AUTOMATIC PRONUNCIATION ERROR DETECTION AND GUIDANCE FOR FOREIGN LANGUAGE LEARNING

Chul-Ho Jo *Tatsuya Kawahara* *Shuji Doshita* *Masatake Dantsuji*†

Department of Information Science,
Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

We propose an effective application of speech recognition to foreign language pronunciation learning. The objective of our system is to detect pronunciation errors and provide diagnostic feedback through speech processing and recognition methods. Automatic pronunciation error detection is used for two kinds of mispronunciation, that is mistake and linguistical inheritance. The correlation between automatic detection and human judgement shows its reliability. For the feedback guidance to an erroneous phone, we set up classifiers for the well-recognized articulatory features, the place of articulation and the manner of articulation, in order to identify the cause of incorrect articulation. It provides feedback guidance on how to correct mispronunciation.

1. INTRODUCTION

Recently, the effective application to aid in acquiring foreign language is getting possible thanks to the tremendous progress of speech processing technology and rapid improvement of computer hardware. We propose an autonomous pronunciation guidance system using the state-of-the-art speech recognition methods.

The primary goal of this research is to develop a foreign language pronunciation learning system to aid non-native speakers studying Japanese language as a second language. To be a meaningful CALL (Computer-aided Language Learning) for pronunciation, two functions, pronunciation error detection and feedback guidance, are indispensable.

In our system, two kinds of word set from non-native speakers are used to evaluate automatic pronunciation error detection : 1) pronunciation errors by mistake and 2) those by linguistical inheritance. The quality of error detection is confirmed by human judgement. In addition, what kinds of mispronunciation they would do in Japanese is also investigated.

†The author is with the Center for Information and Multimedia Studies, Kyoto University

Furthermore, we not only detect pronunciation errors but also offer feedback guidance to correct it. Until now there is no effective feedback method, especially in consonants. With the cooperation of the linguistic expert, we propose a novel feedback method based on the place and manner of articulation.

2. SYSTEM OVERVIEW

Figure 1 illustrates the basic design of speech processing performed in the system. The system consists of three kinds of processing modules : (1) pronunciation error detection, (2) feedback for vowels, (3) feedback for consonants.

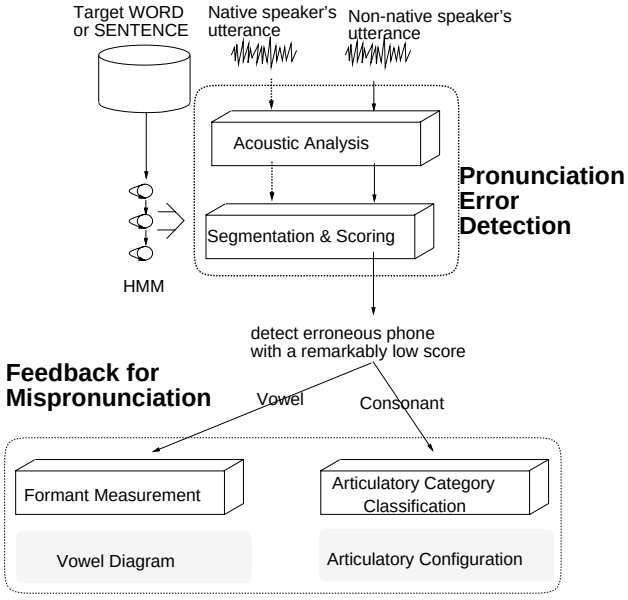

Figure 1: Block diagram of the automatic pronunciation error detection and guidance system

In automatic pronunciation error detection, we use an HMM log-likelihood probability normalized by its time duration. Viterbi segmentation is performed under a given transcription to produce accurate phonetic segmentation. However, this scoring method does not

always correlate well with the perceptual evaluation by human listeners due to the speaker variability. Thus, we focus on a great degradation in scores that suggests pronunciation errors.

As for the phone that is detected with a remarkably low score, we isolate the most stable frame of its corresponding speech using the HMM segmenter and trigger the feedback module. We use different feedback strategies for consonant and vowel, because their articulation is utterly different. For vowel feedback, we directly measure the formant frequency $F_1 - F_2$, and give feedback by plotting it on the articulatory vowel diagram mapped onto the open/close position on the jaw and front/back on the tongue[1]. For consonant feedback, we classify the sample with the articulatory categories on the place and manner of articulation to see whether his articulation is correct or not.

