Manual and Semi-manual Comparison in the Annotation of CAPT Speech Corpus

Yanlu Xie, Xing Wei, Wei Wang,
Beijing Advanced Innovation Center for Language Resources Beijing Language and Culture University, Beijing 100083, China
xieyanlu@blcu.edu.cn

Abstract. Annotation plays an important role in a robust computer aided pronunciation training (CAPT) system. Manual annotation is time and annotator consuming. Automatic annotation is more efficient, but the accuracy is lower. This paper proposes to provide phoneme level labelling candidates with ASR models, and compare this method with manual annotation. In the previous work, phoneme level mispronunciation patterns were modelled and detected to provide readable pronunciation erroneous tendencies (PETs) feedback. In this paper, it is proposed to use this model to provide phoneme level labelling candidates, then annotators could choose the appropriate labels and make final decision. Experimental results show that compared with manual annotation, the consistency rate of semi-manual annotation increased from 88.52% to 92.88%. In addition, the false positive rate (FPR) reduced by 3%, the posterior F1 score declined more than 3%. The reported results demonstrated the efficiency of the proposed method.

1. Introduction
In recent years, Computer Aided Pronunciation Training (CAPT) system has received more and more attentions, where multi-level feedback (e.g., pronunciation score and phone substitution) is given to guide second language (L2) learners practice their pronunciation. More and more inter-language speech databases are developed for a robust CAPT system, and the scale of database becomes larger. For instance, the Cross Towns consists of 72,000 utterances from 161 speakers who are from 5 different counties [1], and the iCALL consists of 90,841 utterances from 305 speakers with a total duration about 142 hours [2]. The number of people learning Mandarin Chinese as a second language (L2) is increasing, yet over two-third of the non-native speech corpora focus on English as target language [3-5], and there is less focusing on non-native Mandarin [2,6]. In addition, the size of these corpora is small, which means it is hard to be used in a CAPT system.

Annotation plays a foremost role in speech database [7], especially for an inter-language corpus. The traditional and most adopted method to annotate speech data is manual annotation. Although manual annotation can achieve a high accuracy, it requires a lot of manpower and material resources, meanwhile, has higher requirements to annotators [7]. The second way is automatic annotation. For instance, CHAT (Codes for the Human Analysis of Transcripts) can be used to produce and analyze data, DARCLE Annotation Scheme (DAS) proposed a workflow for annotating long natural language recordings, SLAM and Speech Analyzer POSCAT even could automatically give phone-level labels [8-10] to the annotators. Compared with manual annotation, automatic annotation with ASR method can save time and manpower significantly. However, limited by the recognition result of ASR model, the accuracy rate of automatic annotation is not as good as manual annotation.
In fact, labeling non-native speech data is much more challenging than labeling native speech data, especially facing non-native mispronunciations, which is pronounced by using irregular articulation manner and place. 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. In the previous work, it is built an articulatory features based mispronunciation detection framework to detect PETs, so this paper mainly focuses on labeling phoneme level mispronunciation patterns [11].

In this study four annotators are divided into two groups. The first group would check the automatic annotation results achieved before and make final decision, and the second group is manual annotation. Because the ground truth of the phone-level labeling is controversial, the performance of annotation could not be indicated by the consistency rate merely. In order to measure the labeling result precisely, a posterior probability annotation evaluation method is used to evaluate annotation performance [12]. In addition, in order to further evaluate the labeling performance of each annotator, precision rate, recall rate and false positive rate (FPR) evaluation method is adopted.

The rest of this paper is organized as follows: Section 2 presents annotation convention and procedure, Section 3 shows experiments and results, and conclusions are given in Section 4.

2. Annotation Method

2.1. Annotation convention
The usual method to annotate corpus is to transcribe sounds in symbols as IPA annotation does. Such method does not make any annotation to the pronunciation errors, which has great restriction to CAPT system. Therefore, in this study, it is suggested that it shall annotate erroneous articulation tendencies instead of transcribing error sounds faithfully [13]. For example, if a Chinese Second Language (CSL) learner has problems of rounded “e”, a diacritic {o} can be used to indicate the erroneous tendencies of lip rounding.

2.2. Automatic annotation procedure
In previous work, an articulatory features based mispronunciation detection system has been built to label boundaries and possible erroneous phonemes automatically. The illustration of the detection system is provided as an example in [11].

The feature extraction module is comprised of a bank of articulatory feature classifiers. Expanded frames of input speech are fed into each model, generating frame likelihoods to each possible articulatory feature within that category. Subsequently, equation (1) is used to calculate phone level log posterior [14]. Then, the top-N consistency rate is achieved to indicate the consistency of all articulatory features within that category. Finally, this top-N results are given to annotators for checking and making secondary annotation. Equation (1) is show as follow:

$$\log P(p|O;t_s, t_e) = \frac{1}{t_e-t_s} \sum_{t_s}^{t_e} \log \sum_{s \in \text{set}} P(s|O_t)$$

where $p$ is the input feature at frame $t$, $t_s$ and $t_e$ are the start and end time of unit $p$, $P(s|O_t)$ is frame level likelihood, {set} is the set of context-dependent units, whose central unit is $p$.

