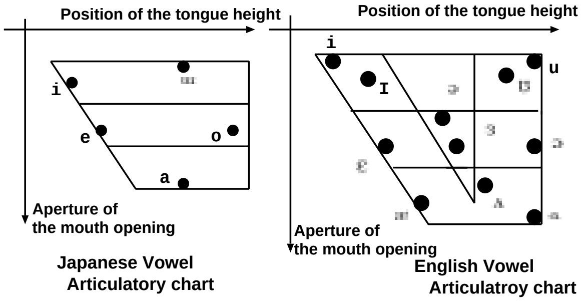

Figure 1: Vowel articulatory chart

vowels /ɑ/,/ʌ/ and /æ/. So, learners often have difficulty in pronouncing /ɑ/,/ʌ/ and /æ/ with distinction.

Japanese learners also focus on the familiar distinctive features. For example, /iː/ and /ɪ/ are different in quality indeed, but learners often think the difference is in length of the vowel. They often mistake /iː/ for /ɪ/. A substitution of this kind can cause a semantic error such as confusion of “live” and “leave”.

We design the system with consideration of these error tendencies and use minimal pairs, which differ in only vowel part, as the training material. Table 1 shows the confusing minimal pairs.

3. USE OF FORMANT STRUCTURE FOR INSTRUCTION

3.1. Vowel Articulation and Formant Frequency

From the articulatory phonetic point of view, vowels can be described with three articulatory elements: the aperture of the mouth opening, the position of the tongue and the amount of the lip rounding. Figure 1 shows both Japanese and English vowel articulatory charts. Formant frequencies are known to have relation with these three elements. Specifically, the first formant frequency is related with the aperture of mouth opening, the second formant frequency with the position of the tongue, and the higher formants with the amount of lip rounding. Our system generates instruction based on the formant measurement and this knowledge.

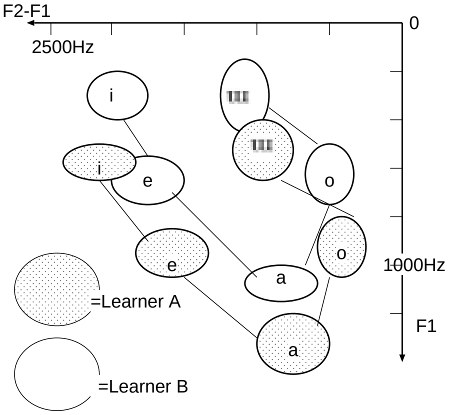

Figure 2: Formant chart

3.2. Adaptation to Learner

Each person has a different vowel formant frequency structure, so sometimes different speakers’ different vowels overlap as indicated in Figure 2. This causes a serious problem in generating appropriate instructions based on the measured frequencies. Some studies use the knowledge that the relative position is always kept among vowels for recognition of vowel sequence [4]. Others construct the vowel standard patterns beforehand and modify them in recognition [5]. However, we cannot introduce these techniques in our system, because learners’ pronunciation of foreign language is usually not stable and should not be regard as correct one.

We make use of the formant structure of Japanese, the learners’ mother tongue, and map it onto English formant structure. Specifically, we measure the formant frequencies of five Japanese vowels and get the maximum and minimum value of the first formant and the second formant. We regard the area as the learner’s unique formant area and map the English vowel structure on it.

3.3. Generating Articulatory Instruction

When ideal formant structure is estimated, it is compared to learners’ actual formant structure and instructions are generated. For example, if the actual first formant value is less than the ideal first formant value, the instruction will be “aperture of mouth opening should be wider.” Chart plotted with the first and second formant frequencies are used to show the instruction to learners, and the value of the third formant frequency is also displayed in the chart. Also, for easy understanding of instruction, explanation of the chart is given in text. Figure 3 shows an example of the system instruction.

