INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Deep Learning based Mandarin Accent Identification for Accent Robust ASR

Felix Weninger[1], *Yang Sun*[2], *Junho Park*[1], *Daniel Willett*[*], *Puming Zhan*[1]

[1]Nuance Communications, Inc., Burlington, MA, USA
[2]Nuance Communications, Inc., Aachen, Germany

{felix.weninger,yang.sun,junho.park,puming.zhan}@nuance.com

Abstract

In this paper, we present an in-depth study on the classification of regional accents in Mandarin speech. Experiments are carried out on Mandarin speech data systematically collected from 15 different geographical regions in China for broad coverage. We explore bidirectional Long Short-Term Memory (bLSTM) networks and i-vectors to model longer-term acoustic context. Starting from the classification of the collected data into the 15 regional accents, we derive a three-class grouping via non-metric dimensional scaling (NMDS), for which 68.4% average recall can be obtained. Furthermore, we evaluate a state-of-the-art ASR system on the accented data and demonstrate that the character error rate (CER) strongly varies among these accent groups, even if i-vector speaker adaptation is used. Finally, we show that model selection based on the prediction of our bLSTM accent classifier can yield up to 7.6 % CER reduction for accented speech.

Index Terms: accent identification, paralinguistics, robust automatic speech recognition, error analysis

1. Introduction

There are manifold varieties of Chinese dialects, which are often not mutually intelligible. Although the only official language is Mandarin in both mainland China and Taiwan, recognizable accents exist under the influence of local dialects. The Standard Mandarin is based on the Beijing dialect, and accents are usually distributed regionally. Geographically, northern dialects in China tend to have fewer distinctions than sourthern ones (cf. Figure 1), while many other factors, such as the history and development of cities, as well as education level, play an important role as well [1]. As accented speech poses a practical challenge to ASR systems, the reliable identification of accented Mandarin speech has immediate applications in robust ASR. In this paper, we evaluate various types of classifiers for the Mandarin accent identification task, and propose the usage of a bLSTM accent classifier to switch automatically between standard and accented Mandarin ASR models.

Relation to prior work The task of accent or dialect identification has generally received less attention than language identification [2]. Foreign accent identification in English was the subject of the INTERSPEECH 2016 Computational Paralinguistics Challenge [3]. The best performing system [4] in the Challenge used an approach based on i-vector [5]. In the same evaluation framework, comparable performance of i-vector and deep learning based systems [6] was found. For the task of classifying US vs. UK accents in English ASR, a bLSTM based system that is integrated with the acoustic model (AM) was proposed [7]. The works [8] and [9] presented multi-accent approaches to Mandarin ASR, similar to multi-lingual ASR. In a similar vein, the work [10] introduced accent-specific acoustic features for a Deep Neural Network (DNN) AM. In [11], the authors evaluated the performance of speaker adaptation for accented Mandarin speech. Yet, these studies did not attempt automatic accent classification; in our paper, we consider the case of model selection based on accent prediction. In general, while there are some studies on the tasks of Mandarin tone recognition (e.g., [12, 13]) and tone mispronunciation [14], which are related to the problem at hand, the literature regarding automatic classification of Mandarin accents is scarce and is limited to few, distinct accents [15, 16, 17].

The novel aspects of our paper are: (i) the usage of deep learning and i-vector based systems for discriminating a large variety of regional accents in Standard Mandarin; (ii) an NMDS-based method for error analysis and subsequent grouping of classes to improve robustness of accent identification; (iii) reducing the ASR error rate for accented Mandarin speech by using an accent classifier for model selection.

2. Data Collection

The speech database used in this study contains 135 k utterances (84.7 hours) from 466 speakers. The language is Standard Mandarin as spoken in various regions across China. 15 collection locations were selected for broad coverage (see Figure 1). All speakers are native speakers of the respective dialect region, i.e., they speak the local dialect / language as their first language and learned Standard Mandarin later. Speakers are balanced by gender and across accents (30–32 speakers per accent).

The speech recordings originate from a scripted in-car human-machine interaction scenario. For a realistic setting, the recording was done in mid-size cars (various models per region) while driving on city roads and highways (approximately equal distribution of environments). Moreover, the data collection setup ensures that differences between groups are of acoustic-phonetic, not linguistic nature. All utterances are manually transcribed.

Each speaker was recorded in one separate consecutive recording session. Audio equipment was turned off during all recordings and windows were closed. All data considered in this study were recorded with the same microphone model (Shidu S-43).

