A Study on Robust Detection of Pronunciation Erroneous Tendency Based on Deep Neural Network

Yingming Gao[1], *Yanlu Xie*[1], *Wen Cao*[2], *Jinsong Zhang*[1,*]

[1] College of Information Sciences, Beijing Language and Culture University, Beijing 100083, China
[2] College of Chinese Studies, Beijing Language and Culture University, Beijing 100083, China

gaoyingming1@sina.com, xyl@blcu.edu.cn, tsao@blcu.edu.cn, jinsong.zhang@blcu.edu.cn

10.21437/Interspeech.2015-242

Abstract

Compared with scoring feedbacks, instructive feedbacks are more demanded by language learners using computer aided pronunciation training (CAPT) systems, which require detailed information about erroneous pronunciations along with phone errors. Pronunciation erroneous tendency (PET) defines a set of incorrect articulation configurations regarding main articulators and uttering manners for the phones respectively, and its robust detection contributes to the provision of appropriate instructive feedbacks. In our previous works, we designed a set of PET labels for CSL (Chinese as a second language) by Japanese learners, and conducted a preliminary detection study with GMM-HMM. This study is aimed at achieving a more robust detection of PETs by two approaches: employing DNN-HMM as the acoustic modeling, and comparing three kinds of acoustic features: MFCC, PLP, and filter-bank. Experimental results showed that the DNN-HMM PET modeling achieved more robust detection accuracies than the previous GMM-HMM, and the three kinds of features behaved differently. A lattice combination of the results of three feature systems led to the best PET results: FRR of 5.5%, FAR of 35.6%, and DA of 88.6%, which showed its efficiency.

Index Terms: CAPT, pronunciation erroneous tendency (PET), mispronunciation detection, deep neural network

1. Introduction

Computer aided pronunciation training (CAPT) systems based on automatic speech recognition (ASR) technology have been attracting considerable attention in recent years. From the view of feedback form, the CAPT systems can be roughly divided into two categories: pronunciation scoring and pronunciation error detection. Pronunciation scoring has made substantial progress since the confidence scores can be obtained fairly easily with an ARS system without confining learners' L1 background. Researchers investigated pronunciation scoring from speaker level to phone level [1-3], and have achieved a high consistency with human experts at sentence level [2] and speaker level [3]. Those researches are more suitable for providing an indication of the candidate's proficiency, but not instructive enough to guide the learners to correct their mispronunciations.

Nowadays, more and more researchers are concentrating on pronunciation error detection, which is aimed at detecting or identifying pronunciation errors or deficiency in a high precision and providing instructive feedbacks. Neri et al. showed that implementing corrective feedback even if in a limited form, did improve the pronunciation quality of students on an individual phoneme level and had a positive impact on user's motivation [4]. In order to identify the location and type of learner's mispronunciations, Harrison et al. implemented a mispronunciation detection and diagnosis prototype which used extended recognition network [5]. In fact, learner's erroneous sound always deviates a little from the canonical sound more than the phone errors (insertion, deletion or substitution). Therefore pronunciation erroneous tendency (PET) was proposed to define a set of incorrect articulation configurations regarding main articulators and uttering manners for the phones respectively, and its corresponding diacritics were designed in our previous work [6]. Moreover, we implemented a PET detection system and evaluated its feasibility using GMM-HMM [7].

However, as an indispensable component of CAPT systems, mispronunciation detection accuracy at phone level remains to be raised. Recently, deep neural network (DNN) has not only reduced the speech recognition errors significantly, but also has been successfully applied to CAPT [8]. Some have applied the DNN to mispronunciation detection and diagnoses in L2 English language learners and obtained significant performance improvement [9-11]. Extending this approach to pronunciation quality scoring has achieved similar performance improvement [3]. Within the framework of our previous research [7], we explored the potential of DNN-based approach for PET detection. Furthermore, this study compared the effects of three kinds of acoustic features for the detection of specific categories of PETs. With the lattice combination technology, we integrated the three systems.

The paper is organized as follows: In Section 2, a brief description of pronunciation erroneous tendency (PET) and its annotation convention will be presented. Section 3 gives an overview of our detect system. This is followed by experimental results and discussion in Section 4. The paper is concluded with our directions for future work in Section 5.

