Correlational Neural Network Based Feature Adaptation in L2 Mispronunciation Detection

Wenwei Dong and Yanlu Xie
Beijing Advanced Innovation Center for Language Resources
Beijing Language and Culture University
Beijing, China
E-mail: dongwenwei_blcu@163.com, xieyanlu@blcu.edu.cn

Abstract—**Due to the difficulties of collecting and annotating second language (L2) learner's speech corpus in Computer-Assisted Pronunciation Training (CAPT), traditional mispronunciation detection framework is similar to ASR, it uses speech corpus of native speaker to train neural networks and then the framework is used to evaluate non-native speaker's pronunciation. Therefore there is a mismatch between them in channels, reading style, and speakers. In order to reduce this influence, this paper proposes a feature adaptation method using Correlational Neural Network (CorrNet). Before training the acoustic model, we use a few unannotated non-native data to adapt the native acoustic feature. The mispronunciation detection accuracy of CorrNet based method has improved 3.19% over un-normalized Fbank feature and 1.74% over bottleneck feature in Japanese speaking Chinese corpus. The results show the effectiveness of the method.**

Keywords-Computer-Assisted Pronunciation Training; Correlational Neural Network; Bottleneck feature

I. Introduction

With the accelerating of globalization, language learning becomes more and more important. CAPT provides an opportunity for learners to practice their pronunciation and give them feedbacks free from time and space limitations, and its main framework is similar to Automatic Speech Recognition (ASR). Because of the difficulty of annotating non-native data, acoustic model of mispronunciation detection is usually trained with native speaker's corpus and using it to evaluate L2 learners' pronunciation, thus there is a mismatch between them. Adding L2 learners' corpus in the training set can help mitigate the mismatch to some extent [1]. To further reduce the impact of it, the speaker and environment adaptation techniques are explored.

The adaptation techniques can be divided into model adaptation that tune-up the acoustic model parameters to fit the test set and feature adaptation that tune-up the feature before training acoustic model to make it more fitted for the test set. Most of model adaptation techniques are used in GMM acoustic models, such as maximum likelihood linear regression (MLLR) [2] and Maximum A Posterior (MAP), but they cannot be applied to DNN acoustic model which have a number of parameters that cannot be adapted with such small data. Many feature adaptation methods are used in DNN acoustic model.

Feature-space MLLR (fMLLR), Vocal Tract Length Normalization (VTLN) and i-vector are effective methods for reducing the mismatch of native and non-native speech [2, 3], Zhang [4] and Huang [5] used selective maximum likelihood linear regression (SMLLR) and MLLR to reduce the speaker difference, Luo [6] compared feature-space MLLR (fMLLR) and factorized fMLLR which factorizing the speaker and environment and other acoustic factors, factorized method performs better in CAPT tasks. Besides, some researchers used deep neural networks as a feature extractor to reduce the influence of mismatch, it can effectively model and feature represent. Gao [7] and Nicolao [8] used bottleneck feature that extracting from a neural network as the input of classifier to improve detection result. And a lot of works tried to add a rescoring or verification process to get further improvements [9−13]. Most of them are use single view neural network as a feature extractor.

With the great progress of neural network in various fields, researchers work on changing the model structure to get used to different tasks. Chandar [14] proposed CorrNet, it's a multi-views model based Common Representation Learning (CRL). CorrNet try to embed different descriptions (views) of data in a common subspace and maximize the relation of them. In this paper, we take CorrNet as feature extractor to adapt feature and reduce the mismatch, CorrNet is trained with a few L2 and L1 speech data, the model can learn the relationship between them. The paper is organized as follows: Section 2 presents two way of improving GOP measures and the frameworks of adding CorrNet. Section 3 introduces the experiment corpus and setup. Section 4 shows the experiment results and discussions, and the conclusions are drawn in Section 5.

II. The Framework of Mispronunciation Detection

In this section, we introduce the CorrNet and the new framework of generating posterior probability, then briefly review the traditional GOP method.

