Developing A Chinese L2 Speech Database of Japanese Learners With Narrow-Phonetic Labels For Computer Assisted Pronunciation Training*

Wen Cao[1], *Dongning Wang*[1], *Jinsong Zhang*[1,2], *Ziyu Xiong*[3]

[1] Center of Studies of Chinese as a Second Language,
[2] College of Information Science,
Beijing Language and Culture University, P. R. China
[3] Institute of Linguitics, Chinese Academy of Social Sciences

tsao@blcu.edu.cn, wangdn03@gmail.com, jinsong.zhang@blcu.edu.cn, xiongzy@cass.org.cn

Abstract

For the purpose of developing Computer Assisted Pronunciation Training (CAPT) technology with more informative feedbacks, we propose to use a set of narrow-phonetic labels to annotate Chinese L2 speech database of Japanese learners. The labels include basic units of "Initials", "Finals" for Chinese phonemes and diacritics for erroneous articulation tendencies. Pilot investigations were made on the annotating consistency of two sets of phonetic transcriptions in 17 speakers' data. The results indicate the consistency is moderately good, suggesting that the annotating procedure be practical, and there is also a room for further improvement.

Index Terms: L2 speech database, narrow phonetic label, computer assisted pronunciation training

1. Introduction

Computer assisted pronunciation training (CAPT) approaches based on automatic speech recognition (ASR) technology have received considerable attentions in recent years [1-5]. They exploit ASR techniques to locate pronunciation errors and provide corrective feedback to language learners, for whom the one-on-one nature of man-machine interactive learning is known to be optimal. To develop such kinds of CAPT systems, L2 speech databases with pronunciations properly labeled are necessities [2,5]. They provide important information like regular mispronunciation statistics, serve as training data of acoustic models for error detection and testing data during performance evaluation. They are also important resources for phonetic-phonology and L2 acquisition studies [6].

The phonetic transcriptions of L2 speech databases are usually obtained by auditory analysis of non-native utterances into phonetic symbols, among them the most important ones are those for mispronunciations by L2 speakers. As it is a hard time-consuming and error-prone job to manually create phonetic transcriptions, researchers have been considering about how to efficiently label the mispronunciations for the purpose of CAPT [2,4,5]. Communicativeness was usually regarded as the most important requirement [2,5]. Efficiency and effectiveness of CAPT systems were also emphasized [5]. Robustness of error detection was also suggested to be a valuable requirement [5]. Directed by these guidelines, experienced experts usually give subjective grades to L2 speech at utterance levels, and label the most frequent pronunciation errors [2,5]. These two kinds of labels are used to develop CAPT systems which can detect categorical pronunciation errors including phoneme substitutions, deletions and insertions, and report the problems to learners by means of scores or scales [2,5].

Despite the progress of CAPT technologies, their applications to second language education are not successful [4]. The CAPT systems usually received criticisms from their users, especially language teachers, that they are lack of pedagogical uses. One major complain is about the feedback of scores, which are said to be not instructive enough to guide the learners to correct their pronunciations. For example, a feedback of "0.5" score for a sound [p^h] can only notify the learner that he has made an inaccurate pronunciation, but cannot help much him to correct his erroneous pronunciation manner.

To remedy such pedagogical defects, we regard that the feedback module of a CAPT system should function more informatively. It should provide not only a low score about an erroneous pronunciation but also a pertinent way for the learner to correct his error. Taking the previous example of the erroneous [p^h] sound, an ideal feedback should indicate that if the error is due to an insufficient aspiration or an inappropriate constriction place. For the purpose, appropriate error diagnoses are necessary in the system to analyze errors and their main characteristics. Therefore, it is necessary to build an L2 speech database with errors properly labeled.

As a guideline of phonetic annotation, we propose that more erroneous information than main phonetic category errors should be annotated in L2 speech. One way we suggested to do this is to use a set of narrow-phonetic labels to represent category errors and some erroneous tendencies [6]. It can even transcribe those gradual variations between two nominal categories, which are usually ignored in most conventional approaches. The demerits of this ambitious approach include hard labeling processes, and questionable reliabilities of phonetic labels.

