Improving native accent identification using deep neural networks

Mingming Chen, Zhanlei Yang, Hao Zheng, Wenju Liu

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

{mmchen,zhanlei.yang,hzheng,lwj}@nlpr.ia.ac.cn

10.21437/Interspeech.2014-486

Abstract

In this paper, we utilize deep neural networks(DNNs) to automatically identify native accents in English and Mandarin when no text, speaker or gender information is available for the speech data. Compared to the Gaussian mixture model(GMM) based conventional methods, the proposed method benefits from two main advantages: first, DNNs are discriminative models which can provide better discrimination on confusion regions of different accents; second, they have the hierarchical nonlinear feature extraction capability which can learn discriminative high-level features for the specified task. In detail, the speech data of all accents is used to train DNNs, and in the testing stage, we first identify the accent label of each frame, then determine the sentence label by the majority voting conducted on the frame labels. The experiments on accented English and Mandarin corpus demonstrate that, compared to the GMM based methods, our proposed method can significantly improve the frame accuracy as well as sentence accuracy on the test set. Moreover, the performance of the proposed method can be further improved by using context information.

Index Terms: automatic accent identification, Deep neural networks, Gaussian mixture model, English accents, Chinese accents

1. Introduction

Automatic accent identification is an active research topic in the speech community because accent is one of the key factors in worsening the performance of practical speech applications(e.g., speech recognition, speech understanding)[1, 2, 3, 4, 5, 6, 7]. There are two kinds of accents: 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 identifying accents caused by native speakers in audio which has no transcripts, no speaker or no gender information.

Methods used in accent identification can be classified in two categories: text-dependent methods[2, 3, 4] and text-independent methods[5, 6, 7]. There are many scenes that no text, gender or speaker information is available for the speech data we need to process, therefore, we focus on the method for these unconstrained speech data. Gaussian mixture model(GMM)-based methods are text-independent methods which have been widely used for accent identification[5, 6, 7]. The GMM is usually trained with maximum likelihood estimation(MLE) and the confusion regions between the resulting GMMs for different accents cannot be well suppressed due to the concentration of maximizing the probability of data in each accent[5]. Although Discriminative training methods, such as minimum classification estimation(MCE)[12, 13], have been utilized to strengthen the discrimination by depressing the confusion regions explicitly, their performance is still limited.

In this paper, we proposed an accent identification method based on deep neural networks(DNNs). We refer DNNs to feed-forward neural networks which have at least two hidden layers. DNNs which have been resurrect recently have become the-state-of-art methods in acoustic modelling of speech recognition[8] and image classification[9]. For the accent identification task, DNNs which are discriminative models can provide better discrimination on confusion regions of different accents than GMM. Furthermore, as stated in [14], multiple hidden layers in the DNNs can be seen as more powerful feature extractor than shallow models in speech recognition. It has also been reported that the context information can improve the performance of speech applications, such as speech recognition and prosodic event detection[8, 14, 15, 16]. We then investigate the use of context information to capture long time span features for accents and improve the performance of accent identification. Our experiments on IViE corpus[17] and Intel accented Mandarin corpus show that DNNs reduce the sentence error rate by 16.5% and 12.3% correspondingly in the three-way native accent identification tasks on the English and Mandarin corpus and further reduce them by 6.0% and 6.8% using context information.

The rest of the paper is organized as follows: Section 2 describes the two data corpus and features used in the experiments. Section 3 presents the DNN based accent identification approach. Section 4 describes the experimental results and analysis. Section 5 concludes the paper and describes future works.

2. Data corpus and features

2.1. Data Corpus: English and Mandarin

The first corpus used in our experiments is the IViE corpus[17], which consists of eight kinds of native British accents. We use speech data of three accents in our study based on geographical origins as in [5]. These accents are from Belfast(Northern Ireland), Cambridge(England) and Cardiff(Wales). For each accent, we use the semi-spontaneous retold version of read texts for our study. There are six male and six female speakers in each accent. We select three male and three female speakers for training; one male and one female speakers for validation and two male and two female speakers for testing. For each speaker, most of the retold sentences are roughly 20-50 seconds and we split these long sentences into several short sentences which each occupies 3-7 seconds long.

