A Semi-manual Annotation Approach for Large CAPT Speech Corpus

Yanlu Xie, Xin Wei, Wei Wang, Jinsong Zhang
Beijing Advanced Innovation Center for Language Resources
Beijing Language and Culture University, Beijing 100083, China

xieyanlu@blcu.edu.cn blcuweixing@163.com vickyyzq@126.com jinsong.zhang@blcu.edu.cn

Abstract

Annotation plays an important roles in speech database. However annotation is time and annotators consuming. This paper proposes to provide phoneme-level labeling candidates with the state-of-the-art ASR models. The annotators could manually choose the appropriate labels and make final decision. Also a posterior probability evaluation method is applied to measure the annotation results. BLCU-SAIT speech corpus, a corpus aimed at computer aided pronunciation training (CAPT) is labeled with the annotation approach. Experimental results show that the mean consistency rate of manual labels is 87.2%. The posterior F1 score is 0.857. The annotation problems are converted from the open-ended questions to multiple-choice questions with the method. And the annotation results meet the requirements of CAPT systems.

Keywords: annotation, manual label, F1 score

1. Introduction

Annotation plays an paramount roles in speech database (Bird, S. 2001), especially in the database for language learning. For instance, the annotation is rewarding in studying the language phenomenon and in developing technology in assisting language learning. More and more interlanguage speech database is developed for the second language learning task recently. The scale of the database becomes larger. For example, the iCALL consists of 90,841 utterances from 305 speakers (Chen, N. F. 2016), the ERJ consists of 68,000 utterances from 200 speakers(Minematsu, N. 2002). It is a difficult task to annotate so much speech data manually. And the accuracy of the annotation results is questionable.

Some researchers have proposed Computer-Aided Annotation methods. CHAT (Codes for the Human Analysis of Transcripts) provided the instrument for producing and analyzing data (MacWhinney, 2000). DARCLE Annotation Scheme (DAS) proposed a workflow for annotating long natural language recordings. The transcripts and speech boundaries could be labeled automatically by some tools(Marisa Casillas, 2017). SLAM and Speech Analyzer POSCAT even could automatic give phone level labels (Kim, B., 2000)(Godwinjones, R. 2009) to the annotators.

These methods help to relief the human burden in annotation of some fields. However in some specific task, transcripts and speech boundaries are insufficient. For example, some CAPT (Computer-Aided Pronunciation Training) systems could detect mispronunciation and provide multi-level feedback (e.g., pronunciation score and phone substitution) to guide L2 learners to practice their pronunciation(Yingming Gao, 2015) (Yanlu Xie, 2016) (Leyuan Qu, 2016).

The performance of these systems is highly dependent on the quality of phone-level labeling of the non-native corpus. In fact, labeling non-native speech data is much more challenging than labeling native speech data, especially facing non-native mispronunciations. Moreover, as phonetic annotation is a subjective task, the familiarity of annotation conventions and psychological factors will also greatly affect the annotation consistency rate. Therefore, it is necessary to develop an automatic speech annotation system to assist human annotation.

Our recently proposed CAPT framework requires a large amount of phoneme labels, so this paper mainly focus on labeling phoneme-level mispronunciation patterns. Therefore, in this paper we attempted to use state-of-the-art ASR models based annotation system to automatically label a Chinese L2 speech corpus, then annotators were asked to check the detection results and make a final annotation.

Due to the difficulty of labeling non-native mispronunciations, the percentage of consistency between annotators is not always so high. Also the ground truth of the phone-level labeling is controversial. The performance of the annotation could not be indicated by the consistency merely. In order to measure the labeling precisely, a posterior probability annotation evaluation method is proposed.

The rest of this paper is organized as follow: Section II presents annotation framework, including automatic labels, manual labels and annotation evaluation criterion. Section III gives a brief description of the annotation corpus. Section IV shows experiments and results. Conclusions are given in Section V.