2.3. Manual annotation procedure
In this study, four phonetic graduate students participated in manual annotating. These four annotators are divided into two groups. The first groups is based on the results of automatic mispronunciation detection model, and the second group is manual annotation.

An inventory of most frequent mispronunciations of Japanese speakers will be provided to each annotator for reference beforehand. Only mispronunciation phones detected by the models are given to annotators, while marked with number for convenient. Each annotator of the first group is asked to check the result first and annotate any mispronunciation errors at the 3rd tier according to the annotation convention.
The annotation example of the second group is the same as the first group, but without the mispronunciation detection result. Annotators in second group would make final decisions by themselves. There are no time limits for the annotators to work on each utterance. Besides, all the annotators in this study will work on the same data. As a result, each utterance was annotated four times. All the work has been done with software “Praat 6.0.26”.

3. Experiments and Results

3.1. Annotation corpus

The annotation corpus used here is a large scale Chinese second language (L2) speech database, which can be referred to as BLCU-Chinese speech corpus [13]. 380 sentences are selected from this corpus, table 1 gives some overall statistics of data used.

| Table 1: Japanese L2 Inter-Chinese Corpus |
|------------------------------------------|
| Text | 301 utterance |
| Speaker | 7 females |
| Number of utterance | 380 |
| Number of phonemes | 5320 |
| Average length per utterance | 14 |
| Number of types of PETs | 65 |

3.2. Evaluation metrics

3.2.1. Mean consistency rate

The annotation labels can be classified into four groups in view of consistency:

- 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.

\[ MCR = \frac{CC + CM}{CC + CM + IM + WM} \]  

3.2.2. Posterior Probability Annotation Evaluation

Higher MCR may come from higher CC or CM, as an extreme situation, if there are very few pronunciation errors and annotators label zero errors, the consistency rate will also be high. In statistical analysis, the F1 score takes both Precision and Recall in consider as a measure of a test’s accuracy. Hence, in this paper, the original annotations are taken as ground truth and compare with the results of four annotators in this study. Therefore, the annotation labels can be classified into four groups:

- True Positive (TP): Correct pronunciation labeled as correct pronunciation.
- True Negative (TN): Mispronunciation labeled as mispronunciation.
- False Positive (FP): Mispronunciation labeled as correct pronunciation.
- False Negative (FN): Correct pronunciation labeled as mispronunciation.

According to these four results, F1 score, Precision and Recall, are calculated. The Precision, in this study, means the percentage of real correct percentage in all the labeled correct pronunciation, the Recall represents the percentage of labeled correct pronunciation in all the correct pronunciation. The corresponding formula is show as follow:

\[ Precision = \frac{TP}{TP + FP} \]
Recall = \frac{TP}{TP + FN} \quad (4)

F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)

However, in terms of non-native mispronunciation annotation, true positives, true negatives, false positives and false negatives are not certain. Thus, Posterior F1 (F1p) [12] is used to measure the F1 score and the MCR together, the formula is shown as follow:

\[ F_{1p} = \frac{2 \times Precision \times Recall}{Precision + Recall} \times MCR \quad (6) \]

In addition, False Positive Rate (FPR) is used to evaluate the percentage of unlabeled mispronunciations, the formula is shown as follow:

\[ FPR = \frac{FP}{FP + TN} \times 100\% \quad (7) \]

3.3. Annotation result

Compared with the previous manual annotation results, the mean consistency rate of the two groups in this study raised from 80.7% to 92.88% and 88.52% respectively, the consistency rate was improved remarkably. The detail of the MCR results are showed in Figure 1. In Figure 1, the original, T1 and T2 represents MCR result of previous manual annotation, group one and group two in this study, respectively. From Figure 1, it is found that the inconsistent labels of the first group is over 7%, in which about 5% are only labeled by one annotator. Besides, the inconsistent rate of the second group is upon 11%, among which about 2% are treated as mispronunciation but labeled differently, the rest are only labeled by one annotator. With the assistance of automatic annotation, the MCR of the first group is higher than the second group. Compared with original manual annotation results, two groups in this study achieved higher MCR is mainly because the consistent rate of these two groups is higher than original manual annotation.

![Figure 1: Mean Consistency Rate Results](image)

In addition, the result of four annotators in this study with original manual labels is compared, and the distributional percentage of phoneme labels are showed in Figure 2. In Figure 2, T11-original represents the consistency rate between the first annotator of the first groups in this study and original manual annotations, T21-original represents the consistency rate between the first annotator of the second group in this study with original manual annotations. The results are analyzed as:

- For the two annotators of the first group, among the overall over 79% (CC+CM) consistent labels, over 76% for correct phonemes, and left about 3% are treated as mispronunciations. The total inconsistent labels account for over 19% (WM+IM), in which about 2% are treated as mispronunciation but labeled differently, the other beyond 17% was labeled by only one annotator.