2. PET definition and annotation

2.1. PET definition

There are some general and salient mispronunciation patterns when CSL learners speak Chinese [12-14]. Most of them result from an inaccurate place of articulation or erroneous uttering manners. When learners practice pronouncing the target speech, they usually use the similar articulation-placement existing in their mother tongue because of the influences of negative transfer of mother tongue. Also, learners have difficulties in mastering uttering manners which do not exist in their mother tongue. Learner's erroneous sound cannot be simply classified into insertion, deletion or

substitution [15]. Their mispronunciation always deviates a little from the canonical sound, rather than the absolute phoneme category substitution. Therefore pronunciation erroneous tendency (PET) defined a set of incorrect articulation configurations regarding main articulators and uttering manners for the phones respectively: raising, lowering, advancing, backing, lengthening, shortening, centralizing, rounding, spreading, labio-dentalizing, laminalizing, devoicing, voicing, insertion, deletion, stopping, fricativizing, nasalizing, retroflexing and so on [6].

2.2. Annotation convention

A set of diacritics were designed for different kinds of general PETs. Several diacritics can also be combined to represent a complex sound variant. Directed by the convention in BLCU-CAPT-1 [6], multi-level phonetic transcriptions including words, syllables, Chinese traditional "phonemes" of "Initials" and "Finals", lexical tones, and high-level prosody events, were annotated. This study only applied the phoneme tier. Table 1 gives a small part of the PETs annotation description in our convention.

Table 1. *Annotation Convention.*

PET	Diacritics	E.g.	Notation
Spreading	w	u{w}	Round sound "u" has a spreading lip
Backing	-	n{-}	The tongue position of phoneme is a little back
Shortening	;	p{;}	The aspiration duration of phoneme p is shorter
Laminalizing	sh	sh{sh}	Balade-palatal phoneme sh is pronounced as Japanese lamina-alveolar

3. Overview of detection system

3.1. Detection framework

The illustration of the detection system is provided in Figure 1 with an example. Firstly, the system prompts learners to speak a given utterance "两块五一斤(Two point five yuan per jin)", which corresponds to the Pinyin ("l iang k uai u i j in"). Secondly, according to the extended pronunciation network, records of learner's speech are recognized via the ASR-based detector. Then the system judges the sound based on the difference between the recognized phone-level transcription (" l iang k uai u{w} i j in") and the canonical one. At last, the corrective feedbacks ("please try to round your lip more when pronouncing the phoneme 'u' ") will be given to learners.

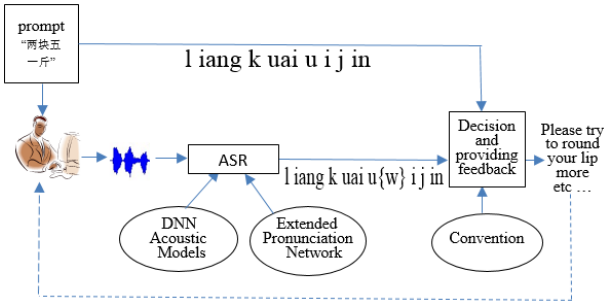

Figure 1: *Flow chart of our detection framework.*

3.2. Acoustic model training

A deep neural network is a feed-forward, artificial neural network with multiple hidden layers between its input and output. The architecture of deep neural network is shown in Figure 2. The DNN is initialized with the deep belief network (DBN) pre-training procedure [16]. A DBN can be efficiently trained in an unsupervised, layer-by-layer manner where the layers are constructed by stacking up multiple Restricted Boltzmann Machines (RBMs). Since input feature vector is a continuous variable, the first two layers (i.e. input and 1st hidden layers) are modeled as a Gaussian-Bernoulli RBM and other hidden layer pairs are modeled as Bernoulli-Bernoulli RBM. Then network can be discriminatively trained in a supervised way with back-propagation algorithm in order to "fine-tune" all the weights and bias. The state transition parameters are obtained from original GMM-HMM system.

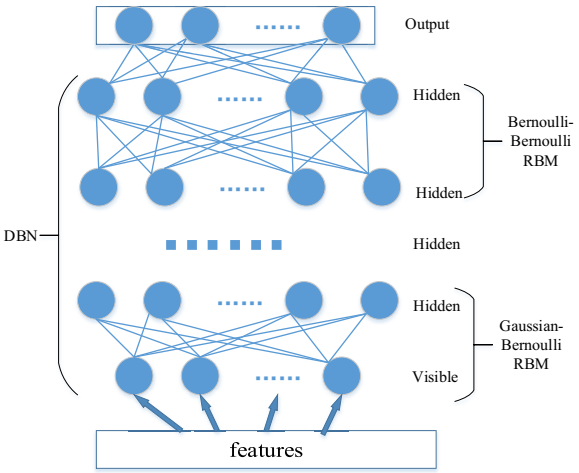

Figure 2: *Architecture of DNN.*

3.3. Extended pronunciation network

Extended pronunciation network is a representation of pronunciation variants in the form of a network. According to the possible mispronunciation in our annotation convention, we extended the pronunciation of every Chinese word (or Character) in the dictionary. When the system gives the prompts, the corresponding extended pronunciation network will be constructed with all possible pronunciations. One example of such network is shown in Figure 3.