2. Annotation Methods

The annotation procedure could be divided into two parts: automatic label and manual label. In the automatic label part, the automatic speech recognition system will identify the possible erroneous (segmental and tonal) and label them. In the manual labeling part, human will decide which erroneous will be labeled finally.

2.1 Automatic Label

The Automatic label procedure will automatically label the boundaries and the possible erroneous phones.

Firstly an automatic speech recognizer is used to force-align the speech data into phonetic segments of Initials and Finals, and different levels of phonetic boundaries are assigned properly(Cao W 2010). After automatic mispronunciation detection is done, the erroneous phones which speech recognizer identified are transcribed in

Pinyin. Thus the mispronunciation types are labeled to assist annotators in making the final decision.
The illustration of the detection system is provided in Figure 1. In the acoustic module, we compare Long-Short Term Memory (LSTM) and Chain model which are both the state-of-the-art methods used in the ASR system.
The chain model we used in this study was introducted firstly by Povey et al. (D.Povey, Vijayaditya Peddinti, 2016) named as 'lattice-free maximum mutual information' (LF-MMI). The chain model has several differences, compared to traditional DNN-HMM model. This model use a three time smaller frame rate at the output of the neural network, which can significantly reduce a quantities of computation required in the test time and make real-time decoding much faster. Because of reducing the frame rate, unlike convential HMM topology, this model use a topology that can be traversed in one frame. During the training procedure of chain model, a forward-backward algorithm is run to estimate the sequence corresponding to the transcript (Juho Leinonen et al, 2018).
In the decoding module, we substitute the original grammar with an expanded grammar according to the length of input speech. For example, the input speech contains two syllables, then the corresponding expanded grammar will be limited to give a detection result with two syllables. After decoding, we will obtain Top-2 results based on the likelihoods of the output layer. The top two recognized results of the two models are given to annotators as the reference. Because of lacking of inter-Chinese databases, we used some Chinese databases to train the acoustic model.

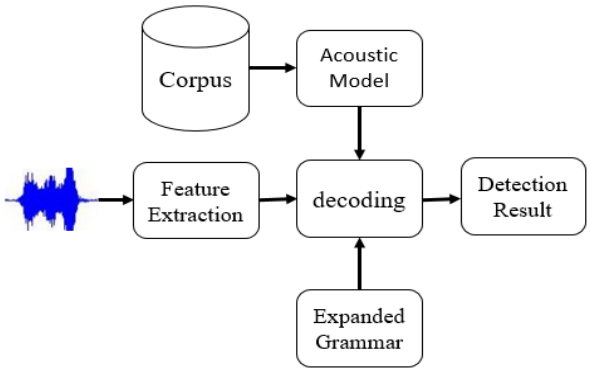

Fig.1 Flow chart of the detection framework

2.2 Manual Label

With the automatic labels, annotators will further check and label the data. The annotation problems are converted from the open-ended questions to multiple-choice questions with the method
Firstly annotators will judge if Initials or Finals in a bi-syllable word is similar with native's pronunciation or not. Then annotators will label the bi-syllable using Pinyin. If Pinyin could not remark the erroneous, PET diacritics will be used to describe the erroneous tendencies(Cao W 2010). For instance, as to the word 'ba ba'(father), if the pronunciation of 'a' is not native-like enough, and is more like 'e' in Chinese. Thus 'e' is used to indicate that the error sound is between 'a' and 'e'. If 'a' sounds like a native-like 'e', '/e' is used to reveal that 'a' is replaced by a standard Chinese 'e' sound. If the place of articulation of 'a' is too far behind, '{-}' is used to show backing of 'a' according to PET annotation conventions(Cao W 2010). All the work is deal with the software Praat.
A annotation example is shown in Fig.2. The first four tiers are corresponding to the orthographic given by speech recognizers. The fifth and the sixth tiers are respectively automatic results and manual annotation tier. In the sixth tier, there are two kinds of annotation symbols. If an error is described in Pinyin, it is annotated outside curly braces '{}'. On the contrary, PET diacritics are entered into '{}'.