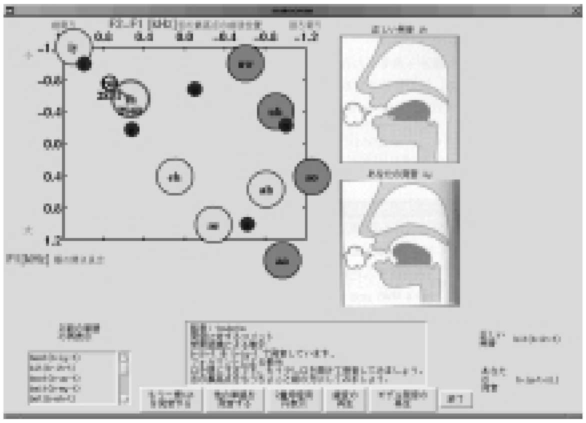

Figure 3: Example of instruction

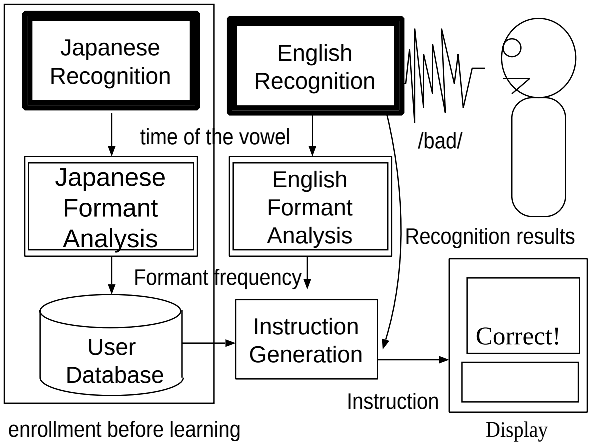

Figure 4: System configuration

4. SYSTEM IMPLEMENTATION

The system configuration is shown in Figure 4. For preliminary setting, five Japanese word utterances from each speaker are segmented with the Japanese phonetic recognizer and the formant of each vowel is analyzed and the ideal formant structure is calculated. The pronounced English word for practice is also segmented and formants are analyzed in the same way. Then, the articulatory instruction is generated by comparison between the actual value of the formant and the estimated ideal value. The instructions are given in three ways, in text, images and charts. These modules are explained in detail.

4.1. Configuration of Automatic Phonetic Recognizer

We use both a Japanese and an English speech recognizer. For Japanese recognition, we use the Japanese dictation toolkit in [6]. As samples for adaptation, five words “karu”, “kiru”, “kuru”, “keru”, “koru” are chosen, and the first vowel are segmented. For English recognition, we use the model trained with TIMIT database. As a sample for evaluation, we use “b-V-t-Iv” [1] pattern.

We perform phonetic recognition and segmentation by Viterbi algorithm. Specification of the acoustic model is summarized in Table 2.

Table 2: Acoustic and language models in both languages

	Japanese	English
training set	160 Male Read Speech	TIMIT DATABASE

feature parameter	MFCC(12)
	+ΔMFCC(12)
	+ΔPow(25)
		MFCC(12)
		+ΔMFCC(12)
		$+\Delta$Pow(25)

CMS	for each utterence	for each utterance
#phone	43	63
#HMM state	3	3
#Gausssian mixture	16	16
Pattern for sample words	$sil - kVrV - sil^2$	$sil - bVt(I_V) - sil$

4.2. Formant Analysis and Instruction Generation

Formant frequencies are estimated by peak-picking in smoothed power spectrum every five miliseconds. In order to remove the influence by the neighboring consonants and make reliable formant estimation, we cut out initial and final portions of vowel segments and compute the mean of formant frequency in the remaining central portions. The system compares the actual formant frequency and the estimated formant frequency on the first, the second and the third formant. Then the values are plotted in the chart and the instruction is made following the rules in Table 3. When vowel insertions are detected in the recognition result, an alert of their presence is given by text.

Table 3: Rules of instruction generation

	$Actual >$ $Estimated$	$Actual <$ $Estimated$
First formant	narrower	wider
Second formant	more frontal	more back
Third formant	more unrounded	more rounded

5. EXPERIMENTS

For the system evaluation, we performed two experiments: verification of formant structure estimation and confirmation of training effect.