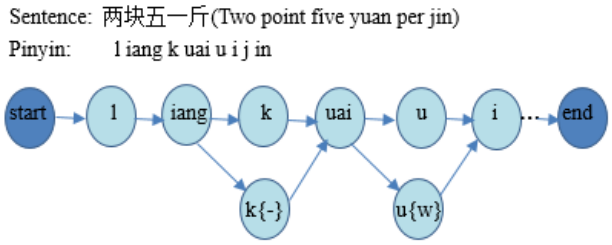

Figure 3: *Extended pronunciation network.*

4. Experiments

4.1. Experiment setup

We collected a large scale of Chinese L2 speech database, which can be referred to as BLCU inter-Chinese speech corpus [6]. In order to keep consistent with our previous research employing GMM-HMM as acoustic model [7], this study focused on the continuous speech of Japanese part. Table 2 gives some overall statistics of corpus. 80% of the data was used for training and the rest for testing.

We trained three kinds of acoustic models employing different acoustic features: Mel-Frequency Cepstral Coefficient (MFCC), Perceptual Linear Predictive Analysis (PLP), and filter-bank (FBANK). Compared with MFCC and PLP, FBANK contains more complete information. 13-dimemsion MFCC, 13-dimension PLP, and 23-dimension FBANK features, with their first and second order derivatives respectively, were extracted from utterances with a 20ms-length window shifted every 10 ms. 11 frame super-vector consisting of 5 preceding frames, current frame and 5 succeeding frames was used as the input of DNN.

Table 2. *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

We evaluated the detection performance with three kinds of metrics based on four outcomes in the experiment: True Acceptance, True Rejection, False Acceptance, and False Rejection. These metrics included:

- 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

Although there are 65 kinds of specific PETs, some of them are rare in the corpus and their corresponding acoustic models are also unreliable. Therefore this research choose the 16 most common mispronunciations in the experiment, which covers 61.4% of all the mispronunciation samples. As for the three kinds of metrics, while we aimed 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 critical to avoid discouraging learners by rejecting their correct pronunciations. Therefore we attached greater importance to DA and FRR while optimizing the performance. In Table 3, we compared the PET detection performance of the DNN-HMM-MFCC and the conventional GMM-HMM-MFCC which was employed in our previous work [7].Though a slight degradation existing in false acceptance rate (FAR), the new model obtained better performance in both false rejection rate (FRR) and diagnostic accuracy (DA). The DNN-HMM model would be used in the following PETs detection experiments.

Table 3. *Detection results of GMM and DNN models.*

Acoustic model	FRR	FAR	DA
GMM-HMM+MFCC	8.0%	32.6%	86.0%
DNN-HMM+MFCC	6.7%	35.9%	87.6%

With further analysis of the 16 most common PETs, we grouped them into four broad headings, i.e.

- Lip rounding and spreading: sounds with spreading lips have problems of rounding tendency or sounds with the rounding lips have problems of spreading tendency.
- Tongue advancing and backing: the tongue position of phonemes is a little advance or back
- Shortening: an insufficient aspiration or an inappropriate constriction.
- Laminalizing: balade-palatal phonemes are pronounced as Japanese lamina-alveolar.

We developed three systems with different kinds of acoustic features (MFCC, PLP, and FBANK) to compare detection performance on the four broad heading PETs. The statistic results are shown in Figure 4.

Figure 4 (a) illustrates that FBANK outperforms MFCC and PLP in detecting lip rounding and spreading tendency, because it has the higher DA and the lower error rates (both FAR and FRR). In the same way, (b), (c), and (d) illustrate that PLP is dominant in detecting tongue advancing and backing tendency, MFCC in detecting shorting tendency, and PLP in detecting laminalizing tendency, respectively.

The findings show that MFCC, PLP, and FBANK have different performances in capturing characteristics of specific PET, though they are all spectrum features. Therefore, we expected to further improve the performance by using system combination technology. We combined three lattices generated by different systems, then the resulting lattice was decoded. Table 4 shows the performance of MFCC, PLP, FBANK and the integration of these three kinds of acoustic features in detecting all the 16 most common PETs, and the combination system obtains the best result (FRR=5.5%，FAR=35.6%，DA=88.6%).