A. *CorrNet Framework*

Common Representation Learning (CRL) focus on embedding different description (or views) of the data into a common subspace, two popular paradigms are Canonical Correlation Analysis (CCA) based approaches and Autoencoder (AE) based approaches. CorrNet is a kind of AE approach. It's a multi-view model and aims to learn a common representation from two views of data. We use its loss function as follows:

$$\mathrm{Jz}(\theta)=\sum_{\mathrm{i=1}}^{N}\mathrm{L}(\mathrm{z_i},\mathrm{g}(\mathrm{h}(\mathrm{x_i})))+\mathrm{L}(\mathrm{z_i},\mathrm{g}(\mathrm{h}(\mathrm{y_i})))-\lambda\mathrm{corr}(\mathrm{h}(\mathrm{X}),\mathrm{h}(\mathrm{Y})) \tag{1}$$

where L is reconstruction error,

$$\mathrm{corr}(\mathrm{h}(\mathrm{X}),\mathrm{h}(\mathrm{Y}))=\frac{\sum_{i=1}^{N}(\mathrm{h}(\mathrm{x_i})-\overline{\mathrm{h}(\mathrm{X})})(\mathrm{h}(\mathrm{y_i})-\overline{\mathrm{h}(\mathrm{Y})})}{\sqrt{\sum_{i=1}^{N}(\mathrm{h}(\mathrm{x_i})-\overline{\mathrm{h}(\mathrm{X})})^2\sum_{i=1}^{N}(\mathrm{h}(\mathrm{y_i})-\overline{\mathrm{h}(\mathrm{Y})})^2}} \tag{2}$$

so we can use CorrNet in our task to benefit GOP in the following aspects:

- Minimize the self-reconstruction error of ASR result.
- Minimize the cross-reconstruction error of native and L2 Fbank.
- Maximize the correlation between the hidden representations of both view's Fbank.

Compared with other multi-task models, CorrNet tries to maximize the correlation of each view and use transfer learning to construct each other. In our task, acoustic model that trained with native speech will perform better in test set of native speech than non-native speech, it is a result of Fbank feature not just containing phone information, but also the speaker and channel et al, and it will affect the accuracy of mispronunciation detection. CorrNet has two input layers, common hidden layer, and two output layers, we can use native speech and L2 speech Fbank as two views to train model. The CorrNet can not only optimize the phone ASR result of each view but also learn the relations of them, the common layer can embed the feature of two kind databases into a common subspace, and it can maximize the phone information and reduce other differences to some extent. The main purpose of this task is detecting mispronunciation, the two output layers can still keep their difference in phone.

In my method, we use CorrNet as feature extractor and combine CorrNet feature and Fbank feature to train acoustic model, then use acoustic model's output of each phone to calculate the score, the framework shows in figure 1,

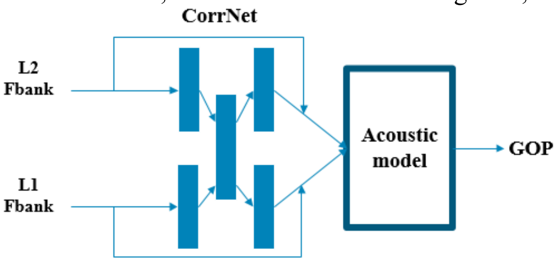

Figure 1. The framework of CorrNet based GOP

B. Bottleneck Framework

Bottleneck feature (BNF) in my experiment use TDNN as feature extractor, we set a bottleneck layer, and the nodes of this layer are less than other hidden layers. Aim to map a high dimensional vector into a low dimensional space with rich information, it also can reduce the mismatch of the acoustic feature, we combine BNF and Fbank to train acoustic model for mispronunciation detection.

We use different feature extractor to compare the ability of reducing native and non-native databases mismatch. The BNF framework is shown in figure 2.

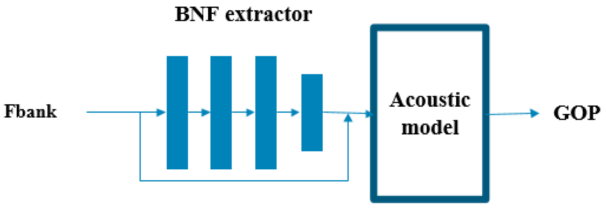

Figure 2. The framework of BNF based GOP

C. GOP Computing Methods

Witt et al proposed GOP in phone level scoring, it's text-dependent, for the acoustic segment X, target phone p,

$$\mathrm{GOP}(p)=\frac{1}{d}log\frac{P(X|p)P(p)}{\sum_{\{q\in Q\}}P(X|q)P(q)} \tag{3}$$

where d is the number of frames, *P(X|p)* is the likelihood of X corresponding to each phone q. Q is the set of phones. In DNN-HMM based system, assuming all phone's prior probability are equal and the sum of all phone posterior probability can be approximated by its maximum [13], in this paper we use the GOP,

$$\mathrm{GOP}(p)=\frac{1}{d}log\frac{P(p|X)}{max_{\{q\in Q\}}P(\mathrm{q}|X)+P(p|X)} \tag{4}$$