Pilot annotations on phonetic segments have been carried out by a group of 6 annotators on a part of Chinese L2 Speech Database of Japanese Learners we have collected, which consists of 4,631 continuous utterances by 17 speakers. This paper presents our investigations on the label qualities, in

* Supported by the China MOE Project 07JJD740060 of Key Research Institute of Humanities and Social Sciences at Universities.

order to make clear the feasibilities of annotation conventions and existing problems for further improvements.

In the following sections, after a brief description of our database and segmental annotation conventions, investigations will be made on inter-annotator consistency in terms of symbol percentage agreement, correlation coefficients of error numbers, and phonetic segmentation deviations. Finally, some discussions will be made.

2. Database and annotation

2.1. Database

The Japanese part of our large scale Chinese L2 speech database (referred to as BLCU inter-Chinese speech corpus) has collected data of more than 100 speakers [6]. Among them, continuous speech of 17 Japanese speakers' speech (8 males and 9 females) has been phonetically annotated at segment level. Each speaker uttered a same sentence set of 301 daily used sentences. The annotators are 6 post-graduate students majoring in phonetics, divided into two groups. The speech data was annotated twice independently by the two groups, with each annotator labeling a continuous 200 utterances on a rotating basis. Table 1 gives some statistics of the database annotated.

Table 1. *Japanese L2 inter-Chinese database studied.*

Text	301 utterances
Speaker	8 males 9 females
Number of utterances	4,631
Number of phonemes	64,190
Average length per utterance	13.9 phonemes
Number of annotators	6
Number of annotations per utterance	2

2.2. Annotation convention

The usual way to annotate speech is to transcribe sounds faithfully in symbols as IPA annotation does. The category to which a sound is assigned can be regarded as an absolute measurement of it based on auditory judgments. Such a procedure demands highly a robust and accurate capability of an annotator to categorizing acoustic-phonetic events. On the other hand, accurate categorization of non-native speech is never demanded in L2 teaching. Teachers usually respond by pointing out articulation problems instead of an accurate description of student's speech. In view of these facts, we suggested that we could annotate erroneous articulation tendencies instead of transcribing error sounds faithfully. The idea can be illustrated by Figure 1: "e" sound has spread lips and "o" sound has rounding lips in Chinese. If a student has problems of rounded "e", a diacritic {o} can be used to indicate the erroneous tendencies of lip rounding. If the problem is with spreading "o" sounds, a diacritic {w} can stand for the spreading tendencies.

Merits of such an annotation approach are assumed as:

- As articulation tendencies are transcribed instead of absolute phonetic categories, workloads of acoustic-phonetic categorization reduced much for the annotators.
- Gradual variations between two nominal categories can be easily transcribed in the same way as categorical mispronunciations. Whereas such transcriptions were difficult in conventional approach.
- Erroneous articulation tendencies are very informative for creating instructive feedbacks to the students. For example, a detection of "spreading" errors will let the system to tell the student to try "lip rounding".

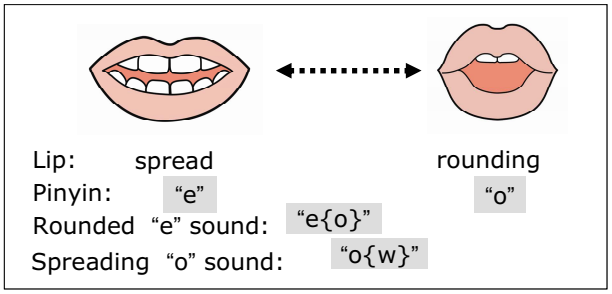

Figure 1. *An illustration of transcribing articulation tendencies. Pinyin is Chinese pronunciation letter.*

Diacritics were designed for the following erroneous articulation tendencies: raising, lowering, advancing, backing, lengthening, shortening, centralizing, rounding, spreading, labio-dentalizing, laminalizing, devoicing, voicing, insertion, deletion, stopping, fricativizing, nasalizing, retroflexing and etc [6]. Several diacritics can be combined to represent a complex sound variation.