The second corpus used is the Intel accented Mandarin speech corpus, which contains six kinds of Mandarin accents. For our study, we select accents from Beijing, Shanghai and Guangzhou where local speakers in these places speak Standard Mandarin, Wu dialectal and Yue Dialectal respectively. For each accent, we selected 50 male and 50 female speakers

for training, two male and two female speakers for validation(or dev) and 10 male and 10 female speakers for testing. For each speaker, we randomly select 50 sentences from the corpus for our study. Each sentence is roughly 3-7 seconds long.

The statistics of data in the two corpus used in the experiments is listed in Table 1.

Table 1: *Statistics of data in the two corpus used in our experiments.*

Language	data set	speakers	sentences	hours
English	train	6	828	3.1
	dev	2	216	0.8
	test	4	605	2.2
Mandarin	train	300	15000	20.3
	dev	12	600	1.1
	test	60	3000	5.6

2.2. Features

Before extracting features from the speech audio in the two corpus, we first use a silence remover to eliminate the long silence in the utterances. The silence remover and feature extractor use a window of 25 milliseconds(ms) and a frame shift of 10 ms. The silence remover is based on the energy of each frame. We then extract 39-dimensional MFCC features(static plus first and second order delta features) of each frame as input features for accent identification. Cesptral mean and variance normalization is then performed on the utterance level. For the context information, we concatenate current frame with a window of previous and following frames of the current frame as the input features and the window size varies from 1 to 5.

3. Deep neural networks based accent identification

3.1. Deep neural networks

DNNs have received a lot attention again in recent years and become the dominant techniques in acoustic modelling in speech recognition[8] and image classification[9]. Compared to feed-forward neural networks used before, DNNs have more hidden layers and more hidden nodes in each hidden layer. In the past, due to the limitation of computation resources and the lack of careful initialization methods, feed-forward neural network with the same configurations could not be trained successfully. Thanks to the method proposed in [18] and with the help of general purpose GPU computing, DNNs have been successfully trained and achieved significantly improvement in many speech processing and image classification tasks.

Given an input observation vector $\mathbf{x}$, DNNs pass it through multiple hidden layers $\mathbf{h}_i$ to get the output vector $\mathbf{o}$ at the output layer, where $i = 1, ..., L$ and L is the number of hidden layers. This procedure can be formulated as follows:

$$\mathbf{h}_1 = \sigma(\mathbf{W}_1 * \mathbf{x} + \mathbf{b}_1) \quad (1)$$

$$\mathbf{h}_{i+1} = \sigma(\mathbf{W}_i * \mathbf{h}_i + \mathbf{b}_i), i = 1, ..., L-1, \quad (2)$$

where $\mathbf{W}_i$ and $\mathbf{b}_i$ are the weight matrix and bias vector for the hidden layer respectively. σ is the activation function which is usually a sigmoid function or rectified linear function. Compared to the sigmoid function, the rectified linear function can be computed faster and generalize better[10]. Therefore, we use the rectified linear function as the activation function in our experiments. The rectified linear function has following form:

$$\sigma(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{else} \end{cases} \quad (3)$$

For the output layer, the softmax function

$$p(y = s|\mathbf{h}_L) = \frac{\exp\left(\mathbf{W}_L^s * \mathbf{h}_L + \mathbf{b}_L^s\right)}{\sum_{y'} \exp\left(\mathbf{W}_L^{y'} * \mathbf{h}_L + \mathbf{b}_L^{y'}\right)}, \quad (4)$$

is used to estimate the label posterior probability $p(y = s|\mathbf{x})$ where s represents the accent label, $\mathbf{W}_L^{y'}$ and $\mathbf{b}_L^{y'}$ are the y'th row of the weight matrix $\mathbf{W}_L$ and the y'th element of bias vector $\mathbf{b}_L$.

3.2. Dropout training

Overfitting is one of the biggest problems in training DNNs. It typically happens when a relatively small training set is used to train a large DNN. Dropout training has been recently proposed to mitigate the the problem[11]. In the dropout training, a certain percentage(e.g., α) of the neurons in each hidden layer is randomly omitted for each batch of samples. This reduces the dependency of each neuron on other neurons. Dropout training essentially can reduce the capacity of the DNNs and thus improve the generalization ability of the resulting model. Futhermore, when a hidden unit is dropped out, its activation is set to 0 and then no error signal will pass through it[11]. This indicates that it doesn't need to change the training algorithm except the random dropout operation.