3. Accent Identification Experiments

3.1. Methodology

For accent identification, we investigate i-vectors [2] as a baseline approach, as well as deep learning methods. The purpose of these experiments is to verify that the performance improvements from deep learning methods justify the increase in computational complexity compared to i-vector. In particular, we assess the usage of bidirectional LSTM networks [18] to capture the longer-term acoustic context within each speech utterance, which is expected to facilitate accent identification. Furthermore, since it has been shown that the speaker information from

[*] Work conducted while the author was at Nuance Communications

http://dx.doi.org/10.21437/Interspeech.2019-2737

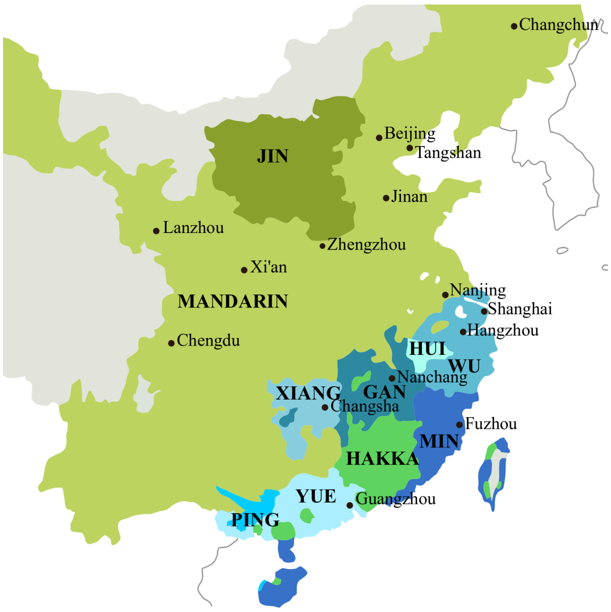


Figure 1: *Locations (indicated by dots) of collection of accented Mandarin speech within Chinese dialect regions.*

i-vector can be used as input feature for DNN acoustic model adaptation [19], we propose to exploit a similar approach for accent ID.

3.1.1. i-vector

i-vectors $\mathbf{v}$ are computed from adapted Gaussian Mixture Models (GMMs) with mean supervector $\mathbf{m}$ and a GMM universal background model (UBM) with mean supervector $\mathbf{u}$,

$$\mathbf{m} = \mathbf{u} + \mathbf{T}\mathbf{v}, \qquad (1)$$

where $\mathbf{T}$ is a total variability matrix [5]. Similar to [9], the GMMs are trained and adapted using Linear Discriminant Analysis (LDA) based features obtained from sliding windows of acoustic input. In accordance with [6], we use accent independent i-vectors. To combine the i-vector with deep learning models, the i-vector is appended to the input features of the DNN or the bLSTM, similar to the method proposed in [19] for speaker adaptation.

3.1.2. DNN and bLSTM Accent Classifiers

The acoustic features used to train deep learning accent identification comprise 45 Mel-frequency Cepstral Coefficients (MFCCs) and 7 fundamental frequency variation (FFV) features [20] extracted at a frame rate of 10 ms and a window size of 25 ms. The FFV features are added to capture variations of tones in accented Mandarin. As a simple deep learning based accent classifier, we explore feed-forward DNNs trained on sliding windows of acoustic input, each spanning 63 contiguous frames (645 ms) of speech. The DNNs have two hidden layers of 512 neurons with rectified linear activation function. The output layer has a softmax activation function and outputs frame-wise posterior probabilities of 15 Mandarin accents and silence, for the center frame of the input window. The input window size and the DNN topology were chosen empirically based on earlier experiments with speaker classification tasks.

Secondly, we investigate bLSTM classifiers which use single frames of acoustic input. The topology consists of two bLSTM layers of size 512, each comprising 256 LSTM memory cells for the forward and backward directions, followed by the softmax output layer. The topology is designed to have a similar number of parameters as the DNN (2.2M vs. 1.9M). The LSTM cells use the hyperbolic tangent activation function. Training is performed on mini-batches of chunks of 64 contiguous frames. As training algorithms for bLSTM, we explored truncated backpropagation through time (BPTT) and context-sensitive chunk BPTT [21, 22]. For the former, the chunks in each mini-batch are continuations of the chunks in the previous mini-batch, and the forward LSTM states are carried over from one mini-batch to the next. For the latter, the order of chunks is completely random, the LSTM states are reset after each chunk, and the number of contextual frames is set to 16 on each side (i.e., chunks overlap by 50%). Note that bLSTMs trained with truncated BPTT are applied to entire utterances at test time, while bLSTMs trained on context-sensitive chunks are evaluated in the same way at test time.