P(p|X) is the phone p's posterior probability of acoustic model, then we set a threshold to make the final decision,

$$\mathrm{GOP}(p)>k\begin{cases}\text{yes,} & \text{correct pronunciation}\\ \text{no,} & \text{mispronunciation}\end{cases} \tag{5}$$

those methods can be affected by alignment result of a model. In practice, we can adjust the threshold for language learners with different level. The main idea of GOP is using classifier's confidence score as the score of pronunciation.

III. Experiments

A. Speech Corpus

Experiment corpus consists of two parts, the native speech database is provided by the Chinese National Hi-Tech Project 863 for Mandarin continuous speech recognition of large vocabulary system development [15]. It contains 166 speakers about 100 hours. We divide it into the training set about 70 hours and development set about 30 hours, and no speakers overlap. We also use 3600 sentences of native Chinese corpus to test GOP algorithm. The non-native speech corpus is BLCU inter-Chinese speech corpus [16]. It has 19 speakers and each speaker has 301 sentences. We add 12 speakers of it as the training set to reduce the mismatch between native and non-native dataset, and 7 speakers corpus as the test set. The test set has been annotated at the phone level. The details are shown in Table 1.

TABLE I. JAPANESE L2 INTER-CHINESE CORPUS

Corpus	**Description**
Text	Conversational Chinese 301
Speaker	7 females
Number of utterances	1899
Number of phones	26431
Average length per utterance	14
Number of annotators	6

B. Evaluation Metrics

There are 4 evaluation indicators:

- False Acceptance Rate (FAR): the percentage of mispronunciation phones that are accepted as correct.
- False Rejection Rate (FRR): the percentage of correctly pronounced phones that are rejected as mispronunciation.
- Diagnostic Accuracy (DA): the percentage of correctly detected.
- The Detection Cost Function (DCF):

$$DCF(\tau) = C_{MISS}FRR(\tau)P_{Target} + C_{FA}FAR(\tau)(1 - P_{Target}) \quad (6)$$

where τ is the threshold of GOP, C_{MISS} is the cost of false rejection, C_{FA} is the cost of false acceptance. P_{target} is a prior probability and in practical application, FAR is more important than FRR, because if too many correct pronunciations are rejected as mispronunciation, it will give users a bad experience.

C. Experiment Setup

Tensorflow toolkit was used to design CorrNet feature extractor, the input feature is Fbank applying CMVN with 10 frames context, a total of 11 frames as an input feature. Common and output layer's nodes are 50, input layer and one hidden layer nodes are 500, 300. The BNF extractor trained with Kaldi, and its input feature same with CorrNet. BNF extractor has 6 layers, each layer has 625 nodes. The last hidden is used as bottleneck layer, and it has 27 nodes.

Kaldi toolkit was used to train Gaussian Mixture Modeling (GMM), Hidden Markov Modeling (HMM) and TDNN, TDNN have 6 hidden layers and each layer have 850 nodes. The alignments generated by GMM-HMM model. Input feature is Fbank, CorrNet, and BNF.

IV. RESULTS AND DISCUSSIONS

A. The DA of Native and Non-native Corpus

We use native and some non-native corpus as training set to test the performance on both native and non-native set. The X-axis of following pictures means different thresholds of GOP. The Y-axis means the accuracy rate of mispronunciation detection. From figure 3, the result shows adding non-native data to training set still have a certain mismatch. In order to further reduce the mismatch, we conducted a comparative experiment with different features.