3. PRONUNCIATION ERROR DETECTION

Generally, pronunciation errors by non-native speakers result from two causes : 1) they make a mistake even though knowing how to pronounce the phone, and 2) they are not able to perceive how to do it since it does not exist in their own native language. We prepared two kinds of test set to evaluate our system : one set consisting of words that are not so hard to pronounce but difficult to memorize (M-set), and the other set consisting of eight types of pronunciation patterns that are difficult for them to pronounce (P-set). The speech samples were collected from totally eight male non-native speakers from different countries (see Table 1).

Table 1: Japanese non-native speakers

ID	SEX	AGE	BIRTH PLACE	RESIDENCE
A	M	24	Korea (Seoul)	1 yr.
B	M	28	China (Beijing)	1 yr.
C	M	26	Taiwan (I-Lan)	1 yr.
D	M	26	France (Toulouse)	3 yr.
E	M	28	Canada (Montreal)	2 yr.
F	M	19	Kazakstan (Almaty)	4 yr.
G	M	34	Indonesia (Bogor)	3 yr.
H	M	36	Kenya (Bungoua)	3 yr.

3.1. M-set (difficult to Memorize)

In Japanese, the numeral+counter combinations show a few different kinds of irregularities in their pronunciation. Typical examples are picked up in Table 2. The initial segment /h/ of K\ [*hoN*] or J, [*huN*] has become */b/* or */p/* according to the final segment of numerals: e.g., 1 K\ [*ippoN*], 2 K\ [*nihoN*], 3 K\ [*saNboN*], etc. In *ichi*+compounds, the mora[1] obstruent appears when the initial segment of counter is */h/* or */k/* : e.g., 1 K\ [*ippoN*], 1 2s [*ikkai*], 1 J, [*ippuN*], etc[2].

[1] Japanese speech rhythm, i.e., a unit of metric timing usually equal to, but sometimes smaller than, a syllable

Table 2: Numeral+counter combinations (M-set)

	K\ [*hoN*] (stick)	2s [*kai*] (times)	J, [*huN*] (minute)
1 [*ichi*]	1 K\ [*ippoN*]	1 2s [*ikkai*]	1 J, [*ippuN*]
2 [*ni*]	2 K\ [*nihoN*]	2 2s [*nikai*]	2 J, [*nihuN*]
3 [*saN*]	3 K\ [*saNboN*]	3 2s [*saNkai*]	3 J, [*saNpuN*]
4 [*yoN*]	4 K\ [*yoNhoN*]	4 2s [*yoNkai*]	4 J, [*yoNhuN*]
5 [*go*]	5 K\ [*gohoN*]	5 2s [*gokai*]	5 J, [*gohuN*]

These words are too difficult for non-native speakers to memorize. We had them read the M-set written only in Chinese characters, not in roman transcription, to see what kinds of mispronunciation they would make in the initial segment of the counter. To detect their mispronunciation, first we have examined by replaying their speech, and then compared the result with their automatically computed scores. Consequently, it is found out that the mispronunciation detection by a human judgement and automatic scoring correlates around a certain threshold score among the M-set. Three samples of */h/* in [*yoNhoN*], */k/* in [*saNkai*], */p/* [*saNpuN*] are detected as shown with a dotted line in Table 2. The use of an absolute threshold on phone scores was turned out to work effectively in the M-set.

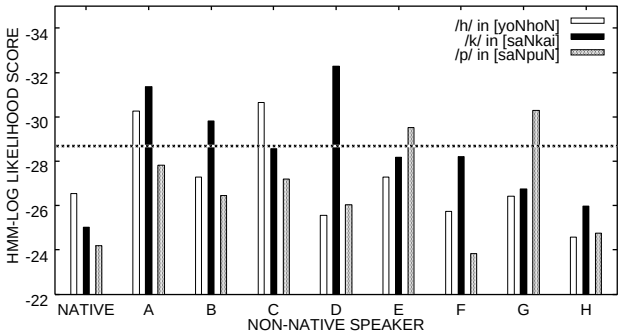

Figure 2: HMM-log likelihood score for M-set (Horizontal dotted line signifies an absolute threshold)

No mispronunciation was found in *ichi*+compounds, which is contrary to our expectations. It may be because most of non-native speakers memorize well this first-begining irregularity. As for */p/* in [*saNpuN*], non-native speaker **E** and **G** did */b/* as we expected that the mispronunciation [*saNboN*]→[*saNpoN*] or [*saNpuN*]→[*saNbuN*] would occur. **A**, **B** and **D** pronounced */g/* for */k/* in [*saNkai*]. It is assumed that the homonym, ;0 3, (third floor) that can be read as [*saNgai*], affected their mispronunciation. Finally, **A** and **C** unclearly pronounced */h/* in [*yoNhoN*]. A voiceless sound after */N/* is known to be hard to pronounce (see **TYPE 7**

in section 3.2).