When testing, we use the "mean network" which contains all of the hidden units but with their outgoing weights multiplied by $1 - \alpha$ to compensate for the fact that $\frac{1}{1-\alpha}$ are active. This means dropout training can also be seen as an efficient way of model averaging of the DNNs.

3.3. DNNs for accent identification

For the accent identification task, the input of DNNs are the features which are the same as those used in the GMM-based methods. In the training procedure, the training speech data of all accents and their corresponding accent labels are used to train a universal DNN whose output layer has N output nodes where N is the number of kinds of accents. In the testing procedure, the input audio is first processed by the silence remover and feature extractor and then passed to the resulting DNN to compute the posterior probability of each accent per frame. After calculating the posterior probabilities for a frame, the label of the frame is determined as the accent label which has the maximum posterior probability. After we get the labels of all frames in an utterance, the label of the utterance is further voted from the labels of all frames. Figure 1. shows the flow diagram of training and testing process of accent identification based on DNNs.

4. Experiments and result analysis

In this section, we present first the results of GMM-based native accent identification system and then results of the proposed DNNs-based system. After that, we investigate the influence of the context information for the DNNs-based systems.

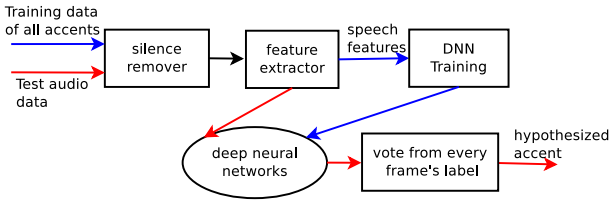

Figure 1: *The training and testing process of DNN based accent identification. The blue arrow represents for the data flow of training process and the red one for the testing. The black arrow represents procedure used for both training and testing phase.*

4.1. GMM-based accent identification

In this part, we use GMM to identify native accents. Before pushing the MFCC features to the GMM, we first use silence remover and feature extractor to process the training speech data. For each accent, we train GMMs with diagonal covariance matrices to model the training data of that accent using the MLE criterion. In order to determine the optimal parameters of GMM, we vary the number of Gaussian mixtures from 256 to 1024. The results of GMM-based accent identification method are presented in Table 2. From the table, we can see that we get

Table 2: *Frame and sentence(sent.) accuracy of English and Mandarin corpus with GMM using different numbers of mixtures.*

Language	Acc(%)	Number of Gaussian mixtures			
		256	512	768	1024
English	frame	52.3	**52.8**	52.4	52.1
	sent.	90.1	**91.5**	90.9	89.8
Mandarin	frame	39.5	40.2	**40.6**	40.4
	sent.	64.0	65.2	**67.6**	67.0

good results on the IViE corpus which is similar as the results in [5] using GMM with MLE. For the Mandarin task, considering the size of the dataset and the similarity between native accented Mandarin, we achieve reasonable frame and sentence accuracy.

Moreover, as stated in [5], discriminative training methods for GMM can improve the classification performance. We use MCE training for GMM to increase the discrimination of GMMs. As we get the best performance on English and Mandarin corpus when the numbers of Gaussian mixtures are set to 512 and 768 separately, we use the same parameters for MCE training on the two corpus. The results of GMM with MCE training are presented in Table 3. The results in table 3 indicates

Table 3: *The performance of GMM-based methods using MCE training on the two corpus.*

Language	Training method	Frame acc(%)	Sent. acc(%)
English	MLE	52.8	91.5
	MCE	**53.4**	**92.1**
Mandarin	MLE	40.6	67.6
	MCE	**41.1**	**68.2**

that MCE training can improve the performance of GMM-based method as it increases the discrimination ability of GMM between different classes. In the following section, we show that better results can be achieved using DNNs which can provide better discrimination between different accents.

4.2. DNN-based accent identification

For DNN-based methods, we examine different configurations of the DNNs. The number of hidden layers varies from 2 to 3 and the number of nodes in each hidden layer increases from 256 to 2048. The input layer has 39 visible units and the output layer has 3 nodes. The networks are initialized by the method proposed in [19]. After initialization, DNNs are trained using mini-batch stochastic gradient descent with the batch size being 128. The initial learning rate is set to 0.01. Momentum is used to speed up learning. The momentum starts off at 0.5 and increases linearly to 0.9 over the first 10000 epochs. The stop condition for the training procedure is determined by performance on the dev sets.