2.3. Annotation procedure

Multi-level phonetic transcriptions including words, syllables, Chinese traditional "phonemes" of "Initials" and "Finals", lexical tones, and high-level prosody events, were first created via automatic methods. The key step to do this was to use an automatic speech recognizer to force-align the speech data into phonetic segments of "Initials" and "Finals". Afterwards, phonetic boundaries were assigned properly to other levels.

In the stage of manual annotation, annotators were asked to first check and adjust the phonetic segmentation boundaries, then annotate any mispronunciations according to the annotation conventions. An inventory of most frequent mispronunciations of Japanese speakers was also delivered to each annotator for reference beforehand. Still, the annotators were free to create new combination labels when they regarded as necessary. There are no time limits for the annotators to work on each utterance. After several orientations led by the first author of this paper, the six post-graduate student annotators worked independently in two groups, on a rotating basis with an arrangement avoiding pairing effects, eg, a pair of annotators always work on the same data. As a result, each utterance was annotated twice. All the work has been done with the software "Praat 5.0.32"[7].

Figure 2 shows a real annotation example, in which there are multi-level labels ranging from phonemes to prosody events. This study only used the phoneme tier, ie. The 3rd tier from the top in the figure.

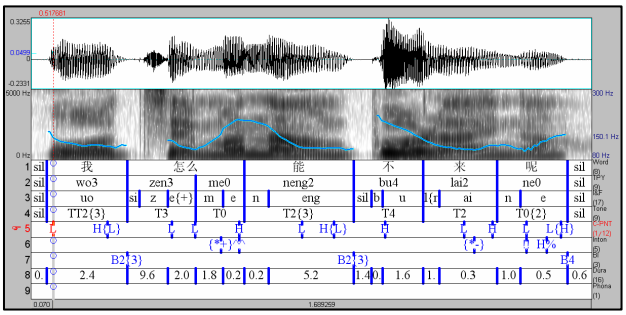

Figure 2. *A real annotation example.*

3. Evaluation Experiments

The annotation quality was evaluated from two aspects: consistency of phoneme labels and phonetic segmentations. As there was a possibility of symbol mismatch, dynamic warping was used to align the two sets of annotations beforehand.

3.1. Consistency of phoneme labels

Consistency of two sets of phoneme annotations was evaluated in percentage agreements with respect to each pair of annotators. The ratios range from 77.0% to 84.3%, and average as 80.7%, as shown in Figure 3. According to literature [8,9], such results can be regarded as moderately good for the nature of narrow phonetic labels.

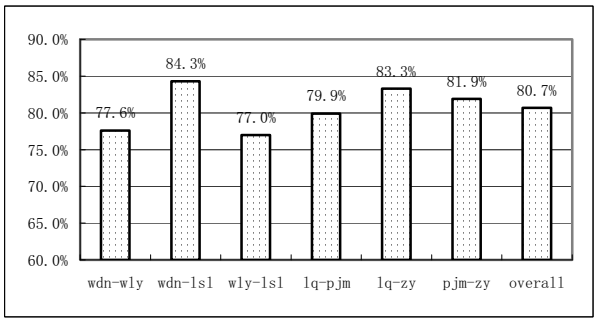

Figure 3. *Percentage agreements of phoneme labels.*

3.2. Analysis of inconsistency

We made an investigation into the annotator pair-wise results of consistency analyses. The annotation labels can be classified into four groups in view of consistency, i.e.

- Consistent correct (CC) phonemes: those regarded as correct by both annotators.
- Consistent mispronunciations (CM): those were annotated as same mispronunciation labels.
- Inconsistent mispronunciations (IM): regarded as mispronunciations by both annotators but annotated in different labels.
- Warning mispronunciations (WM): regarded as mispronunciations by only either one of the two annotators.