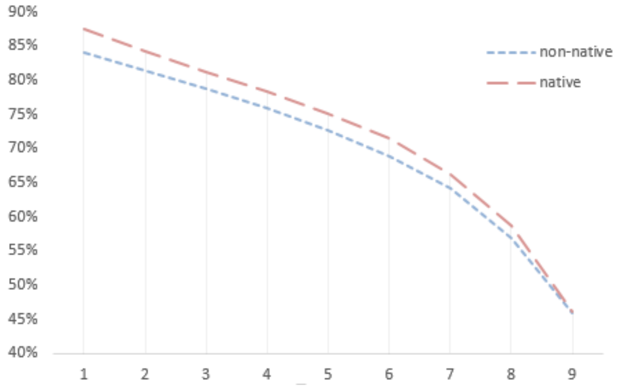

Figure 3. The DA of different dataset

B. Different Feature for TDNN Acoustic Model

We train four kinds of TDNN-HMM acoustic model with Fbank (baseline system), bottleneck (BNF), ivector and CorrNet feature, and use the way to calculate GOP as (4), the DA shown in Figure 4.

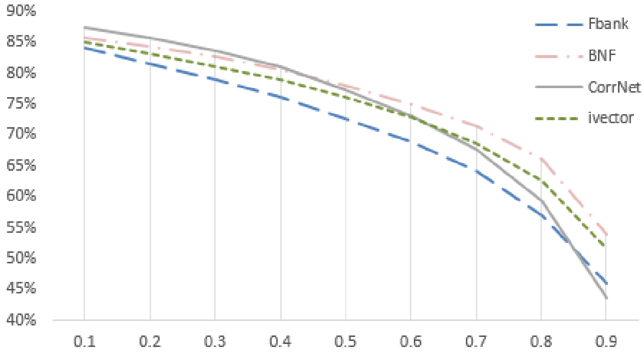

Figure 4. The DA of different feature

BNF is commonly used in language recognition and mispronunciation detection, this method does not consider the relations between two kind databases. In general, BNF and CorrNet features are better than baseline Fbank in all threshold. CorrNet is better than BNF for the threshold of less than 0.45. When the threshold is 0.1, CorrNet DA is 87.31%, BNF DA is 85.57% and Fbank DA is 84.12%. the ivector feature is mostly used in the speaker recognition task, it in mispronunciation detection task can be used to reduce speaker mismatch. When we want to evaluate the pronunciation of non-native speaker, we can use the most similar native speaker as the evaluation criteria. The DA of ivector is 85.04%. The result shows both CorrNet and BNF can reduce the mismatch of two kinds of corpus. The deep neural network as feature extractor can further reduce it than ivector. For the beginning learners of L2

Chinese, setting the threshold a little smaller is good to their language learning, so CorrNet method is more suit for them.

The ROC corves is shown in figure 5.

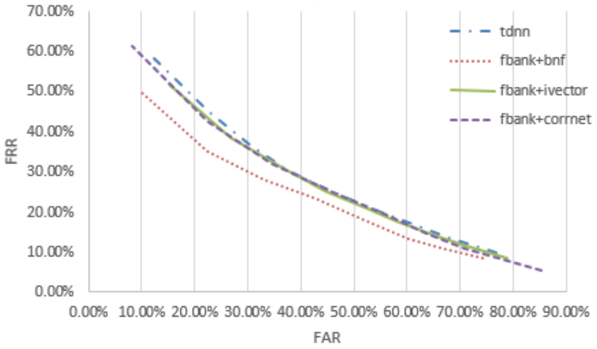

Figure 5. The ROC of different feature

The BNF feature extractor is trained with all native and about 2 hours non-native corpus. And due to the limitations of the model structure, the CorrNet only trained with 2 hours native and 2 hours non-native corpus. The performance of BNF is better than CorrNet.

C. DCF of Different Methods

The total score is 1, we set τ =0.5 for CorrNet and 0.6 for other methods. And P_{target} = 0.8. FRR is the percentage of mistake making by model. If too many correct pronunciations are judged by the system as errors will affect the user's belief in the system. Thus, In practice, we more care about FRR. Table 2 shows the DCF of different methods,

TABLE II. THE DCF OF DIFFERENT METHODS

Methods	DCF (%)
Fbank	31.83
BNF	27.14
CorrNet	26.71
Ivector	29.17

The DCF of BNF and CorrNet are close, however, CorrNet's DA is 77.3%, and it is higher than BNF 74.98%. The Ivector method can help to reduce the mismatch to some extent, and the DA of it is 72.94% in τ =0.6.

From the experiment results, both BNF and CorrNet methods can reduce the influence of mismatch. The thresholds mean the degree of rigors. The higher the threshold, the stricter it is and CorrNet adaption method is more suit for the beginner of language learners. Due to the data limitation, we only use about 2 hours non-native data to adapt native data, the CorrNet requires the same number of frames per view, so if more non-naïve data can be used in training set, the adaptation result could be better.