Table 1: *Accent identification accuracy (15 classes) for SVM baseline with speaker i-vectors, and deep learning systems with and without i-vector input. c-bLSTM: bLSTM trained and evaluated on context-sensitive chunks.*

Model	i-vector	Accuracy [%]	
		Speaker	Frame
SVM	✓	15.0	–
DNN	–	25.8	13.40
c-bLSTM	–	32.2	16.22
bLSTM	–	32.2	20.74
DNN	✓	**34.1**	21.73
bLSTM	✓	28.5	**26.09**

3.1.3. Training and Evaluation

The performance of the accent classifiers is evaluated in speaker-independent three-fold cross-validation on the accented Mandarin speech database. The folds are stratified by accent and gender. DNNs and bLSTMs are trained by minimizing the frame-wise cross-entropy loss. The UBM mean vector $\mathbf{u}$ and total variability matrix $\mathbf{T}$ for i-vector estimation (1) are calculated on the training set of each fold. In *training*, i-vectors are estimated using all available data per speaker. To improve generalization, i-vectors are randomly dropped out in training. In *testing*, we perform frame- and speaker-level classification. The speaker-level classification is mainly motivated from potential applications of the accent classifier in robust ASR (cf. above). It is performed by averaging the posterior probabilities of the 15 accent classes over all frames classified as non-silence. We repeat this procedure utterance by utterance, stopping once the total length of the processed utterances exceeds a given maximum number of frames $T_{\max}$. The test i-vectors are calculated per speaker on the same amount of frames. In subsequent experiments, we will explore various settings for $T_{\max}$. First, we report on results with $T_{\max} = 6\,000$ frames (one minute of speech).

3.2. Performance of Regional Accent Identification

Table 1 shows the accuracy of DNN and bLSTM classifiers on the 15-class accent classification task. The frame-level results are given on the 15 accent classes, excluding frames classified as silence. It can be seen that both types of bLSTM models outperform the standard DNN. The bLSTM classifier trained and evaluated on context-sensitive chunks (c-bLSTM) performs considerably worse than standard bLSTM (trained with truncated

BPTT and evaluated on entire utterances) on frame level, but not on speaker level. This can be explained by the smoothing effect of preserving the LSTM state between chunks, which, however, does not contribute to the speaker level accuracy where multiple frame-level scores are averaged. Furthermore, the performance of c-bLSTM is superior to DNN, although both exploit similar information at the input layer. This in accordance with the findings of [22] obtained on an acoustic modeling task.

To assess the performance of speaker i-vectors, we performed an experiment using Support Vector Machine (SVM) classification, yielding 15.0 % accuracy. In a control experiment (not shown in the table), the accuracy of utterance-level i-vectors was estimated at 17.0 % with SVM. Thus, it is evident that for our task, deep learning approaches outperform pure i-vector modeling by a large margin. However, the addition of i-vector to the DNN input yields a significant improvement and achieves the overall best speaker level accuracy. We also note that the combination of bLSTM and i-vector leads to additional improvement in frame accuracy. Yet, speaker accuracy is degraded compared to either DNN + i-vector or bLSTM without i-vector; in fact, the speaker and frame level accuracy are close to each other. We conjecture that the constant i-vector input per frame further adds to the smoothness of the bLSTM output, which has a detrimental effect in the end since the averaging of the outputs will rarely change the speaker level result compared to the frame level outputs.

3.3. Error Analysis and Accent Grouping

While the speaker level accuracy obtained on 15 accent classes (up to 34.1 %) is greatly above random pick chance (6.7 %), it is still not suitable for accent identification in practical applications. Conversely, a coarser grained but robust classification could be sufficient, e.g., to identify accented data for accent-specific AM training, and select AMs accordingly at test time.

Our hypothesis is that a significant share of the error rate can be explained by confusions of the accents of regions which are geographically close to each other, which is founded by the fact that the varieties of the Chinese language family form a dialect continuum [1]. To confirm this hypothesis, we visualize the class confusions of the bLSTM classifier on speaker level, applying the following method. First, we normalize the confusion matrix $\mathbf{C} = (c_{i,j})$ to sum to one: $\overline{\mathbf{C}} = \mathbf{C}/\sum_{i,j} c_{i,j}$. