A Preliminary study on ASR-based detection of Chinese mispronunciation by Japanese learners

DUAN Richeng [1], *ZHANG Jinsong* [12], *Cao Wen* [2], *Xie Yanlu* [1]

[1] School of Information Science, Beijing Language and Culture University, Beijing 100083, China
[2] Center for Studies of Chinese as a Second Language, Beijing Language and Culture University Beijing 100083, China

rcduan@hotmail.com, jinsong.zhang@blcu.edu.cn, tsao@blcu.edu.cn, xyl@blcu.edu.cn

Abstract

Detecting mispronunciations produced by non-native speakers and providing detailed instructive feedbacks are desired in computer assisted pronunciation training system (CAPT), as it is helpful to L2 learners to improve their pronunciation more effectively. In this paper, we present our preliminary study on detecting phonetic segmental mispronunciations on account of the erroneous articulation tendencies, including the place of articulation and the manner of articulation. Through modeling and detecting these error patterns, feedbacks based on articulation-placement and articulation-manner could be given. Moreover, Japanese learners of Chinese are focused on in this study. The experimental results show that the approach can detect the mostly representative pronunciation errors moderately well, achieving a false rejection rate of 8.0% and a false acceptance rate 32.6%. The diagnostic accuracy is 86.0%.

Index Terms: pronunciation error patterns, ASR-based mispronunciation detection, CAPT

10.21437/Interspeech.2014-360

1. Introduction

Computer assisted pronunciation training based on automatic speech recognition (ASR) has made considerable progress in the last few years [1-12]. In the beginning, since ASR-based confidence measures (CMs) have an advantage of easy implementation and can be applied to all L2 learners without restricting their L1 background, relevant studies have been extensively conducted [7-9]. These kinds of methods include different levels from speaker level to phone level, and have achieved a high consistency with human experts in speaker level [8]. They usually give a feedback of how confident the detector is when a given target speech is pronounced. According to the score, L2 learners can know their pronunciation proficiency. Besides the pronunciation scoring, if more detailed corrective feedbacks can be added, students will improve their pronunciation more effectively. In such a way, CAPT system can meet the pedagogical demands better.

Nowadays, more and more researchers are concentrating on how to give more detailed detection results and instructive feedbacks. Most of them try to define or select the erroneous types frequently produced by language learners. Some compare the actual phonemes that learners pronounce with orthographic transcription [13]. Others derive these erroneous types according to mispronunciation rules made by experts in second language learning [1]. A typical scenario is that the system prompts the learner to read pre-designed materials, and then pronunciation feedback is presented. Taking the word "red (/r e d/)" for example, a learner may mispronounce "red (/r e d/)" as "led (/l e d/)" and a feedback like: "You mispronounced phoneme /r/ as/ l/" will be given. In reality, as it is described in Yoon's study [14], there are many "distortion errors", i.e. the erroneous sound is always between two phoneme categories, rather than the absolute phoneme category substitution. In fact, it always deviates a little from the canonical sound /r/ and, consequently, we proposed to use a set of narrow-phonetic labels to annotate the erroneous tendencies of L2 learners in our previous work [15]. Having these annotations, a key point that whether the detecting task of these mispronunciations really works is put forward. Thus, based on the previous annotation work mentioned above, this paper aims to validate the feasibility of the methods we proposed by showing a detailed analysis of detection results, including 16 specific mispronunciations.

The paper is organized as follows: In the second section, how we define these mispronunciation patterns and our annotation work are briefly described. Section three presents our detection framework. Experimental results are reported and discussed in the fourth section. The last section is about the conclusions.

2. Mispronunciation pattern definition and annotation

Chinese L2 speech acquisition studies have been explored in recent years. There are some general and salient mispronunciation patterns when foreigner language speakers learn Chinese [16-18], such as an insufficient-aspiration mispronunciation manner or an inappropriate constriction. Most of them result from an inaccurate place of articulation or an erroneous manner of articulation. Hence, if we could define and detect these types of error patterns, it will enable the CAPT system to generate more detailed informative feedbacks, which are related to the speech organs. At the same time, according to these guidelines, students can target at their erroneous pronunciations.

