Improving Deep Neural Networks Based Multi-Accent Mandarin Speech Recognition Using I-Vectors and Accent-Specific Top layer

Mingming Chen[1], *Zhanlei Yang*[1], *Jizhong Liang*[2], *Yanpeng Li*[2], *Wenju Liu*[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[2]Electric Power Research Institute of ShanXi Electric Power Company, China State Grid Corp

{mmchen,zhanlei.yang,lwj}@nlpr.ia.ac.cn, {eprjzl,eprypl}@126.com

10.21437/Interspeech.2015-718

Abstract

In this paper, we propose a method that use i-vectors and model adaptation techniques to improve the performance of deep neural networks(DNNs) based multi-accent Mandarin speech recognition. I-vectors which are speaker-specific features have been proved to be effective when used in accent identification. They can be used in company with conventional spectral features as the input features of DNNs to improve the discrimination for different accents. Meanwhile, we adapt DNNs to different accents by using an accent-specific top layer and shared hidden layers. The accent-specific top layer is used to adapt to different accents while the share hidden layers which can be seen as feature extractors can extract discriminative high-level features between different accents. These two techniques are complementary and can be easily combined together. Our experiments on the 400-hours Intel Accented Mandarin Speech Recognition Corpus show that our proposed method can significantly improve the performance of DNNs-based accented Mandarin speech recognition.

Index Terms: Accented speech recognition, deep neural networks, model adaptation, i-vectors, KL-divergence regularization

1. Introduction

Accent is one of the key factors in worsening the performance of practical speech recognition system [1]. We can classify accents into two kinds: accents caused by foreign or non-native speakers and accents caused by native speakers who speak the dialects of the language. In this paper, we focus on the accented Mandarin speech recognition task in which the speakers are from different regions of China.

For accented speech recognition, lexicon adaptation methods[1, 2, 4, 5, 6] and model adaptation methods [3, 7] have been used to improve the performance in the GMM-HMM based system. It has been also reported that model adaptation methods is typically found to be more effective than the lexicon adaptation[1, 8]. Recently, deep neural networks(DNNs) have become dominant techniques for acoustic modelling in automatic speech recognition(ASR) [9, 10, 11]. DNNs as powerful feature extractors can mitigate the degradation of performance of recognizing accented-speech. Nevertheless, there are still large performance gap between the accented speech and the native speech in the DNN-HMM based ASR system[8]. Some studies[8, 18] have conducted on adapting DNNs to improve the performance of recognizing accented speech. In [8], they propose a multi-accent deep neural network acoustic model with an accent-specific top layer and shared bottom hidden layers. This method has been proved very effective for foreign accented speech recognition and is similar to the method that adapts top layers for different languages in the DNNs-based multilingual speech recognition[12].

Besides adaptation methods, much work conducted on the accented speech recognition task has adopted features beyond acoustic features to supply more discriminative information between different accents[1, 2]. In recent years, i-vectors features[13, 19] which are speaker-specific features have been used to adapt deep neural networks to the target speaker[14]. It has been reported that using i-vectors as additional features can not only improve the the performance of speaker-independent speech recognition system but also speaker-dependent system. Furthermore, i-vectors has been used to identify native accents[15] and proved to be useful features for accent identification[16]. Therefore, it is sensible to use i-vectors as complementary input features in accented speech recognition.

In this paper, we propose a method combining i-vectors and model adaptation to improve the performance of deep neural networks based native accented Mandarin speech recognition. We use i-vectors in parallel with acoustic features to supply accent-specific information for accented mandarin speech recognition. For model adaptation method, we use the method proposed in [8] which uses deep neural networks with bottom shared hidden layers and accent-specific top layer. Our experiments on the 400-hours Intel Accented Mandarin Speech Recognition Corpus show that our proposed method achieves a 11.8% relative improvement in word error rate over the DNN-based baseline system.