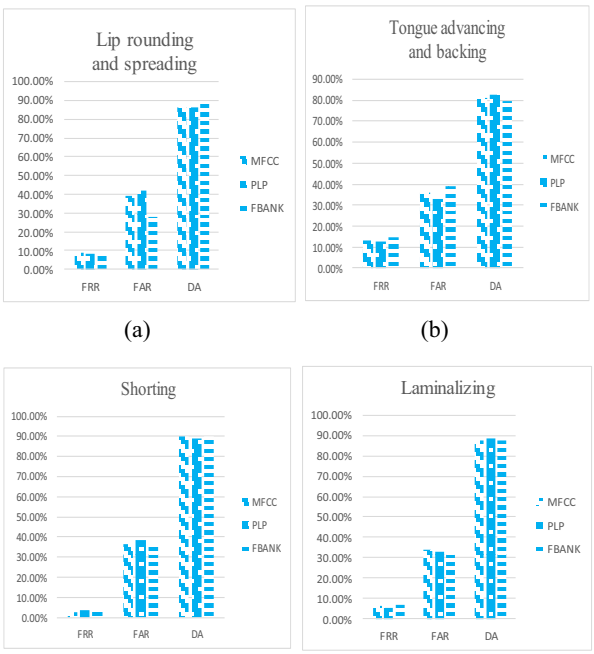

Figure 4: *Detection results of four broad heading PETs with different acoustic features.*

Table 4. *Detection results of different acoustic models.*

Acoustic model	FRR	FAR	DA
DNN-HMM+PLP	6.1%	39.4%	87.4%
DNN-HMM+MFCC	6.7%	35.9%	87.6%
DNN-HMM+FBANK	6.8%	34.6%	87.8%
System combination	5.5%	35.6%	88.6%

5. Conclusions

Aiming at implementing robust detection of PET, we employed DNN-HMM as acoustic modeling. The DNN-HMM achieved better detection rate over the GMM-HMM. Furthermore, we developed three detection systems with different kinds of acoustic features to compare their detection performance. The characteristics of the different PETs were captured by the specific acoustic feature more efficiently. With the lattice combination technology, we combined these three systems, and obtained the best results, achieving FRR of 5.5%, FAR of 35.6%, and DA of 88.6%. There are several areas where further research could be made: more distinctive phonetic features can be employed to detect the specific PET. Besides, based on phonological rules and statistics of mispronunciations, extended pronunciation lexicons can be constructed, which reflects how likely each type of error might occur as language models for ASR [5][17]. 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

This work is supported by National Nature Science Foundation of China (61175019), and Beijing Higher Education Young Elite Teacher Project (YETP0879). The asterisked author is the corresponding author.

7. References

[1] S.M. Witt and S.J. Young. "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication* 30.2 (2000): 95-108.

[2] J. Zheng, C. Huang, M. Chu, and F. K. Soong, "Generalized segment posterior probability for automatic Mandarin pronunciation evaluation," *Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[3] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," *INTERSPEECH,* 2013.

[4] A. Neri, C. Cucchiarini, and H. Strik, "ASR-based corrective feedback on pronunciation: does it really work?" *INTERSPEECH,* 2006.

[5] A. M. Harrison, W. K. Lo, X. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," *SLaTE,* 2009.

[6] W. Cao, D. Wang J. Zhang, and Z. Xiong, "Developing a Chinese L2 speech database of Japanese learners with narrow-phonetic labels for computer assisted pronunciation training," *INTERSPEECH,* 2010.

[7] R. Duan, J. Zhang, W. Cao, and Y. Xie, "A Preliminary study on ASR-based detection of Chinese mispronunciation by Japanese learners," *INTERSPEECH,* 2014.

[8] K. Li and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multi-distribution Deep Neural Networks," *Chinese Spoken Language Processing (ISCSLP),* 2014.

[9] W. Hu, Y. Qian, and F. K. Soong, "A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training," *Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[10] X. Qian, H. Meng, and F. K. Soong, "The Use of DBN-HMMs for Mispronunciation Detection and Diagnosis in L2 English to Support Computer-Aided Pronunciation Training," *INTERSPEECH*. 2012.

[11] W. Hu, Y. Qian, and F. K. Soong, "A new Neural Network based logistic regression classifier for improving mispronunciation detection of L2 language learners," *Chinese Spoken Language Processing (ISCSLP)*, 2014.

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

[13] F. Y. Li and W. Cao, "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.

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

[15] S. Y. Yoon, M. Hasegawa-Johnson and R. Sproat, "Landmark-based automated pronunciation error detection," *INTERSPEECH,* 2010.

[16] G. Hinton, S. Osindero and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation* 18.7 (2006): 1527-1554.

[17] D. Luo, X. Yang, and L. Wang, "Improvement of segmental mispronunciation detection with prior knowledge extracted from large L2 speech corpus," *INTERSPEECH*, 2011.