[1] IV means Inserted Vowel, null or an English vowels.

[2] V means vowel set in the language. In this case, Japanese five vowels.

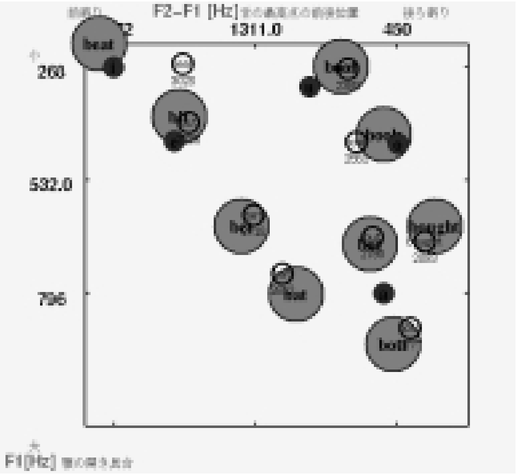

Figure 5: Estimated and actual formant structures

5.1. Verification of Formant Structure Estimation

To verify that ideal formant structure estimation is correct, we had ten bilingual speakers of Japanese and English uttered both Japanese and English words. From the Japanese words, their ideal English formant structure was estimated and compared to their actual English formant structure.

Four out of the six male speakers had good correspondence, but two speakers had a non-ignorable gap between the ideal and actual formant values. Three of the four female speakers also had a large gap, only one female speaker had a good match. Figure 5 shows a good example of the results.

The main cause of mis-matches seems to be measurement errors of formant frequency. Higher formant frequencies are sometimes mistaken as the lower frequency and the mean of formant frequency is jumped to a large value. We must use smoothing method for more robust measurement. We use vowel Japanese and English articulatory charts in [7]. As indicated in [7], there are also dialectual difference in Japanese /ɯ/. So charts do not always indicate the accurate learners' articulation. The value of formant frequency of a learner also changes in each utterance. For more robust estimation, we need to use more Japanese word samples.

5.2. Verification of Learning Effect

To measure the effect of training with this system, we conducted a listening test, in which learners listen to a number of words from the minimal pair and select the appropriate one. Five learners took this listening test before and after using the system for approximately one hour. Table 4 shows the results. Four of them improved their scores. Actually, three learners noted that they became more aware of the difference between various English vowels.

Table 4: English listening test results before and after using the system

Name	A	B	C	D	E
Before Learning	60	80	85	75	85
After Learning	85	95	90	75	90

6. CONCLUSION

We present a CALL system to generate effective instructions based on the formant structure. Although the proposed model is promising, the system needs further improvement in precision of measurement of formant frequencies. More robust formant structure estimation is now on going.

Acknowledgment: We are greatly indebted to Prof. Shuji Doshita of Ryukoku University for his enlightening guidance, Prof. Scott Johnson of Kansai University and to Masaaki Shimizu, Research Associate in the Center for Information and Multimedia Studies for their invaluable help. This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research on Priority Areas, No.12040106, 2000.

References

[1] Peter Ladefoged. *A Course in Phonetics*, chapter 8. Harcourt Brace College Publishers, 1993.

[2] Terence Odlin. *Language Transfer*, chapter 7. Cambridge Applied Linguistics, 1990.

[3] Peter Averu and Susan Ehrlich. *Teaching American English Pronunciation*, chapter 8. Oxford University Press, 1997.

[4] Rong Yu and Masayuki Kimura. Speaker-independent vowel recognition based on a mutual relational model of vowels (in japanese). *Trans. IEICE*, Vol. J69-D, No. 9, pp. 1320–1327, 1986.

[5] Masahide Sugiyama and Masaki Kohda. An unsupervised speaker adaptation techniques for vowel templates using speech recognition results (in japanese). *Trans. IEICE*, Vol. J69-D, No. 8, pp. 1197–1204, 1986.

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

[7] Yayoi Honma. *Acoustic Phonetics in English and Japanese (in Japanese)*, pp. 1–21. Ymaguchi shoten, 1985.