2.1. Mispronunciation-pattern definition

We have designed diacritics for different kinds of general erroneous articulation tendencies: raising, lowering, advancing, backing, lengthening, shortening, centralizing, rounding, spreading, labio-dentalizing, laminalizing, devoicing, voicing, insertion, deletion, stopping, fricativizing, nasalizing, retroflexing and so on [15]. Several diacritics can also be combined to represent a complex sound variation. They are all based on articulation-placement and articulation-manner. Table 1 gives a small part of the mispronunciation-pattern definition in our convention.

2.2. Annotation

Directed by the convention in BLCU-CAPT-1 (such as conventions in Table 1) [15], multi-level phonetic transcriptions including words, syllables, Chinese traditional

Copyright © 2014 ISCA

"phonemes" of "Initials" and "Finals", lexical tones, and high-level prosody events, were annotated. A real annotation example is shown in Figure 1. There are 9-level labels ranging from phonemes to prosody, and this study only uses the phoneme tier, i.e. the third tier from the top.

Table 1. *Annotation Convention.*

(BLCU-CAPT-1-PART)

General Error Tendency	Diacritics	Specific Example	Notation
Advancing	+	e{+}n	The tongue position of phoneme e is a little advance
shortening	;	p{;}	The aspiration duration of phoneme p is shorter
Rounding	o	e{o}	Spread sound "e" has a rounding lip
Laminalizing	sh	sh{sh}	Balade-palatal phoneme sh is pronounced as Japanese lamina-alveolar

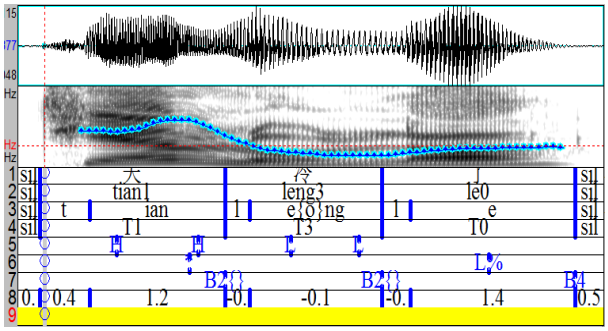

Figure 1: *A real annotation example.*

3. Extended pronunciation network based detection strategy

In order to detect mispronunciations, we use the extended pronunciation network in this study, which includes canonical pronunciation and every possible mispronunciation. In Section 3.1, a detailed description is given. The Flow chart of detection framework is shown in Section 3.2.

3.1. Construction of extended pronunciation network

Extended pronunciation network is a representation of pronunciation variants in the form of a network. When it comes to constructing extended pronunciation network, a dictionary with multiple pronunciations will be used. According to the possible mispronunciation in our annotation convention, we extend the pronunciation of every Chinese word (or Character). Table 2 gives the pronunciation representation of Chinese Character "看 (see)". Having this dictionary, when system gives the prompts, all possible pronunciations can be obtained by looking into the dictionary and the corresponding extended pronunciation network will be constructed automatically. One example of such network is shown in Figure 3.

Table 2. *Pronunciation Representation of "看(see)" in the dictionary.*

Words or Characters	Pronunciations(Pinyin)
看	k an
看	k{;} an
看	k an{-}
看	k{;} an{-}

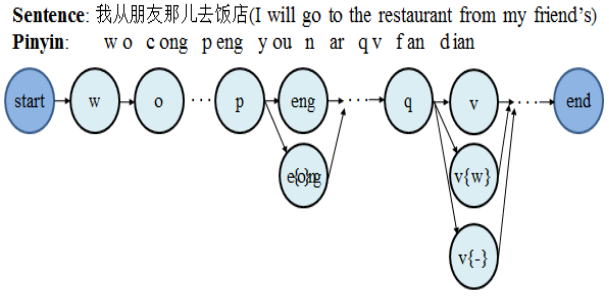

Figure 2: *Extended pronunciation network.*

3.2. Overview of detection framework

Figure 4 provides a flow chart of our detection framework: firstly, system prompts learners to speak a given utterance and records of their speech are recognized via the ASR-based detector. Then phone-level transcription will be output. At last, a list of feedbacks are given to learners.

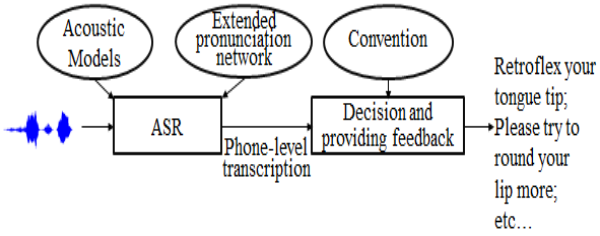

Figure 3: *Flow chart of our detection framework.*

4. Experiments

4.1. Experiment setup

We have collected a large scale of Chinese L2 speech database, which refers to as BLCU inter-Chinese speech corpus [15]. The Japanese part of our speech database has collected data of more than 100 speakers. The corpus used in this paper is the continuous speech of the Japanese part, including 7 female speakers. Each learner uttered a same set of 301 daily-used sentences. The recordings were also annotated by 6 graduate students who majored in phonetics and checked by the professor when they are inconsistent. All the works were done with the software "Praat 5.0.32" from University of Amsterdam. Table 3 gives some overall statistics of corpus.

The acoustic models are HMM tri-state models. 39-dimensional feature vectors, consisting of 12-dimensional MFCC, log-energy, and their first and second derivatives, are extracted from utterances using a 20 ms-length window shifted

every 10 ms. Minimum Phone Error (MPE) training criteria are adopted in this experiment. The entire experiment process is implemented by the HTK toolkit from Cambridge University.

Table 3. *A Japanese L2 inter-Chinese corpus.*

Text	301 utterance
Speaker	7 females
Number of utterance	1899
Number of phonemes	26431
Average length per utterance	14
Number of annotators	6
Number of annotations per utterance	2

4.2. Evaluation metric

There are totally 4 outcomes in the experiment: True Acceptance, True Rejection, False Acceptance, and False Rejection. Based on these outcomes, many different metrics can be used to evaluate the effectiveness of ASR-based CAPT. Here we use three kinds of metrics which are commonly used:

- False Rejection Rate (FRR): The percentage of correctly pronounced phones that are erroneously rejected as mispronounced.
- False Acceptance Rate (FAR): The percentage of mispronounced phones that are erroneously accepted as correct.
- Diagnostic Accuracy (DA): The percentage of detected phones that are correctly recognized, i.e. the detection result is consist with the human annotations.

4.3. Experimental results

As the 301 daily-used sentences we use are oral corpus, frequency of some kinds of specific erroneous articulation tendencies is fewer and their corresponding acoustic models are also unreliable. Hence, we rank the 65 kinds of specific pronunciation error patterns in descending order of occurrence frequency. The statistical result is shown in Figure 4 and top 16 are focused in this study. We split the whole corpus into five parts, and a K-fold Cross Validation (K=5) is adopted to validate our method and the overall detection performance is represented in figure 5. The average FRR and FAR are 8.0% and 32.6% respectively. Of the phones detected by the system, 86.0% of them are diagnosed correctly. As for the three kinds of metrics, while we aim to maximize the DA and minimize both error rates (FAR and FRR), there is an inherent trade-off between the two error rates. Considering the purpose of CAPT, it is essential to decrease the FRR since it is better not to discourage learners by rejecting their correct pronunciations due to the defection of CAPT's identifying functions. In this study, an 8.0% FRR and a 32.6% FAR sound good.

To analyze the results more specifically, Figure 6 gives the detection performance of those 16 specific mispronunciations. From the detecting results, we can see that among the 16 mispronunciations, one is not well-detected, i.e. the velar nasal /ing/, whose detection results (FRR=7%, FAR=88%) are rather worse than other phones. Perhaps it is because the phonetic values of the velar nasal /ing/ and its advancing error which sounds like the alveolar nasal /in/ are quite similar [19], or because the feature vectors we use cannot discriminate this kind of nasal sounds [20].

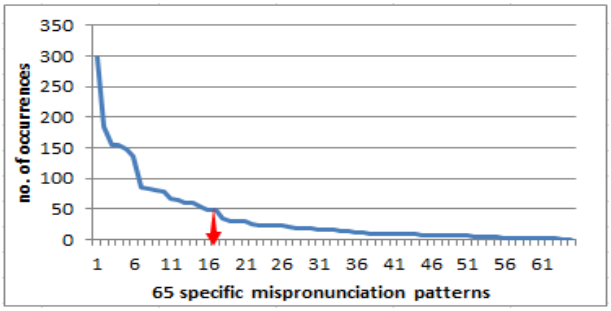

Figure 4: *Rank of mispronunciation patterns by occurrences.*

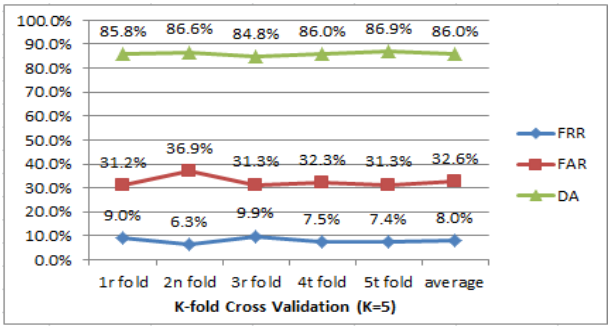

Figure 5: *Performance of our detection system.*

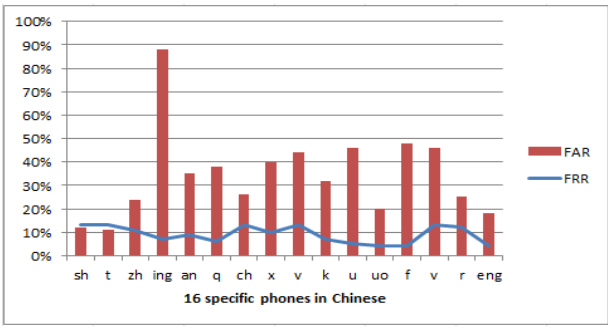

Figure 6: *Two kinds of error rates of 16 specific mispronunciations.*

On the other hand, among the other 15 specific mispronunciations, three broad heading error tendencies are dominant, i.e.

- Shortening: an insufficient aspiration or an inappropriate constriction.
- Laminalizing: balade-palatal phonemes are pronounced as Japanese lamina-alveolar.
- Lip rounding and spreading: sounds with spread lips have problems of rounded sound or sounds with the rounding lips have problems of spreading lips.

All erroneous tendencies above are typical and salient mispronunciation patterns when Japanese speakers learn Chinese [16-18]. For example, the plosives consonant /d, t/ are both present in Chinese and Japanese but the differences between them are not the same. In Chinese, they are both voiceless plosives and the difference is that phoneme /t/ is an aspirated plosive while /d/ is an unaspirated one. However, /d/ is a voiced plosive while /t/ is voiceless in Japanese. The Chinese distinctive feature, aspiration, is absent. As a result of

the language transfer, when Japanese learners pronounce the Chinese character "他" (the Pinyin is t a), the aspiration duration of segmental "t" is often shorter than Chinese speakers. Similarly, there are 5 vowels in Japanese phonetics, the diacritics of whom are partially the same with the Chinese ones, such as the phoneme "e" and phoneme "o". However, the shapes of lips are different when pronouncing those phones. As for the Japanese ones, the shapes are more natural as a whole. But in terms of the Chinese ones, they are characterized by spreading or rounding. So Japanese often neglect the shape of their mouth. Figure 7 shows the detection results of these three general error tendencies respectively.

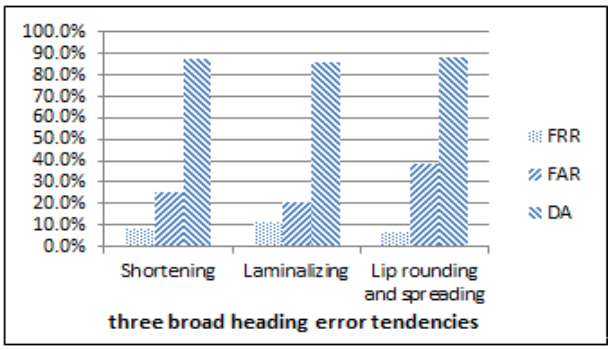

Figure 7: *Experimental results of three broad heading error tendencies.*

5. Conclusions

Aiming at realizing informative and instructive CAPT system, we propose a new way to select phonetic segmental error patterns on account of the erroneous articulation tendencies, including the place of articulation and the manner of articulation. The point lies in that we use a set of diacritic symbols to define and transcribe erroneous articulation tendencies and then model and detect these error patterns, then feedbacks based on articulation-placement and articulation-manner will be given to the learners. Totally, 16 specific phone-level mispronunciation tendencies have been detected moderately well, and results (FRR=8.0%, FAR=32.6%, DA=86.0%) are obtained. The performance convinces us the validity and feasibility of proposed methods. Moreover, we acknowledge that there is much room for improvement of the system's performance. We see several areas where further improvement could be gained: more distinctive phonetic features can be used to detect the specific erroneous tendency, such as VOT, formant and etc. Besides, utilizing posterior probability scores for decision and weighting different kinds of detection networks will also be helpful. In the near future, further efforts will be made to improve the system and more data will be used to develop CAPT system.

