DNN based detection of Pronunciation Erroneous Tendency in data sparse condition

Yingming Gao, Yanlu Xie, Ju Lin and Jinsong Zhang
College of Information Sciences, Beijing Language and Culture University, Beijing 100083, China
E-mail: gaoyingming1@sina.com, xieyanlu@blcu.edu.cn, linjucs@163.com, jinsong.zhang@blcu.edu.cn

Abstract— **Detecting pronunciation erroneous tendency (PET) can provide second languages learners with detailedly instructive feedbacks in the computer aided pronunciation training (CAPT) systems. Due to the data sparseness, DNN-HMM achieved limited improvement over GMM-HMM in our previous work. Instead of directly employing DNN-HMM to detect PETs, this paper investigated how to further improve the performance by DNN based features extracting in data sparse condition. Firstly, the probabilities of articulatory features derived from the top layer of DNN were fed into DNN-HMM. Secondly, the bottleneck features (BNF) extracted from the middle hidden layer were incorporated with original MFCC and then fed into SGMM-HMM. The experimental results showed that the new features converted from original acoustic features with DNN were more discriminative, and SGMM with BNF outperformed DNN in detecting PETs. The SGMM-HMM obtained the best detection results, achieving FRR of 5.3%, FAR of 29.6% and DA of 90%.**

I. INTRODUCTION

With accelerating process of globalization, there is an increasing need for learning a second language. As an important component of computer assisted language learning (CALL) systems, computer assisted pronunciation training (CAPT) has been attracting considerable attention in recent years [1-5]. As a crucial technology in CAPT, mispronunciation detection can detect learners' pronunciation errors, ideally provide corrective feedbacks, and is still a challenging research area. 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 [6]. Extended pronunciation lexicon or recognition network including standard pronunciations and common mispronunciations was employed to detect mispronunciation and provide diagnostic information [7-9]. This method can give corresponding phone substitution feedback when learners mispronounce phone /A/ as /B/.

Primary foreign language learners frequently suffer from phoneme substitution errors. However, the mispronunciations produced by intermediate and senior learners are neither target-like nor absolute phoneme categorical substitutions. Their erroneous pronunciation always deviate a little from canonical sound [10]. Compared with the canonical pronunciation, the mispronunciations are some acoustic variations but not different categories. In our previous work, pronunciation erroneous tendency (PET) was proposed to define a set of incorrect articulatory configuration regarding main articulation-placement and articulation-manner, such as shortening errors which can describes an insufficient aspiration [11]. Detailedly corrective feedback (e.g. “please try to round your lip more”) can be derived by comparing the difference between the detected PETs and their corresponding canonical articulatory configurations. The acoustic model in this framework was implemented by GMM-HMM [12], and subsequently replaced by DNN-HMM [13]. However, the DNN-HMM system achieved limited improvement over the GMM-HMM one due to the data sparseness problem. For further exploring the use of DNN for PETs detection in data sparse condition, this paper was motivated by the following two aspects.

1) Articulatory attributes refer to those abstract classes which reveal the positions and movements of different articulators during speech production. With speech attribute modeling, Li et al. improved mispronunciation detection and enriched diagnostic feedback [14]. On the one hand, articulatory attributes are consist with the definition of PET; on the other hand, they are more word, speaker and even language independent, which allows to share data to train acoustic model. Therefore, the articulatory features (AFs) extracted from articulatory attributes may be helpful to detect PETs.

2) Bottleneck (BN) features generated from multiple layer perceptron (MLP) [15], particularly deep neural networks [16-17], achieved significant improvement in ASR because bottleneck layer can learn some latent patterns of the input features. However, the effect of BN features is not obvious while dealing with small/medium-scale tasks in that limited training data is insufficient to train a complicated deep neural network. Therefore, Qi et al. investigated subspace Gaussian mixture model (SGMM) for BN features based ASR systems and obtained significant performance improvement [18].

Therefore, this paper investigated two approaches to further improve the performance of DNN based PET detection in data sparse condition. Firstly, the probabilities of articulatory features derived from the top layer of DNN were fed into DNN-HMM to detect PETs. Secondly, the bottleneck features were extracted from the middle hidden layer of DNN and then were fed into SGMM-HMM. The rest of the paper is organized as follows: Section II gives a brief description of PET detection framework. Section III and Section IV present articulatory features and bottleneck features for PET detection, respectively, which is followed by experiment and results in Section V. Finally, conclusions are given in Section VI.