Table 4: *Frame and sentence(sent.) accuracy on the test set of English and Mandarin corpus with DNNs using different configurations.*

Language	No.layer	Acc(%)	Number of hidden layer nodes			
			256	512	768	1024
English	2	frame	56.1	58.6	59.1	58.4
		sent.	86.3	91.7	91.3	90.9
	3	frame	57.8	**60.1**	59.9	59.6
		sent.	91.5	**93.0**	92.6	92.1
Mandarin	2	frame	43.1	44.4	45.0	44.8
		sent.	64.8	67.5	68.6	68.9
	3	frame	44.8	45.7	**46.2**	45.4
		sent.	67.6	69.1	**70.2**	67.9

The results of the DNNs with various hidden layers and hidden nodes on the two corpus are shown in Table 4. As is indicated in the table 4, all of the DNNs produce substantial improvement of frame accuracy on the test set compared to the baseline GMM-based system and for most of them the sentence accuracy is improved. Furthermore, we can see that the DNN-based accent identification system achieves best performance on the two corpus with different architectures due to the amount of the training data. This is similar to the GMM-based system. Finally, with the consideration of training efficiency, we choose DNN with 3 hidden layers which has 512 nodes in each hidden layer for the IViE corpus and DNN which has 3 hidden layers with each having 1024 units in each hidden layer for the Intel corpus.

Then we use dropout training for the two selected DNN architectures to improve the performances. For dropout training, we set the dropout rate(α) from 0.1 to 0.5 and get the best performance for the two corpus when it is set to 0.1. We present the results of DNN using dropout training in the table 5.

Table 5: *Frame and sentence(sent.) accuracy on the test set of English and Mandarin corpus with DNNs using dropout training.*

Language	Model	Frame acc(%)	Sent. acc(%)
English	DNN	60.1	93.0
	DNN+dropout	**60.8**	**93.4**
Mandarin	DNN	46.2	70.2
	DNN+dropout	**47.4**	**72.1**

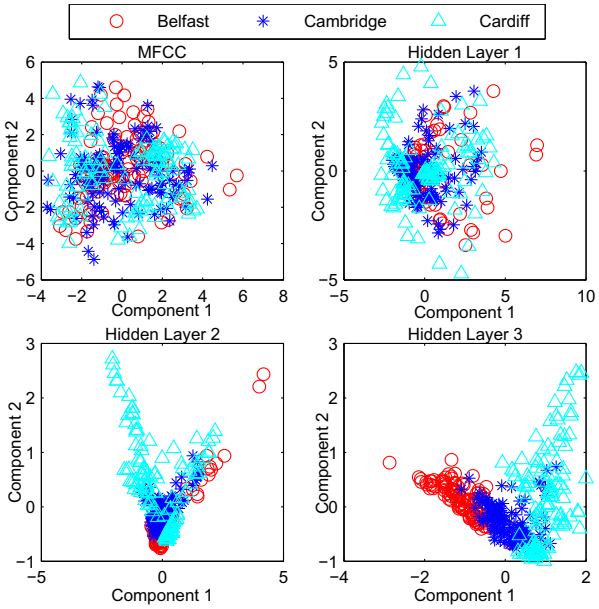

Figure 2: *The plots of component 1 and 2 of PCA results on MFCC raw features, hidden layer 1, 2,and 3 derived features. The samples are randomly selected from IViE corpus. Each accent has 128 samples and each hidden layer has 512 nodes.*

From table 5, we can see that dropout training can further improve the frame accuracy and sentence accuracy. Compared to the best results of GMM-based method, DNN with dropout training reduce the sentence error rate by 16.5% and 12.3% relatively. The results presented in table 5 confirm that DNNs which are discriminative models can provide better discrimination between confusion regions of different accents.

To illustrate that DNNs can learn discriminative high-level features, we randomly select 128 samples for each accent from IViE corpus and pass them through the best resulting DNN model described above. Then we use PCA to select two resulting principle components from the MFCC raw features, hidden layer 1, 2 and 3 derived features. Finally, we visualize the two principle components for the features of the samples in Figure 2. From Figure 2, we can see that as the raw features pass through the multiple hidden layers, hidden layers derive more discriminative features than the raw features. Therefore, DNNs-based method can learn discriminative features through non-linear hidden layers and produce better performance than its GMM-based counterparts.

4.3. The influence of context information

As stated above, the context information is useful for speech applications such as speech recognition and prosodic event detection[8, 14, 15, 16]. In this part, we investigate the use of the context information for DNNs-based accent identification in English and Mandarin. We use the same DNN configurations as that used in the systems producing the best results without using contextual information. They are also trained using dropout training to get better results. The context information we use are stated in Section 2.2. Table 6 presents the results of different window sizes of frames used in DNNs with dropout training.

The results in the table 6 indicate that context information is helpful in identifying native accents for English and Mandarin.

Table 6: *Frame and sentence(sent.) accuracy on the test set of English and Mandarin corpus with DNNs using different context information.*

Language	Acc(%)	Window size				
		1	2	3	4	5
English	frame	64.8	67.9	69.6	70.1	70.6
	sent.	93.6	93.8	93.7	93.1	92.9
Mandarin	frame	48.2	49.1	49.7	50.0	50.4
	sent.	72.9	73.5	74.0	73.8	73.6

They can significantly improve the frame accuracy which can benefit the sentence accuracy on the test set. However, if the window size is too large, it can harm the performance of DNNs which may be explained by that the context information weights too much than the current frame and misleads the models.

5. Conclusions

In this paper, we propose a DNNs-based method for native accents identification in English and Mandarin. This work shows that DNNs-based method can provide better discrimination between confusion regions of different accents than GMM-based method and learn discriminative high-level features through multiple non-linear hidden layers. The performance of DNNs can be further improved by using context information. In the future, we plan to investigate methods for determining sentence label from the labels of each frame and fuse different features using DNNs to further improve the performance of the proposed method.

6. Acknowledgements

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

7. 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] Pongtep, A., and Hansen,J.H., "Advances in phone-based modeling for automatic accent classification." Audio, Speech, and Language Processing, IEEE Transactions on 14.2 (2006): 634-646.

[4] Chen, N. F., Shen, W., and Campbell, J. P. (2010, March), "A linguistically-informative approach to dialect recognition using dialect-discriminating context-dependent phonetic models", In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on (pp. 5014-5017). IEEE.

[5] Huang, R.Q., and Hansen, J. H., "Unsupervised discriminative training with application to dialect classification.", Audio, Speech, and Language Processing, IEEE Transactions on 15.8 (2007): 2444-2453.

[6] Ghinwa, C., Zweig, G., and Nguyen, P., "An empirical study of automatic accent classification." Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE, 2008.

[7] Lei, Y., and Hansen, J. H. (2011)., "Dialect classification via text-independent training and testing for Arabic, Spanish, and Chi-

nese". Audio, Speech, and Language Processing, IEEE Transactions on, 19(1), 85-96.

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

[9] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In NIPS (Vol. 1, No. 2, p. 4).

[10] Nair, V., and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10) (pp. 807-814).

[11] Hinton, G.E., Srivastava, N., Krizhevsky, A, Sutskever. I., and Salakhutdinov, R., "Improving neural networks by preventing co-adaptation of feature detectors", The Computing Research Repository(CoPR), vol.abs/1207.0580, 2012.

[12] Juang, B.W., Wu, H., and Lee, C.H., "Minimum classification error rate methods for speech recognition.", Speech and Audio Processing, IEEE Transactions on 5.3 (1997): 257-265.

[13] Juang, B.W., and Katagiri, S., "Discriminative learning for minimum error classification [pattern recognition]." Signal Processing, IEEE Transactions on 40.12 (1992): 3043-3054.

[14] Yan, Z., Huo, Q., and Xu, J. (2013). "A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR.", In Proceedings of Interspeech.

[15] Jeon, J. H., and Liu, Y. (2010, September), "Syllable-level prominence detection with acoustic evidence.", In INTERSPEECH (pp. 1772-1775).

[16] Ni, C., Liu, W.,and Xu, B.(2012). "From English pitch accent detection to Mandarin stress detection, where is the difference?.", Computer Speech and Language, 26(3), 127-148.

[17] Grabe, E., Post, B., and Nolan, F. (2001). The IViE Corpus. Department of Linguistics, University of Cambridge.

[18] Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.

[19] Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In International Conference on Artificial Intelligence and Statistics (pp. 249-256).