- For the two annotators of the second group, more than 76% (CC+CM) are consistent labels, upon 73% for correct phonemes, and left about 3% for mispronunciations. Besides, the inconsistent labels take up over 21% (WM+IM), including more than 2% for inconsistent mispronunciation and about 20% for warning mispronunciation.
Compared with the two annotators of the second group, the annotators of the first group achieved higher consistent rate and lower inconsistent rate.

**Fig 2: Distribution of Phoneme Labels**

According to Figure 2, compared with original manual labels, the percentage of warning mispronunciation of all annotators in this study is a bit high. As mentioned before, the mean consistency rate was not comprehensive in measuring the binary classification, original manual labels are taken as the ground truth and calculate F1p to evaluate the annotation quality of the four annotators. In addition, FPR is calculated to measure the percentage of unlabeled mispronunciations of each annotator, the calculation results are shown in Table 2.

**Table 2: F1p and FPR Results**

|        | T11-original | T12-original | T21-original | T22-original |
|--------|--------------|--------------|--------------|--------------|
| Precision | 84.31%       | 84.54%       | 82.81%       | 81.16%       |
| Recall  | 96.26%       | 95.45%       | 96.15%       | 94.62%       |
| F1      | 89.89%       | 89.66%       | 88.98%       | 87.37%       |
| F1p     | 72.13%       | 71.44%       | 68.04%       | 66.82%       |
| FPR     | 71.46%       | 73.96%       | 75.93%       | 76.01%       |

Table 3 shows that compared with the second, annotators in the first group have better Precision, Recall and F1 score, which means annotators in the first group have better accuracy. After measure the F1 score and the mean consistency rate together, the F1p has a distinct decline. However, compared with the annotators of the second group, the results of the annotators of the first group are better. In addition, the FPR of all four annotators are high which means there are lots of mispronunciations are treated as correct pronunciation by these four annotators. Like the results of F1p, the FRP of the annotators of the first group are about 3% less than the second group. Furthermore, the distribution of unlabeled mispronunciations are counted by the four annotators, the results are showed in Figure 3.

As show in Figure 3, the annotators of the first group have less unlabeled mispronunciations than the second group. For the first group, unlabeled mispronunciation account for around 35% are Initial of Chinese phonemes, left about 65% are Finals. For the second group, about 45% of the overall are Initials.

**4. Conclusions**

This paper mainly focuses on improving phoneme level annotation efficiency of inter-Chinese corpus. Two methods named manual annotation and semi-manual annotation are explored. Compared with manual annotation, the consistency rate of the semi-manual method raised from 88.52% to 92.88%. In addition, it is found the FPR of semi-manual method reduced by about 3%. The results illustrate that compared with manual annotation, semi-manual annotation can improve the consistency rate and of phoneme level annotation significantly, while providing more labeled speech data within a limited time.

In the near future, further efforts will be made to improve the system and more data will be used to develop CAPT system.

**Acknowledgements**
This work is supported by Intelligence Research Funds of State Language Commission(ZDI13-51) and the Fundamental Research Funds for the Central Universities (18YJ030004).

References
[1] Chen N F, Tong R, Wee D, et al. iCALL corpus: Mandarin Chinese spoken by non-native speakers of European descent[C], Sixteenth Annual Conference of the International Speech Communication Association, 2015.
[2] Carletta J, Ashby S, Bourban S, et al. The AMI Meeting Corpus: A Pre-announcement [M], Machine Learning for Multimodal Interaction. Springer Berlin Heidelberg, 2006:28–39.
[3] Christopher Cieri, David Miller, Kevin Walker, The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text, Proc. LREC 2004.
[4] S. Pigeon, W. Shen, and D. van Leeuwen, Design and characterization of the non-native military air traffic communications database, in ICSLP, Antwerp, Belgium, 2007.
[5] Yow-Bang Wang and Lin-Shan Lee, “Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 5049–5052.
[6] Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. Speech Communication, 23-60.
[7] Zhang B. The State of the Art, Problems, and Coping Strategies of the Chinese Interlanguage Corpus Development[J]. Applied Linguistics, 2010.
[8] Pye C. The CHILDES Project: Tools for Analyzing Talk, by Brian MacWhinney[J]. Child Language Teaching & Therapy, 2000, 8(2):156.
[9] Godwinjones R. Emerging Technologies Speech Tools and Technologies.[J]. Language Learning & Technology, 2009, 13(3):4-11.
[10] Cha J. POSCAT: A Morpheme-based Speech Corpus Annotation Tool[J].
[11] Wei X, Chen J, Wang W, et al. A study of automatic annotation of PETs with articulatory features[C], Asia-Pacific Signal and Information Processing Association Summit and Conference. 2017:1608-1612.
[12] Xie Y, Wei X, Wang W, et al. A Semi-manual Annotation Approach for Large CAPT Speech Corpus [C], 11th edition of the Language Resources and Evaluation Conference, 7-12 May 2018.
[13] Cao W, Wang D, Zhang J, et al. “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. 2010.
[14] W. Hu, Y. Qian, F. K. Soong and Y. Wang. “Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning based Logistic Regression Classifiers”. Speech Communication, 67, pp. 154-166, 2015.