3.2. P-set (hard to Pronounce)

The mispronunciation that is mentioned in previous section can be corrected once a non-native speaker realizes his mistake. However, the mispronunciation caused by disparity between his own native language and the target language cannot be easily corrected even if he realizes it. In this experiment, we aim to detect such kinds of mispronunciation. It is needed to investigate their speech statistically to detect such an inveterate mispronunciation. Eight kinds of pronunciation, which are known as the ones that non-native speakers have much difficulty in pronouncing, are prepared as below[2].

TYPE 1 */u/↔/o/* distinction
tsukue (desk) *mokuteki* (purpose)

TYPE 2 */sh/↔/s/* distinction
shiNyou (trust) *moushikomu* (propose)
shurui (kind) *shourai* (future)
gakusei (student) *yakusoku* (promise)

TYPE 3 */j/ ↔ /z/* distinction
jikoku (time) *ojigi* (bow) *jitsuryoku* (ability)
tazuneru (visit) *guuzeN* (accidently) *zeNbu* (all)
chikazuku (approach) *katazukeru* (dispose)

TYPE 4 Semi-vowel
yuubiN (mail) *yoyuu* (margin)

TYPE 5 Consonant after a mora obstruent
kekkyoku (after all) *gekkyuu* (salary)

TYPE 6 Voiced consonant */g/* and */gy/*
kagayaku (shine) *hogaraka* (cheerful)
gyougi (manners) *kougyou* (industry)

TYPE 7 Voiceless sound after */N/*
aNshiN (relief) *kiNchou* (tension) *shiNkou* (faith)

TYPE 8 Voiceless sound between vowels
ichigatsu (January) *shochi* (treatment)
shuukyou (religion)

First, we calculated the average score and standard deviation of speech samples from 20 male native speakers to set up the tolerance range for each phone. From the preliminary experiments, this tolerance range, a sort of relative threshold, produces better results for the P-set. The tolerance range of each phone is represented in Figure 3 with a thick line. The score far above it means an existence of inveterate mispronunciation on the corresponding phoneme. To the contrary,

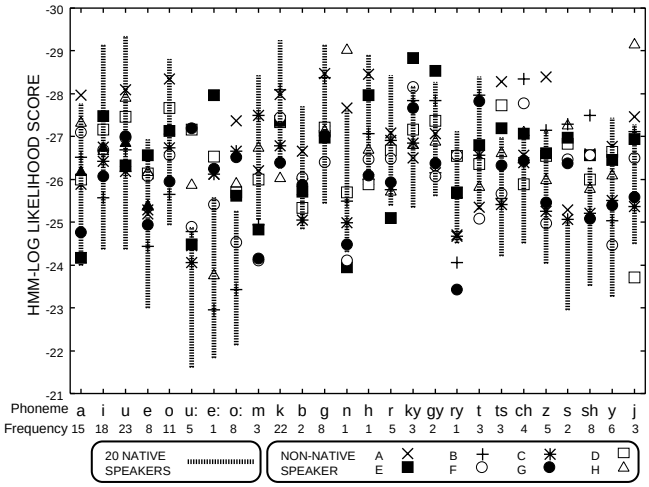

Figure 3: An average HMM score and a tolerance range for P-set

some non-native speakers show better scores than 20 native speakers in */m/*, */n/*, */r/*, */ry/*.

Based on the scoring result, we looked into their inveterate mispronunciation and found four types of mispronunciation. First, */s/* shows large difference between native speakers and non-native speakers, as it is generally known that non-native speakers tend to pronounce */s/* for */sh/* or */sh/* for */s/* (**TYPE 2**). Second, **H** makes very poor score for */j/*. By human judgement, it is confirmed that he has mispronounced */j/* in *ojigi* and *jitsuryoku*, respectively (**TYPE 3**). Third, they are not accustomed to pronounce voiceless sounds such as */ch/* and */ts/* after */N/* or between vowels (**TYPE 7** and **TYPE 8**). Through the experiment, an inveterate mispronunciation were successfully detected by our statistical automatic scoring, except a few samples that are considerably characterized by accent and rhythm.

Apart from eight types of pronunciation, it is noteworthy that long vowels such as */u:/*, */e:/*, and */o:/* show large degradation in their scores. We found a very interesting phenomenon : non-native speakers pronounce it as a diphthong, whereas many native speakers do it as a long vowel, e.g., [*shourai*] instead of [*sho:rai*] for the word *shourai*. In modern Japanese, not a few diphthong is pronounced as a long vowel regardless of the orthographic notation : e.g., [*ei*] → [*e:*], [*ou*] → [*o:*], etc[2].