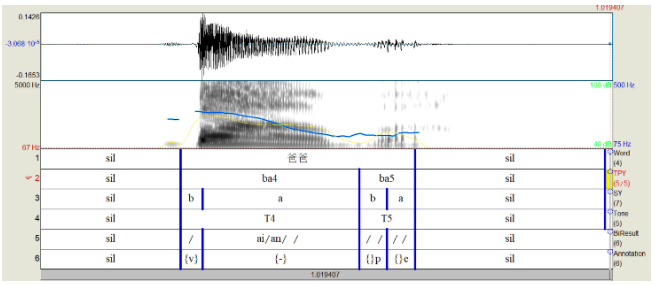

Figure 2. An annotation example

2.3 Posterior Probability Annotation Evaluation

The Mean consistency rate (MCR) with respect to each pair of annotators is widely applied in measuring the annotations results and the agreements. However the consistency rate is not comprehensive in measuring the binary classification. As an extreme case, if the erroneous is very little and one annotator is lazy and labels zero erroneous. The consistency rate will also be high.
In statistical analysis, the F1 score is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score.

$$F_1 = \frac{2\,Precision*Recall}{Precision+Recall} \qquad (1)$$

The F1 score assumes that false negatives, true negatives, false positives and true positives are certain. But in terms of annotation, especially for non-native mispronunciations, the four values are not certain. Thus we proposed Posterior F1(F1p). F1p could measure the F1 score and the consistency rate together. So the uncertainty could be considered in the formula.

$$F_1p = \frac{2\,Precision*Recall}{Precision+Recall} * MCR \qquad (2)$$

3. Annotation Corpus

The annotation corpus used here is BLCU-SAIT corpus. BLCU-SAIT is an interlanguage speech corpus aiming at Chinese learning. This corpus is composed of four sections which cover most of the Chinese phoneme types and tri-tone types bounded by prosodic boundary using a 103 sentence set.
The corpus is divided into four parts.

(1)103 declarative sentences.
(2)237 bi-syllable words. These words cover 97% of segmental phonemes and all the 20 kinds of bi-tone types in Mandarin.
(3) 1520 tonal syllables .
(4)A discourse (The North Wind and the Sun).

302 non-native speakers have been recruited to record the corpus. The mean age of all the speakers is 23 years with a standard deviation of 4.1 years. In that, 66% are female speakers, 34% are male. All the speakers stay in China and have studied Chinese for a few years.
The corpus was recorded in studio using USB M-audio sound card, a SHURE microphones, a software of recorder PC3.0. The data was recorded into 16 bits pulse-code modulation (PCM), sampled at 16 kHz. In order to minimize the effect of growing familiarity with the order of difficulty affecting the quality of the recording, the subpart of the data was presented in a randomized order.

4. Annotation Results

18 native speakers from north China are selected to annotate the corpus. The annotation is divided into two phases. In the first phase, 237 bi-syllable words spoken by 156 speakers are annotated. There are totally 44,304 words had been annotated. The speakers were from four countries shown as table 1.

Table 1: Speaker numbers of annotated data

		Speaker number
Country	Korea	19
	Russia	44
	Japan	45
	Kazakhstan	48
Totally number		156

In the automatic label phase, the bi-syllable words are decoded by the Long-Short Term Memory (LSTM) and Chain models. The acoustic feature used in this study is Mel-Frequency Cepstral Coefficient (MFCC). The input feature is a 39-dimension MFCC+Δ+ΔΔ vector. After forced-alignment, context-dependent (CD) feature labels are used to train corresponding neural network, which containing 6 hidden layers and 625 nodes. The Long-Short Term Memory (LSTM) and Chain models are trained with non-native speech corpus such as BLCU inter-Chinese corpus (Cao W 2010). Because of lacking of inter-Chinese databases, we also used several native speech corpus to train the acoustic model, including the Chinese National Hi-Tech Project 863, which containing 94000 utterances spoken by 160 speakers at about 100 hours (Sheng Gao 2010), and THCHS-30 (Dong Wang 2015), etc.
In the manual label phase, two annotators of each speaker are randomly assigned to avoid pairing effects. Each annotator will judge the automatic labels and label the bi-syllable using Pinyin. The third annotator will check and verify two annotators' results. It took about five months to finish the phase.
The final annotation results are shown in table 3, figure 3 and figure 4.