The reminder of this paper is organized as follows. In section 2, we introduce i-vectors used in our paper. We show the details of the model adaptation method in Section 3. Section 4 presents the framework of our proposed method. Experiments and results are presented in Section 5. We conclude the paper in Section 6.

2. I-vectors

I-vectors[7] are a popular technique for speaker recognition and speaker verification because they can encapsulate all the speaker's relevant information in a low-dimensional fixed-length representation[19]. In this paper, we use i-vectors to supply accent-specific information. I-vectors are used in parallel with acoustic features(such as MFCC, PLP features) as the input features to deep neural networks. The details of i-vectors technique is introduced below.

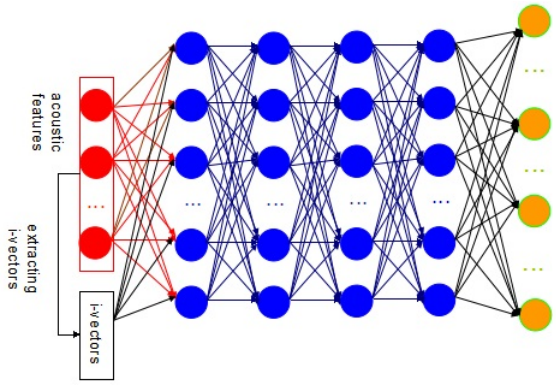

Figure 1: I-vectors extraction and intergration with a neural network. Note: I-vectors is extracted by utterance, spectral features in an uttrerance have same i-vectors features.

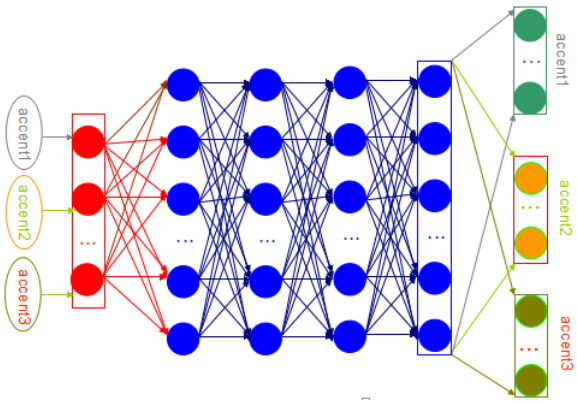

Figure 2: Framework of model adaptation method for DNN-based accented speech recognition. Note: The set of language senones is same for different kinds of accented speech.

I-vectors are low-dimensional features used in speaker recognition which can characterizes speaker-relevant information. In this paper, we extract the i-vectors for each utterance and stack it with acoustic features as the input of a DNN-based accented Mandarin ASR system.

I-vectors approach was originally motivated by the Joint Factor Analysis(JFA) framework introduced in [13]. For the speaker recognition tasks, factor analysis is used to generate a low-dimensional subspace which is called the total variability space. This space includes factors of both speaker and channel variability. Unlike JFA, all the variability is constrained in a single subspace in the I-vector representation whereas each kind of variability is modelled in an explicitly separate subspace in JFA. They utilize an effective and elegant way of decreasing the large-dimensional input data to a small-dimensional feature vector. The resulting feature vector retains most of the useful information of speaker entity.

In the framework of i-vectors approach, given a universal background model(UBM) trained on data from multiple speakers, we can adapt it to a given utterance and generate an utterance-dependent Gaussian mixture model(GMM). The eigenvoice adaptation technique used in the i-vectors extraction assumes that there is a matrix T contains speaker and channel variability information. The utterance GMM supervector M is obtained as

$$M = m + T * i; \tag{1}$$

where m is the segment-independent component of the mean supervector taken from a GMM-UBM trained on a large number of speakers; T is a low-rank rectangular matrix spanning the subspace covering the relevant variability; i is a low-dimensional latent variable representing coordinates in the subspace and has normally distributed prior N(0,I). After iteratively estimating matrix T on a large training corpus, we can use the lower-dimensional vectors i as a speaker model to replace a large GMM. i is referred to as an i-vector. More details of the i-vector algorithm are fully described in [13].

The procedure of using i-vectors as complementary input features in a neural network is shown in Figure 1. In the procedure, we first use i-vectors extractor to extract each utterances' i-vectors features , then they are concatenated to each frame in the utterance to form the input features for neural network.

3. Model Adapation Method

In this paper, we use the model adaptation method which adapt the top layer of conventional DNNs to accent-specific top layer[8]. This method is motivated by the shared hidden layers of deep neural networks in the multilingual speech recognition[12]. In multilingual speech recognition, shared-hidden-layers multilingual deep neural networks(SHL-MDNN) can extract high-level cross-lingual features through hierarchical shared hidden layers while using language-dependent top softmax layers to recognize different languages. Compared to multilingual speech recognition, multi-accented speech recognition can use the similar method to improve the performance of accented speech, furthermore, we can use the same set of language senones for different accented speech, we just need to adapt the same set of language senones to different accented speech. The framework of the model adaptation method for accented speech is shown in Figure 2.

In this method, the bottom hidden layers are shared between different accents while extract high-level cross-accent features between different accented speech through multiple hidden layers. This process can allow maximal knowledge sharing between different accented speech and can be seen as a type of regularization[8]. When using this architecture in decoding phase, the computation of hidden layers can be shared between different accented speech and we only need to evaluate the accent-specific top layer separately for each accent.

4. Framework of the proposed method

In this section, we introduce the framework of our proposed method. The method can be seen as the combination of two parts: (1) using i-vectors as the complementary features for the input of DNNs and (2) using accent-specific top layer for the output layer of DNNs. For the i-vectors part, we first extract i-vectors for each utterance and then concatenate the i-vectors to the acoustic features for each frame, then we use the concatenated features as the input features of DNNs. For model adaptation part, we adapt the top layer to different accented speech to generate accent-specific top layers. The framework of the proposed method is depicted in Figure 3.

Our proposed method has two advantages: 1) compared to the previous methods[5, 8], it combines the accent-specific information from the input layer and the top layer of deep neu-

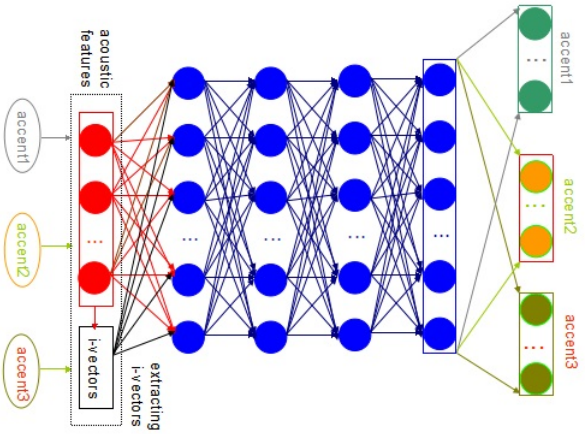

Figure 3: Framework of our proposed method.

ral networks which can improve the discrimination of DNNs-based acoustic model between different accented speech; 2) the combination of the two techniques is very straight and easy, we can first train DNNs using acoustic features and i-vectors as input features, then just adapt the parameters in the top layer to accent-specific layer leaving the input layer and hidden layers unchanged.

5. Experiments and Results

We used the open source Kaldi speech recognition toolkit[17] in all experiments and conducted our experiments on the 400+ hours Intel Accented Mandarin Speech Recognition Corpus. The corpus consists of six different kinds of accented Mandarin speech. The speakers of different accents are native residents from Beijing, Shanghai, Guangdong, Haerbin, Wuhan and Chengdu respectively. Accented Mandarin speech spoken by speakers from Beijing and Haerbin is close to Standard Mandarin(Putonghua) speech while accented Mandarin speech spoken by residents living in Shanghai, Guangzhou, Chengdu and Wuhan is strongly affected by Wu dialects, Cantonese, Chengdu dialects and Wuhan dialects which were much different from Standard Mandarin.

The statistics of the train data in the Corpus we used is listed in Table 1.

Table 1: *Statistics of training data in Intel Accented Mandarin Speech Recognition Corpus used in our experiments.*

Accent	# speakers	# utterances	# hours
Beijing	225	88176	88.2
Haerbin	115	44565	44.3
Shanghai	219	84911	85.1
Chengdu	120	45881	45.9
Wuhan	116	45917	46.0
Guangzhou	243	93965	94.0
Total	1038	403415	403.5

In our experiments, we selected 100 utterances from two speakers as developing data and 1000 utterances from ten speakers as testing data in each accented speech. Speakers in train, developing and test set were mutually exclusive. Therefore, we had 600 utterances($\sim$ 1 hour) in the developing set and 6000($\sim$ 10 hours) utterances in the test set.

5.1. Frontend processing

We code the speech data into 25 milliseconds(ms) frames with a frame-shift of 10 ms. Each frame is represented by a feature vector of 39 dimensional MFCC features (static plus first and second order delta features) which are mean normalized per utterance. In our experiments, every 11 consecutive cepstral frames are spliced together to be the input features of DNNs.

5.2. I-vectors extraction

We use the Kaldi online i-vectors extraction script to extract i-vectors for each utterance [17]. The extraction process is as follows: 1) train a diagonal covariance UBM on the features which are extracted by apply LDA and MLLT transformation on the 13-dimensional static MFCC features. The number of mixtures in the UBM was set to 1024. 2) convert the resulting diagonal covariance UBM trained in phase 1) to full covariance UBM as initial extractor model. 3) use the same features in phase 1) to estimate the parameters in the initial extractor model and generate an i-vectors extrator when the parameters converged. As stated in [14], for an 300 hours speech recognition tasks, it is reasonable to use 100-dimensional i-vectors to get a good performance. Therefore, we use 100-dimensional i-vectors in our experiments. If there were more training data including a larger numbers of speakers, it may need to use higher-dimensional i-vectors.

5.3. DNN training

In our experiments, we trained four deep neural networks and they all had 6 hidden layers with p-norm activation functions[21]. The definition of p-norm activation function is :

$$y = ||x||_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}} \quad (2)$$

The value of p can be set to different values; we use p=2 in our experiments. In our experiments, for each hidden layers, the input dimension was set to 2500 and the output dimension was 250, that is to say, i in equation (2) is 10.

The four DNNs we trained were different in whether using i-vectors and accent-specific top layer in DNNs or not. We show the details of four models in Table 2.

Table 2: *Four deep neural networks trained in our experiments.*

Models	usage of i-vectors	usage of accent-specific top layer
DNN Baseline	No	No
DNN model I	Yes	No
DNN model II	No	Yes
DNN model III	Yes	Yes

All four kinds of DNN models use consecutive 11 frames of spectral features as their input. The softmax output layer has 4983 units which correspond to the context-dependent HMM states generated by the procedure of decision tree clustering in GMM-HMM framework.

The training recipes and methods outlined in [21] are used in our experiments. We first randomly split the training data into blocks each containing 1000000 frames and then train the networks with stochastic gradient descent(SGD) on mini-batches

of 256 frames and a cross-entropy criterion. The initial learning rate is set to 0.01 and the final rate is 0.001. In the mix-up procedure, we mix the 4963 output layer units up to 10000 units.

When using accent-specific top layer in DNN model II and III, we first trained a DNN model using the training data of all six kinds of accents before the mix-up procedure, then mixed up the top layer and adapted the top layer of the model to accent-specific top layer for each accent using training data of the corresponding kind of accented Mandarin speech. The adaptation procedure usually took 1 or 2 iterations before the parameters in the accent-specific top layer converged.

5.4. Hybrid DNN-HMM decoding

In the decoding phase, for DNN Baseline and Model I, we used the trained model to compute DNN output scores for all the utterances in the test set; for DNN Model II and III, we used DNN model with different accent-specific top layers to compute the output scores for the corresponding accented test speech data. The vocabulary used has 60K words and the decoding language model is a 3-gram LM with 80M n-grams.

5.5. Experimental results and analysis

After we trained these four DNN models described above, we used them to recognize the test set separately. We list the word error rates(WER) for these models on the test set in Table 3.

Table 3: *Word error rates for these four DNN models on the test set.*

Accent	Baseline	Model I	Model II	Model III
Beijing	12.44%	10.89%	11.90%	10.85%
Haerbin	15.02%	14.53%	14.70%	14.36%
Shanghai	18.01%	16.12%	17.07%	15.70%
Guangzhou	25.70%	23.51%	24.04%	22.69%
Chengdu	18.28%	16.12%	17.21%	15.74%
Wuhan	19.80%	17.98%	18.42%	17.13%
Average	18.21%	16.52%	17.22%	16.07%

For all the four DNN models, we use the pronunciation dictionary of Standard Mandarin. The pronunciation dictionaries of Wu Dialects, Cantonese, Chengdu Dialects and Wuhan Dialects are much different from that of Standard Mandarin, therefore accented Mandarin speech in these regions had much difference with Standard Mandarin speech. Dialects used in Beijing and Haerbin are similar with Standard Mandarin, then accented speech from these two regions sounds similarly to Standard Mandarin.

From Table 3, we can see that for Baseline model which only use MFCC features as input and no accent-specific top layer, test data from Beijing and Haerbin get much lower word error rates than these from other regions which imply that accented speech from these two regions are weaker than those from other regions.

For Model I, we use MFCC features concatenated with 100-dimensional i-vectors as the input features to deep neural networks which had no accent-specific top layer. From Table 3, we can conclude that using i-vectors as complementary features can significantly improve the performance of accented Mandarin speech recognition task which achieves 9.3% relative improvement in word error rate. Furthermore, we can see that except for the test data from Haerbin, significant improvements on other test data have been achieved. For the accented speech from Haerbin, we think relatively small amount of training data and weak accent is one of the reason for the relatively small improvement. The other reason we think is accented speech from Beijing and Haerbin had much similarity and the trained model was bias to Beijing accented speech according to the imbalance of the amount of data.

For Model II, we use MFCC features as the input features to DNNs which had accent-specific top layer. By using accent-specific top layer, we achieve about 5.4% relative improvement over the Baseline model. The improvement is moderate, however, we can see from the result that accented speech strongly affected by local dialects achieve higher improvements than others.

For Model III which is trained by method we proposed, we achieved about 11.8% relative improvements in word error rates than the Baseline Model. Compared to the results of Model I and II, we find that i-vectors features play a dominant roles in improving the the performance of accented Mandarin speech recognition. Furthermore, our proposed method combine i-vectors with accent-specific top layer to obtain the complementary information from the two techniques. The experimental results prove the effectiveness of our method.

6. Conclusions

We have presented an effective way to improve the performance of native accented Mandarin speech recognition system. In our proposed method, we use i-vectors as additional input features to DNNs which also have accent-specific top layer. I-vectors containing speaker information have been proved effective in accent identification task. Accent-specific top layer which can be seen as model adaptation method for DNNs to accented speech have been used to improve the performance of accented English speech recognition. These two techniques can be easily combined in the DNNs-based speech recognition system and can further improve the performance of the system. In this paper, we find that use these above two techniques separately can boost the performance of multi-accent Mandarin speech recognition; furthermore, the performance is further improved when we combine these two techniques. For future work, we plan to use the KL-regularized model adaptation method to improve the performance of the model adaptation part in our proposed method and try higher-dimensional i-vectors to test its performance.

7. Acknowledgements

This research was supported in part by the China National Nature Science Foundation (No.91120303, No.61273267, No.61403370 and No.90820011).

8. References

[1] Huang C, Chen T, and Chang E., "Accent issues in large vocabulary continuous speech recognition", International Journal of Speech Technology, 2004, 7(2-3): 141-153.

[2] Kat, L. W., and Fung, P., "Fast accent identification and accented speech recognition", In Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on (Vol. 1, pp. 221-224). IEEE.

[3] Wang, Z., Schultz, T., and Waibel, A.,"Comparison of Acoustic Model Adaptation Techniques on Non-native Speech", in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[4] Vergyri, D., Lamel, L.,and Gauvain, L., "Automatic Speech

Recognition of Multiple Accented English Data", in the Proceedings of the 2010 Interspeech conference, 2010.

[5] Zheng, Y.L., Sproat, R., Gu L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr R., and Yoon, S.,"Accent Detection and Speech Recognition for Shanghai-accented Mandarin", in the Proceedings of the 2005 Interspeech conference, 2005.

[6] Nallasamy, U., Metze, F., and Schultz, T., "Enhanced Polyphone Decision Tree Adaptation for accented-speech Recognition", in the Proceedings of the 2012 Interspeech conference, 2012.

[7] Arslan, L. M. and Hansen, J. L., "A Study of the Temporal Features and Frequency Characteristics in American English Foreign Accent", Journal of the Acoustic Society, America, December, 1996

[8] Huang, Y., Yu, D., Liu, C.J., and Gong, Y.F., "Multi-Accent Deep Neural Network Acoustic Model with Accent-Specific Top Layer Using the KLD-Regularized Model Adaptation", in the Proceedings of Interspeech 2014.

[9] Dahl, G.E., Yu, D., Deng, L., and Acero, A., "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing (TASLP) - Special Issue on Deep Learning for Spech and Language Processing, Volume: 1, No. 1, Page(s): 33-42, Jan 2012.

[10] Seide, F., Li, G., and Yu, D., "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks", in the Proceedings of Interspeech 2012.

[11] Yu, D., Seltzer, M., Li, J. ,Huang, J., Seide, F., "Feature Learning in Deep Neural Networks - Studies on Speech Recognition Tasks", in the Proceedings of 2013 International Confernece on Learning Representation, 2013.

[12] Huang, J., Li, J., Yu, D., Deng, L., and Gong, Y., "Cross-Language Knowledge Transfer Using Multilingual Deep Neural Network With Shared Hidden Layers", in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[13] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P., "Frontend factor analysis for speaker verification", IEEE Trans. Audio, Speech and Language Processing, vol. 19, no. 4, May 2011.

[14] Saon, G., Soltau, H., Nahamoo, D., Picheny, M., "Speaker Adaptation of Neural Network Acoustic Models Using I-Vectors", in the Proceedings of ASRU, 2013.

[15] DeMarco, A., Cox, S.J., "Native Accent Classification via I-Vectors and Speaker Compensation Fusion", in the Proceedings of Interspeech 2013.

[16] Bahari, M.H., Saeidi, R., Hamme, H.V., Leeuwen, D.V., "Accent Recognition Using I-vector, Gaussian Mean Supervector And Gaussian Posterior Probability Supervector For Spontaneous Telephone Speech", in the Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[17] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y.M., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K., "The Kaldi Speech Recognition Toolkit", in the Proceedings of ASRU, 2011.

[18] Chen, X., Cheng, J., "Deep Neural Network Acoustic Modeling for Native and Non-native Mandarin Speech Recognition", in the Proceedings of 2014 9th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2014.

[19] Glembek, O., Burget, L., Matejka, P., Karafiat, M., and Kenny, P., Simplification and optimization of i-vector extraction, in the Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011.

[20] Karafiat, M., Burget, L., Matejka, P., Glembek, O., and Cernozky, J., "iVector-based discriminative adaptation for automatic speech recognition", in the Proceedings of ASRU, 2011.

[21] Zhang, X.H., Trmal, J., Povey, D., Khudanpur, S., "Improving Deep Neural Network Acoustic Models Using Generalized Maxout Networks", in the Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014.