Figure 4 shows the distributional percentages of the four groups of phoneme labels with respect to pair-wise annotators. We analyze the results as:

- Among the overall 80.7% (CC+CM) consistent labels, about 74% accounts for correct phonemes, and left 6.6% for mispronunciations.
- The total inconsistent labels account for 19.3% (WM+IM), in which about 2.8% IM phonemes are treated as mispronunciations with consensus but labeled differently, the other 16.5% was regarded as correct by either of one annotator.
- Inconsistency due to a mistaking use of phoneme labels should be less than 2.8% (IM). In other words, differences of personal judgments account to a maximum of 19.3%.

If we relax our thresholds discriminating what is a mispronunciation, those 16.5% (WM) with one correct judgment can be regarded as correct. Then among a total of 9.4% (CM+IM) labeled mispronunciations, 6.6% CM labels account for a 70.2% relative consistency rate.

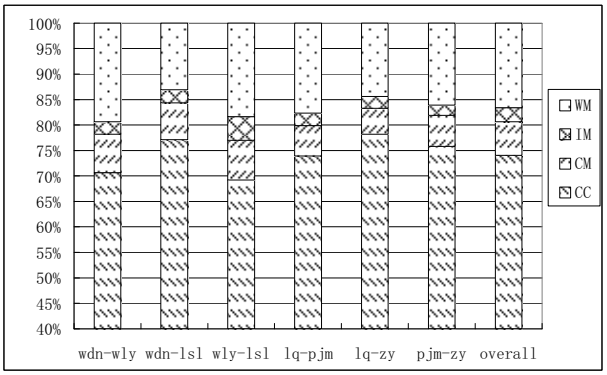

Figure 4. *Percentage distributions of phoneme labels.*

3.3. Correlation of mispronunciation phonemes

In order to see if there exist any subjective bias effects on judging which phoneme as errors, correlation coefficients were computed for the phoneme based mispronunciation rates for the two groups, shown in Figure 5.

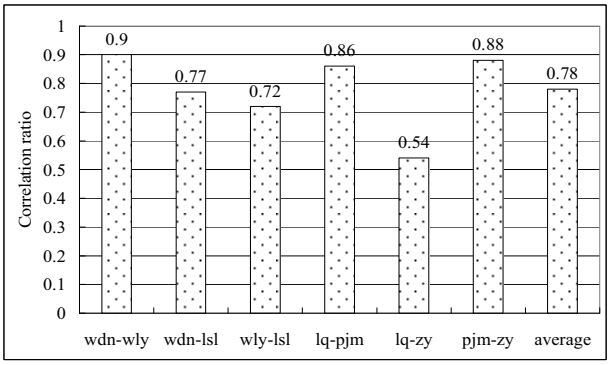

Figure 5. *Correlation ratios of phoneme based mispronunciation rates.*

A high correlation ratio suggests that the pair of annotators did reliable judgments of mispronunciations. When a phoneme has more problems, both the two annotators made more appropriate labels, despite of the fact that they might have different judge threshold levels and the absolute amounts of labels might differ much. The results show that the average ratio approaches 0.78, with most pairs higher than 0.7 except

"lq-zy" pair with 0.54, which suggests a subjective difference existing between that pair of annotators.

3.4. Phonetic segmentation

The quality of phonetic segmentations was evaluated using the standard deviation measurement of absolute timing gaps of category boundaries, as done in [9]. Definitions of phoneme categories are given in Table 2, and results are in Table 3 together with those of [9] for a comparison. The tendencies are: boundaries before null-Initial syllables as well as laterals and nasals have larger deviations, while others have smaller deviations. When our results are compared to those in [9], most of them are in the similar patterns and ranges (except boundaries before null-Initial syllables), despite of the different nature of non-native and native speech. This suggests that human segmentation of inter-language data can achieve similar results with native language data.

Table 2. *Phoneme categories for segmentation check*.

Category	Symbol	Example
Vowel	V	a
Nasal ending Finals	VN	an
Aspirated stops	AP	p
Non-aspirated stops	UAP	b
Aspirated affricatives	AA	q
Non-aspirated affricatives	UAA	j
Fricatives	F	s
Nasals	N	n
Laterals	L	r,l

Table 3. *Standard deviations (SD) of timing gaps for different category boundaries*

segment boundary	SD in ms		segment boundary	SD in ms	
	ours	[9]		ours	[9]
VN+V	23	8	V+F	8	7
VN+VN	21	8	UAA+V	8	
V+V	20	15	VN+F	8	6
V+VN	19	15	UAA+VN	8	
VN+L	16		V+N	8	9
VN+N	15	16	AA+VN	7	
VN+AP	13	10	N+V	7	
V+UAP	12	12	VN+UAA	7	
L+VN	11		AP+V	7	
L+V	11	8	AP+VN	7	
V+UAA	10		UAP+VN	7	
VN+UAP	10	10	F+V	7	7
V+L	9	14	F+VN	7	
V+AA	9		N+VN	6	
VN+AA	8		AA+V	6	
UAP+V	8				

3.5. Discussion

As a summary, measurements through symbol percentage agreements and phonetic boundary deviations show that: the two sets of manual annotations have an overall 80.7% consistency rate for the data of 17 Japanese inter-Chinese utterances, and the phonetic segmentation boundaries are annotated in a similar accurate level of those done to native speech. The results of these two measurements ensure the validity and feasibility of the annotation convention and procedure we developed to build L2 Chinese speech corpus.

Distributional analyses showed that less than 2.8% inconsistency might be due to incorrect use of symbols, and a maximum of 19.3% inconsistency might result from different judgments of a pair of annotators. When we relax the criterion to judge mispronunciations, the 16.5% phonemes with one correct judge can be taken as correct sounds. Then among the labeled mispronunciations, 70.2% ones are consistently labeled by the two groups of annotators. Phoneme based correlation check showed that most annotations were reliable except some colored with subjective differences.

4. Conclusion

Aiming at realizing informative and instructive CAPT technology, we proposed a new way to annotate mispronunciations of L2 Chinese speech databases. The point lies in that we use a set of diacritic symbols to transcribe erroneous articulation tendencies. Continuous speech of 17 Japanese speakers has been annotated twice by two groups of annotators, and quality checks showed that annotation consistency are moderately good. The results convince us the validity and feasibility of proposed annotation methods. Further efforts will be made to improve the annotation efficiency and use the data to develop CAPT systems.

5. Acknowledgements

We would like to appreciate much the hard annotation work by those student annotators from the center of studies of Chinese as a second language in BLCU.

6. References

[1] C.Cucchiarini et al, "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms", Speech Communication, 2000, 30: 109-119

[2] Y. Tsubota, T. Kawahara, M. Dantsuji, "Practical use of English pronunciation system for Japanese students in the CALL classroom", Proc. Of ICSLP2004.

[3] R. Hincks, "Speech recognition for pronunciation feedback and evaluation", ReCALL 2003, 15(1): 3-20

[4] 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.

[5] A. Neri, C. Cucchiarini, H. Strik, "Segmental errors in Dutch as a second language: how to establish priorities for CAPT", InSTIL/ICALL Symposium, 2004

[6] W. Cao, J. Zhang, "The Establishment of a CAPL Inter-Chinese Corpus and Its Labeling". Proc. Of NCMMSC, 2009.

[7] Paul Boersma, David Weenink: http://www.fon.hum.uva.nl/praat/

[8] B. Eisen, "Reliability of Speech Segmentation and Labelling at Different Levels of Transcription", Proceedings of Eurospeech 1993, Berlin, Germany. 673-676.

[9] Wesenick, M.-B., A. Kipp "Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals", Proceedings of ICSLP 1996, Philadelphia/USA.