V. CONCLUSIONS

This paper proposed use different methods to reduce the database mismatch for mispronunciation detection task. Adding non-native data to training set is a common way, but the experiment result need have a further improvement, so we try to use CorrNet to map the L2 and native acoustic feature into a common subspace. Reduce the influence of channel, speaker and others and maximize the relationship between two kinds of corpora. Bottleneck feature was proposed to do the same thing, but in the training process, BNF's model all the data share the weights, it did not consider the difference and relations of input feature. Experiments result show CorrNet as feature extractor are outperform than TDNN extractor. We also compare the ivector feature, the experiment result shows neutral network methods are better than it. In this experiment, because of the limitation of model structure and experimental corpus, we only use 4 hours training corpus that includes native and non-native datasets. In the future, more unlabeled non-native corpora will be collected and participate in training CorrNet.

ACKNOWLEDGMENT

This work is supported by National social Science foundation of China (18BYY124), Advanced Innovation Center for Language Resource and Intelligence (KYR17005), the Fundamental Research Funds for the Central Universities (18YJ030004), the Graduate Innovation Fund of Beijing Language and Culture University (19YCX130), and the project of "Intelligent Speech technology International Exchange". Yanlu Xie is the corresponding author

REFERENCES

[1] R. Tong, B. P. Lim, N. F. Chen, B. Ma, and H. Li, "Subspace Gaussian Mixture Model for Computer-Assisted Language Learning", in ICASSP, pp. 5347–5351, 2014.

[2] H. Huang, H. Xu, Y. Hu, et al. "A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection", Journal of the Acoustical Society of America, 142(5):3165, 2017.

[3] Gales M J F. Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language, 1998, 12(2):75-98.

[4] C. Huang, F. Zhang, F K. Soong, et al. "Mispronunciation detection for Mandarin Chinese", IEEE International Conference on Acoustics. IEEE, 2008.

[5] G. Huang, J. Ye, Z. Sun, Y. Zhou, Y. Shen and R. Mo, "English mispronunciation detection based on improved GOP methods for Chinese students", 2017 International Conference on Progress in Informatics and Computing (PIC), Nanjing , pp. 425-429, 2017.

[6] D. Luo, C. Zhang, L. Xia and L. Wang, "Factorized Deep Neural Network Adaptation for Automatic Scoring of L2 Speech in English Speaking Tests", in INTERSPEECH, 2018.

[7] Y. Gao, Y. Xie, W. Cao, and J. Zhang, "A study on robust detection of pronunciation erroneous tendency based on deep neural network," in INTERSPEECH, pp. 693–696, 2015.

[8] Nicolao, Mauro, A. V. Beeston, and T. Hain. "Automatic assessment of English learner pronunciation using discriminative classifiers", IEEE International Conference on Acoustics, IEEE, 2015.

[9] W. Li, K. Li, S. M. Siniscalchi, N. F. Chen, and C. Lee, "Detecting mispronunciations of l2 learners and providing corrective feedback using knowledge-guided and data-driven based decision trees," in INTERSPEECH, 2016.

[10] W. Li, S. M. Siniscalchi, N. F. Chen, and C. Lee, "Improving Non-Native Mispronunciation Detection and Enriching

Diagnostic Feedback with DNN-Based Speech Attribute Modeling," in ICASSP, pp. 6135–6139, 2016.

[11] Kim, Yoon, H. Franco, and L. Neumeyer. "Automatic Pronunciation Scoring of Specific Phone Segments for Language Instruction", European Conference on Speech Communication & Technology DBLP, 1997.

[12] S. M. Witt, and S. J. Young. "Phone-level pronunciation scoring and assessment for interactive language learning", Speech Communication, vol. 30, no. 2, pp. 95-108, 2000.

[13] 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, vol. 67, pp. 154–166, 2015.

[14] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran, "Correlational Neural Networks." Neural Computation, vol. 28 pp. 257–285, 2016.

[15] S. Gao, et al. "Update Progress Of Sinohear: Advanced Mandarin LVCSR System At NLPR." In proc. ICSLP, 2000.

[16] W. Cao, et al. "Developing a Chinese L2 speech database of Japanese learners with narrow-phonetic labels for computer assisted pronunciation training", in INTERSPEECH, 2010.