II. PET DETECTION FRAMEWORK

The illustration of the detection system is provided in Fig. 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.

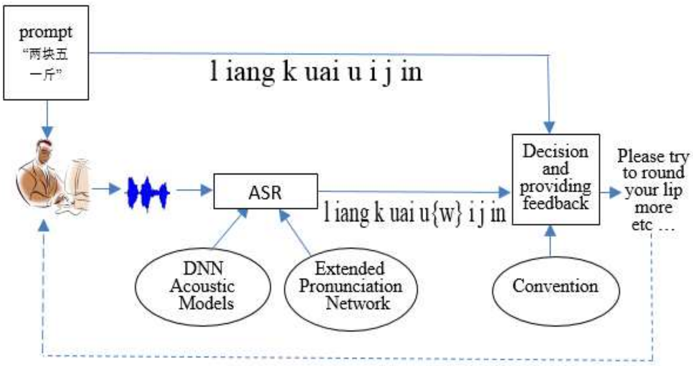
prompt
“两块五一斤”
l iang k uai u i j in
ASR
l iang k uai u{w} i j in
Decision and providing feedback
Please try to round your lip more etc ...
DNN Acoustic Models
Extended Pronunciation Network
Convention

Fig. 1 Flow chart of PET detection framework

III. USING ARTICULATORY FEATURES FOR PET DETECTION

A. *Articulatory categories*

Chinese syllables can be divided into two parts: Initials and Finals. Phonological studies suggest that both the Initials and the Finals can be further divided into a series of detailed categories based on articulatory movements such as manner of articulation, place of articulation, namely, articulatory features (AFs). For Initials, they are divided into 4 groups. As for Finals, there are 5 groups. The detailed information is shown in Table I.

TABLE I
CATEGORIES OF AFS

Feature groups		Feature values
Initials	Voicing	Voiced, unvoiced, null, sil
	Place	Bilabial, Labiodental, Alveolar, Dental, Retroflex, Palatal, Velar, null, sil
	Manner	Stop, Fricative, Affricative, Nasal, Lateral, null, sil
	Aspiration	Aspirated, Unaspirated, null, sil
Finals	Front-back of tongue	Front, Middle-Front, Middle, Middle-back, back, null, sil
	High-low of tongue	High, Middle-high, Middle, Middle-low, low, null, sil
	Rounding	+Round, -Round, null, sil
	Four hu	Kaikouhu, Hekouhu, Qicihu, Cuokouhu, null, sil
	location of dominant vowel	head-dominant, centre-dominant, tail-dominant, null, sil

B. *Extraction of articulatory features*

Since manual AF annotations of speech signals are rather difficult and costly to produce, one reasonable way of generating training material for the articulatory classifier is to convert phone-based training transcriptions to feature transcriptions [19]. This can be achieved by using a canonically defined phone-feature conversion table (such as Table I). In this paper, we used the posterior probabilities of the articulatory categories as the articulatory features. As shown in Fig. 2, to obtain the articulatory features, a bank of deep neutral network (DNN) classifiers were trained. We directly used the posterior probabilities of articulatory categories from the output of softmax layers. The input features of AF extractors were MFCC parameters, which consisted of the 6 preceding frames, the current frame and the 6 succeeding frames. “Append&Expand Module” block stacks together with the outputs delivered by the articulatory feature extractors and generates a supervector, which is fed into next module.

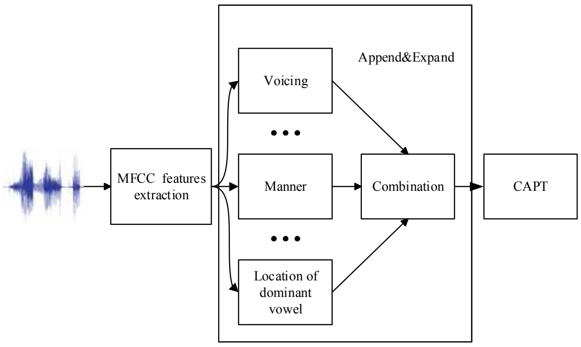

Fig. 2 Flow chart of AF extractors

IV. SGMM FOR BOTTLENECK FEATURE

A. *Bottleneck features*

Bottleneck feature is obtained from the bottleneck layer of an MLP structure. A particular property of the bottleneck MLP is that the bottleneck layer can learn some eminent patterns of the in the training phase. As shown in the left part of Fig. 3, there are 5 layers in total and 3 of them are hidden. The units of input layer (at the bottom) stand for a long-context feature vector. The feature vector is derived by three steps: 1) concatenating 9 consecutive frames of the primary feature followed by de-correlation and dimensionality reduction to 40 using linear discriminant analysis (LDA) [20]; 2) the obtained features were further de-correlated by using the maximum likelihood linear transform (MLLT) method [21]. It was followed by speaker normalization using feature-space maximum likelihood linear regression (fMLLR) [22].The fMLLR was estimated by using the GMM-based system applying speaker adaptive training (SAT) [22-23]; 3) the current frame features concatenated 5 preceding frames and 5 succeeding frames. Therefore, the number of units at input layer of the MLP is 440. The output layer (at the top) corresponds to the tied states of HMM, namely, sub-phones ("senone"). The three hidden layers are constructed following a 1024-42-1024 configuration, where the 42-unit layer (in the middle) is the "bottleneck layer", and the activations of the units yield the BNF.

The network is initialized with the deep belief network (DBN) pre-training procedure [24]. 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). After pre-training, all weights and bias were discriminatively trained by optimizing the cross entropy between the target (correspond to context-dependent HMM states) probability and actual output of softmax output with Back-Propagation (BP) algorithm [25].

B. *The framework of SGMM for BNF*

The traditional GMM-HMM framework for ASR assumes that the covariance matrices of the Gaussian components are diagonal. This assumption is obviously strong but it is necessary for small/medium-scale tasks which endure limited training data. BNF are highly sparse, i.e., most of the mass of feature concentrates on a few dimensions. This in turn leads to high correlation among the dimensions of the feature [18]. We firstly used LDA and MLLT de-correlated and reduced the feature dimensionality. Then we resorted to a more systematic approach, i.e. modeling the correlation using non-diagonal Gaussians. SGMM is a more feasible choice, which not only relaxes the diagonal covariance assumption and thus can model the correlation, and but also it assumes some shared structures, thus model the correlation in a parsimonious way. The SGMM is also effective in training acoustic model with limited amounts of training data.

The framework of SGMM for BNF is shown in the right part of Fig. 3, where a conventional GMM-HMM system is first constructed, and then a universal background model (UBM) is generated by clustering the Gaussian components. Later, the SGMM is initialized by copying the UBM, and then trained by an E-M algorithm similar to the GMM [26]. The input features of GMM-HMM consisted of the primary features and the bottleneck features.

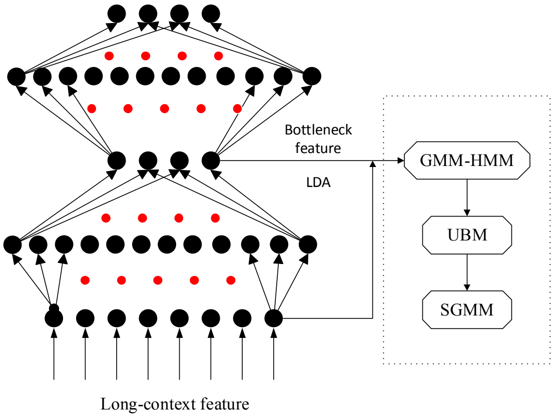

Fig. 3 Framework of SGMM model with bottleneck features

V. EXPERIMENTAL AND RESULTS

A. *Corpus*

In order to keep consistent with our previous work, this study was conducted on the continuous speech of Japanese part of BLCU inter-Chinese speech corpus [11]. Table II gives some overall statistics of corpus. 80% of the data was used for training and the rest for testing. For extracting the AFs, 12 (6 males and 6 females) native speakers' data of the BLCU inter-Chinese speech corpus were employed for training DNN based extractors with a portion (10%) of the training data withheld as a cross validation set.

TABLE II
THE DETAIL INFORMATION OF CORPUS

Text	301 utterance
Speakers	7 females
Number of utterance	1899
Number of phonemes	26431
Total duration	1.57 h
Number of types of PETs	65

B. *Experimental setup*

As for AF, firstly, the boundary information of Initials and Finals of the training data was generated from forced alignment with a recognizer based on DNN. Then, the training targets were obtained by converting phone-based training transcriptions to articulatory transcriptions. For the articulatory DNN classifiers training, we randomized the order of the training utterances lest the DNN training fall into a local optimum. We tuned the number of hidden nodes and hidden layers for the best frame classification accuracy for each articulatory features extractor. Frame accuracy is defined as the ratio of the number of correctly classified frames over the total number of frames, where classification is considered to be correct if the highest output of the DNN corresponds to the correct target. This is a fine preliminary indicator of system performance as well as an efficient way to tune the parameters without running the whole system. The final frame level accuracy of each group on the cross-validation data is shown in Table III.

The network was trained for 100 epochs using stochastic gradient descent (SGD) with a mini-batch size of 128, 20% dropout [27] in the input layer, 40% dropout in the hidden layers, and a cross-entropy objective.

After obtained the AFs, we fed them into DNN-HMM model. The DNN–HMM model was initialized with the deep belief network (DBN) pre-training procedure [24] and fine-tuned with Back-Propagation (BP) algorithm [25].

As for SGMM, we trained three kinds of SGMM-HMM based models with only MFCC, only BN features converted from MFCC, and both of them. Moreover, the models employing Perceptual Linear Predictive Analysis (PLP) or filter-bank (FBANK) were explored. For the first stage of training GMM-HMM, original acoustic features (13-dimemsion MFCC, 13-dimension PLP, and 23-dimension FBANK), with their first and second order derivatives respectively, were extracted from utterances with a 20ms length window shifted every 10 ms.

TABLE Ⅲ

FRAME ACCURACY OF THE ARTICULATORY DNN CLASSIFIERS ON THE CROSS VALIDATION

Articulatory Group	Accuracy (%)	Articulatory Group	Accuracy (%)
Voicing	98.29	High-low of tongue	96.25
Place	96.84	Rounding	98.03
Manner	96.84	Four hu	96.98
Aspiration	96.12	location of dominant vowel	97.21
Front-back of tongue	95.18		

C. *Evaluation metric*

Three kinds of metrics are used to inspect the evaluation performance:

- 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;
- Detection Accuracy (DA): The percentage of detected phones that are correctly recognized, i.e. the detection result is consist with the human annotations.

D. *Results*

We compared the system using MFCC and the one using AF for the PET detection. As shown in Table Ⅳ, though a slight degradation existing in FRR, the DNN-HMM-AF model obtained better performance in both FAR and DA over DNN-HMM+MFCC model.

TABLE Ⅳ

THE RESULTS OF AF-BASED SYSTEM AND MFCC-BASED SYSTEM

Acoustic model	FRR	FAR	DA
DNN-HMM+MFCC [13]	6.7%	35.9%	87.6%
DNN-HMM+AF	7.3%	31.1%	88.1%

We developed a series of SGMM-HMM based systems to demonstrate the effects of SGMM using BNF in detecting PETs. The statistic results are shown in Table V. As we can see from Table V, the SGMM-HMM+bnf (converted from MFCC) system outperformed the SGMM-HMM-MFCC one, which still attained comparable results to the DNN-HMM+MFCC system. Further improvement was obtained in the system incorporating BN feature and its corresponding original acoustic feature MFCC, which illustrated the advantages of SGMM model and the BN features generated by DNN.

Moreover, this research compared three SGMM-HMM+bnf based systems differing in acoustic features. FBANK outperformed MFCC and PLP in the SGMM-HMM+bnf system, which is consist with the DNN-HMM system in [13]. A lattice combination of the results of three feature systems led to the best PET detection performance: FRR of 5.3%, FAR of 29.6% and DA of 90%.

TABLE Ⅴ

THE RESULTS OF SGMM-BASED SYSTEM AND DNN-HMM-BASED SYSTEM

Acoustic Model	FRR	FAR	DA
DNN-HMM+MFCC [13]	6.7%	35.9%	87.6%
SGMM-HMM+MFCC	6.1%	39%	87.4%
SGMM-HMM+bnf	6.7%	31.8%	88.4%
SGMM-HMM+MFCC+bnf	5.7%	33.1%	88.9%
SGMM-HMM+PLP+bnf	6.1%	31.1%	89%
SGMM-HMM+FBANK+bnf	5.5%	29.6%	89.8%
SGMM-HMM System combination	**5.3%**	**29.6%**	**90%**

VI. CONCLUSIONS

In this paper, we investigate how to improve the performance of DNN based PET detection in data sparse condition. Instead of constructing hybrid DNN-HMM model, DNN was mainly used to extract more discriminative features. On the one hand, with the target of articulatory features, the posteriors probabilities were extracted from the top layer of DNN and then fed into DNN-HMM model; on the other hand, bottleneck features were derived from the middle “bottleneck layer” of DNN, incorporated with original acoustic features and then fed into SGMM-HMM model. The results showed that the system based on AF, which was converted from MFCC with DNN, outperformed the original one in detecting PETs and the SGMM-HMM model with BNF and MFCC was more discerning than DNN-HMM model. With lattice combination technology, best detection result was obtained by combining three SGMM-HMM based systems, achieving FRR of 5.3%, FAR of 29.6% and DA of 90%. We plan to apply our approaches and analyses to larger speech corpora like iCALL [28] for future work

ACKNOWLEDGMENT

This work is supported by Wutong Innovation Platform of Beijing Language and Culture University (16PT05) and BLCU support project for young researchers program (16YCX163) (supported by “the Fundamental Research Funds for the Central Universities”). The last author is corresponding author.

REFERENCES

[1] S. Wei, G. Hu, Y. Hu, R.H. Wang, “A new method for mispronunciation detection using support vector machine based on pronunciation space models,” Speech Communications, vol. 51, no. 10, pp. 896–905, 2009.

[2] H. Franco, L. Neumeyer, Y. Kim, O. Ronen, H. Bratt, 1999. Automatic detection of phone-level mispronunciation for language learning. In: Proc. European Conference on Speech Communication and Technology, pp. 851–854.

[3] S. M. Witt, & S. J. Young. Phone-level pronunciation scoring and assessment for interactive language learning. Speech communication, 30(2), 95-108, 2000.

[4] N. F. Chen, R. Tong, D. Wee, P. Lee, B. Ma, and H. Li,“iCALL Corpus: Mandarin Chinese Spoken by Non-Native Speakers of European Descent,” *INTERSPEECH*, 2015.

[5] 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.

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

[7] A. M. Harrison, W. K. Lo, X. Qian, & Meng, H. (2009, September). Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In SLaTE (pp. 45-48).

[8] H. Meng, Y. Y. Lo, L. Wang, & W. Y Lau. Deriving salient learners’ mispronunciations from cross-language phonological comparisons. In Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on (pp. 437-442). IEEE.

[9] W. K. Lo, S. Zhang, & H. M. Meng. Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system. *INTERSPEECH, 2010.*

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

[11] 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.

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

[13] Y. Gao, Y. Xie, J. Zhang and W. Cao, "A Study on Robust Detection of Pronunciation Erroneous Tendency Based on Deep Neural Network," *INTERSPEECH*, 2015.

[14] W. Li, S. M. Siniscalchi, N. F. Chen, & Lee, C. H. “Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling”. *ICASSP*, 2016.

[15] F. Grézl, M. Karafiát, S. Kontár, & J. Cernocky, Probabilistic and bottle-neck features for LVCSR of meetings. *ICASSP*, 2007.

[16] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.

[17] D. Yu, & M. L. Seltzer. Improved Bottleneck Features Using Pretrained Deep Neural Networks. *INTERSPEECH, 2011.*

[18] Qi, Jun, D. Wang, and J. Tejedor. "Subspace Models for Bottleneck Features." *INTERSPEECH*, 2013.

[19] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," Speech Communication, vol. 37, no. 3, pp. 303-319, 2002.

[20] R. O. Duda, P. E. Hart, and David G. Stork, "Pattern classification," in Wiley, 2000.

[21] R. Gopinath, “Maximum likelihood modeling with Gaussian distributions for classification,” *ICASSP*, 1998.

[22] M. J. F. Gales, “Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition,” Comp. Speech & Language, vol. 12, no. 2, pp. 75–98, 1998.

[23] S. Matsoukas, R. Schwartz, H. Jin, and L. Nguyen, “Practical implementations of speaker-adaptive training,” in DARPA Speech Recognition Workshop, 1997.

[24] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural Computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[25] D. E. Rumelhart, Geoffrey E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors." Cognitive modeling, vol. 5, no.3, pp. 1, 1988.

[26] D. Povey. “A tutorial introduction to subspace Gaussian mixture models for speech recognition” MSR-TR-2009-11, Microsoft Research, Tech. Rep., 2009.

[27] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co- adaptation of feature detectors," arXiv preprint arXiv: 1207. 0580, 2012.

[28] N. F. Chen et al., “Large-Scale Characterization of Non-Native Mandarin Chinese Spoken by Speakers of European Origin: An Analysis on iCALL”, Speech Communication, vol. 84, pp. 46-56, 2016.