4. FEEDBACK GUIDANCE FOR CONSONANT ARTICULATION

The pronunciation error detection provides them with the information what kinds of phones are mispronounced. But it does not give any hints or cues on

Table 3: Categorization of Japanese consonants by the place and manner of articulation

MANNER	PLACE CATEGORY	Bilabial VLAB	Bilabial ULAB	Dental VDEN	Dental UDEN	Post-alveolar POST	Palatal PALA	Velar VVEL	Velar UVEL	Glottal GLOT
Plosive	PLOS	/b/	/p/	/d/	/t/			/g/	/k/	
	PLOY	/by/	/py/	/dy/	/ty/			/gy/	/ky/	
Nasal	NASA	/m/		/n/						
	NASY	/my/		/ny/						
Fricative	FRIC		/f/	/z/	/s/					/h/
	FRIY					/j/ /sh/				/hy/
Affricate	AFFR				/ts/	/ch/				
Tap or Flap	TAPF			/r/						
	TAPY			/ry/						
Approximant	APPR	/w/					/y/			

how to correct it. In order to provide effective guidance, we perform the classification based on the articulatory features.

Table 3 lists the articulatory categories for consonants in Japanese. The horizontal axis represents manners of articulation, and the vertical axis places of articulation[3]. We use pair-wise classifiers that are specifically designed to extract optimal features for every pair of the categories, e.g., the classifier for fricative vs. plosive is dedicated to discriminate their differences. Stable discrimination is performed for the most of pairs[1]. The imperfection of consonant recognition is also avoided by performing our classification on the articulatory category rather than phoneme recognition itself.

By the HMM segmenter, the frame with a low score is segmented out and given to the pair-wise classifiers to look into its articulatory features from the aspect of both the place of articulation and the manner of articulation.

Figure 4 shows the classified results by the place of articulation and the manner of articulation for */sh/* of non-native speaker **B** that appears 8 times in the P-set. It is realized that he tends to pronounce *voiceless dental* not *post-alveolar* for */sh/* on the place of articulation (see **TYPE 2** in section 3.2), on the other hand appropriately *fricative* on the manner of articulation.

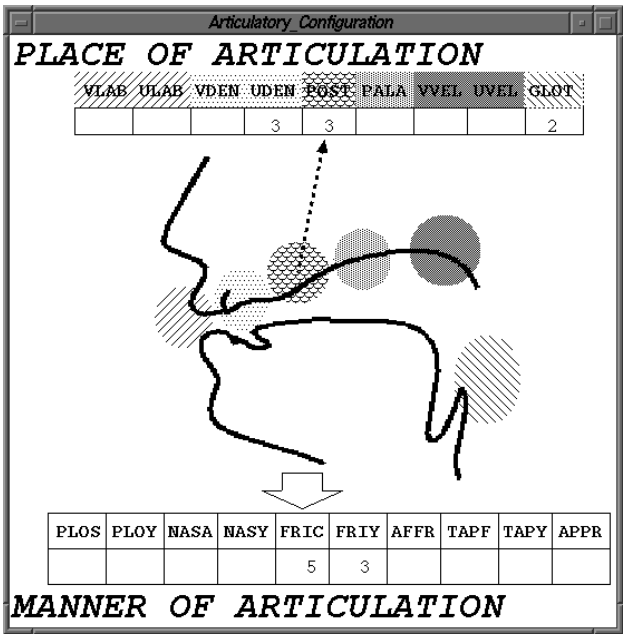

Figure 4: The consonant feedback using the articulatory configuration for the */sh/* sound of non-native speaker **B**

5. SUMMARY

We have demonstrated the feasibility of applying speech recognition techniques into foreign language pronunciation by detecting non-native speakers' pronunciation errors and giving appropriate feedback to them.

In the experiment, we successfully detected their pronunciation errors in the M-set, but in the P-set some disagreement between human judgement and automatic detection appeared due to the effect of prosodic features. Our scoring algorithm deals with the phonetic information. So the algorithm including prosodic features such as accent and rhythm needs to be explored.

The use of articulatory classification on the place and manner of articulation for consonants is advantageous because they are directly linked to the articulatory configuration.

References

[1] Chul-Ho Jo, T.Kawahara, S.Doshita, and M.Dantsuji. Japanese pronunciation training system with HMM segmentation and distinctive feature classification. In *ICSP*, volume 1, pages 341–346, 1997.

[2] Timothy J. Vance. *An Introduction to Japanese Phonology*. State University of New York Press, 1987.

[3] Masatake Dantsuji. Trends in articulatory phonetics. The Bulletin of the Phonetic Society of Japan, 1996.