6. Acknowledgements

We would like to appreciate the hard annotation work by those student annotators from the center of studies of Chinese as a second language in Beijing Language and Culture University. Besides, this work is supported by National Nature Science Foundation of China (61175019), Beijing Higher Education Young Elite Teacher Project(YETP0879) and Youth Independent Research Program Projects of Beijing Language and Culture University (Special Funds of Basic Research Costs for the National University) (10JBT01).

7. References

[1] H. Meng, Y. Lo, L. Wang, W. Yiu, "Deriving salient learners' mispronunciations from cross-language phonological comparisons", Automatic Speech Recognition & Understanding, 2007.

[2] H. Meng, WK. Lo, AM. Harrison, P. Lee, KH. Wong "Development of Automatic Speech Recognition and Synthesis Technologies to Support Chinese Learners of English", APSIPA Annual Summit and Conference (ASC) 2010

[3] Y. Tsubota, T. Kawahara, M. Dantsuji, "Practical use of English pronunciation system for Japanese students in the CALL classroom", Proc. of ICSLP 2004

[4] A. Neri, C. Cucchiarini, H. Strik, L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training", Computer assisted language learning, 2002, 15:441:467.

[5] YB. Wang, LS. Lee, "Toward unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning", ICASSP 2013.

[6] A. Lee, YD.Zhang, J. Glass, "Mispronunciation Detection via Dynamic Time Warping on Deep Belief Network-Based Posteriorgrams", ICASSP 2013

[7] S.M. Witt and S.J. Young. "Phone-level pronunciation scoring and assessment for inter of Rules for interactive language learning", In Speech Communication, Vol.30, pp. 95-108 , 2000.

[8] J. Zheng, C. Huang, M. Chu, F.K. Soong, W. Ye, "Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation", Proc. ICASSP, Hawaii, USA, 2007: 201-204.

[9] F. Zhang, C. Huang, F.K. Soong, M. Chu, R.H. Wang. "Automatic mispronunciation detection for Mandarin", In Proceedings of the International Conference on Acoustic, Speech and Signal Processing, p.2077-2080,2008.

[10] A. Neri, C. Cucchiarini, H. Strik, L. Boves, "The pedagogy-tochnology interface in computer assisted pronunciation training", Computer assisted language learning, 2002, 15:441-467.

[11] AM. Harrison, WK. Lo, X. Qian, H. Meng "Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training", SLaTE 2009.

[12] T.M. Zhao, A. Hoshino, M. Suzuki, N. Minematsu, K. Hirose "Automatic Chinese pronunciation error detection using SVM trained with structural features", SLaTE 2012.

[13] D. Luo, X.Yang, and L. Wang, "Improvement of Segmental Mispronunciation Detection with Prior Knowledge Extracted from Large L2 Speech Corpus", Interspeech , 2011.

[14] S. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Landmark-based Automated Pronunciation Error Detection" Interspeech Proc., 2010.

[15] W. Cao, D. Wang, J. Zhang, Z. Xiong, "Developing A Chinese L2 Speech Database of Japanese Learners With Narrow-Phonetic Labels For Computer Assisted Pronunciation Training" Interspeech Proc., 2010.

[16] Xie, X. L., "A study on Japanese Learner's Acquisition Process of Mandarin Balade-Palatal Initials", Journal of Jilin Teachers Institute of Engineering and Technology, 2010.

[17] Li, F. Y., Cao, W., "Comparative study on the acoustic characteristic of phoneme /u/ in mandarin between Chinese native speakers and Japanese learners", Chinese Master's Thesis Full-text Database, No.S1,2011.

[18] Wang, Y. J., Shanggguan, X. N., "How Japanese learners of Chinese process the aspirated and unaspirated consonants in standard Chinese", Chinese Teaching in the World, 2004.

[19] Y.Wang, "An experimental study on the perception and production of nasal codas by Japanese Learners of Chinese Putonghua", Chinese Teaching in the World, 2002.

[20] H. Xu, Z. Meng, "Acoustic Parameters of Distinctive Features for Nasal Final in Mandarin" Dissertation, Communication University of China.,2009.