Table 2: Phoneme annotation results

	MCR (Mean consistency rate)	F1-a1	F1p-a1	F1-a2	F1p-a2
Japan	85.9%	0.981	0.842	0.981	0.843
Korea	87.1%	0.995	0.867	0.995	0.866
Kaza-khstan	87.6%	0.981	0.860	0.981	0.860
Russia	88.2%	0.972	0.858	0.972	0.857
means	87.2%	0.982	0.857	0.982	0.857

The consistency rate of phoneme annotations with respect to each pair of annotators was evaluated in percentage agreements. The ratios range from 72% to 97% and average as 87.2%. As shown in table 3 and figure 2. The results can be regarded as good for the nature of phonetic labels. Compared with the previous manual annotation results, the consistency rate of the two annotators in this study raised from 80.7% to 87.2%, the consistency rate is improved remarkably(Cao W 2010). The main reason of the improvement maybe that annotation problems are converted from the open-ended questions to multiple-choice questions. The annotators could choose the right answers from the automatic labels.

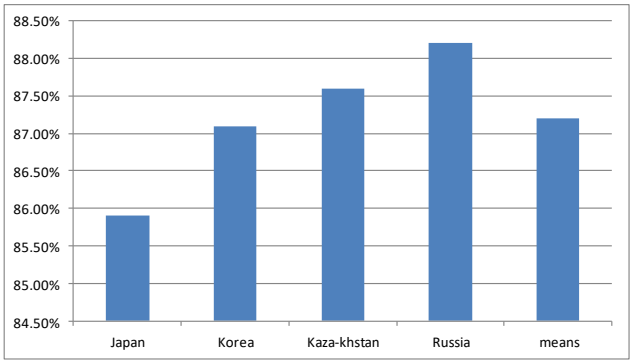

Fig.3 Mean consistency rate of each annotators

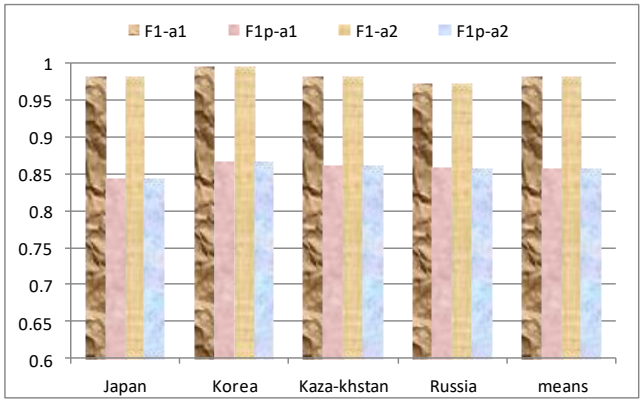

Fig.4 two F1 scores of the two annotators

Furthermore, F1 score is calculated to evaluate the performance. Since the standard answers of mispronunciation are unknown. Granted that the third annotator's label result is the ground truth. Thus we can get two F1 score. F1-a1 and F1-a2 are the F1 score of the first annotator and the second annotator respectively. As shown in table 2 and figure 4, F1-a1 and F1-a2 are extremely high. It means the difference between the two annotators is slight. In fact the third annotator's label result is still unreliable. F1-a1 and F1-a2 is not the reliable scores.

In order to measure the results more reliable, Posterior F1 is proposed. From formula (2), F1p-a1 and F1p-a2 are calculated. The results show that the new F1 drops a little as to the original F1. The ground truth is unknown as to the label problem. So the original F1 is not so precisely. The real F1 is unable to be calculated and shall be smaller than the original F1. Thus the Posterior F1 will be more reasonable. It considers the variance between the third annotator's label result and the ground truth. Even the mean consistency rate(MCR) is not equal to the actual variance. It is proportional to the actual variance. Thus the F1p-a1 and F1p-a2 could reflect the true F1. F1p-a1 and F1p-a2 is similar. It shows that the two annotators' performance is similar. The method could reduce the label ability between the annotators and make the results more objective.

5. Conclusion

This paper mainly focus on labeling phoneme-level mispronunciation patterns. In order to lighten the workload of human annotators, we attempt to use state-of-the-art ASR models based annotation system to automatically label a Chinese L2 speech corpus, then annotators could check the detection results and make a final annotation.

156 speakers' bi-syllable from 4 language backgrounds as pilot data were manually labeled, using both Pinyin, and the PET labeling system. The results show that mean consistency rate of manual labels is 87.2%. The posterior F1 score is 0.857. The results consistency rate is higher than previous report. So the annotation database could applied in the CAPT system. The posterior F1 score shows that the performance of the annotation could be improved further. Also the alternative labels annotated by the ASR models could help human annotators making final decision. They could choose one from four answers. Without the alternative labels, they will choose one from all the initials and finals. And the automatic labels also provide the possible mispronunciation for the human annotators. The annotation problems are converted from the open-ended questions to multiple-choice questions with the method.

In the near future, further efforts will be made to improve the system and more data will be used to develop CAPT systems. Also, the other part of BLCU-SAIT corpus, such as 103 declarative sentences will be labeled.

6. Acknowledgements

This work is supported by Wutong Innovation Platform of Beijing Language and Culture University (16PT05), Advanced Innovation Center for Language Resource and Intelligence(KYR17005), Research Funds of State Language Commission(ZDI135-51). The first author is corresponding author.

7. Bibliographical References

B. MacWhinney,(2000) The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates,

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. Speech Communication,, 23-60.

Cao W, Wang D, Zhang J, et al.(2010) "Developing a Chinese L2 speech database of Japanese learners with narrow-phonetic labels for computer assisted pronunciation training". Eleventh Annual Conference of the International Speech Communication Association, .

Chen, N. F., Wee, D., Tong, R., Ma, B., & Li, H. (2016). Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin. Speech Communication,, 46-56.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur (2016), "Purely sequence-trained neural networks for ASR based on lattice-free MMI". INTERSPEECH.

Godwinjones, R. (2009). Emerging Technologies Speech Tools and Technologies.. Language Learning & Technology, 13(3), 4-11.

Juho Leinonen, Peter Smit, Sami Virpioja, Mikko Kurimo (2018), "New Baseline in Automatic Speech Recognition for Northern Sámi". Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages.

Kim, B., Lee, J., Cha, J., & Lee, G. (2000). POSCAT: A Morpheme-based Speech Corpus Annotation Tool.. language resources and evaluation.

Leyuan Qu, Yanlu Xie, Jinsong Zhang (2016) , "Senone log-likelihood ratios based articulatory features in pronunciation erroneous tendency detecting" in Proc.ISCSLP

Marisa Casillas etc(2017)A New Workflow for Semi-automatized Annotations: Tests with Long-Form Naturalistic Recordings of Childrens Language Environments INTERSPEECH , Stockholm, Sweden

Minematsu, N., Tomiyama, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M., & Makino, S. (2002). English Speech Database Read by Japanese Learners for CALL System Development.. language resources and evaluation.

S. Sheng Gao, Bo Xu, Hong Zhang, et al (2010), "Update of Progress of Sinohear: Advanced Mandarin LVCSR System At NLPR," in Proc. ICSLP.

Wang, D., & Zhang, X. (2015). Thchs-30 : a free chinese speech corpus. Computer Science.

Yanlu Xie, Mark Hasegawa-Johnson, Leyuan Qu, Jinsong Zhang (2016) "Landmark of Mandarin nasal codas and its application in pronunciation error detection" IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Yingming Gao, Yanlu Xie, Wen Cao, Jinsong Zhang, (2015)"A Study on Robust Detection of Pronunciation Erroneous Tendency Based on Deep Neural Network", INTERSPEECH .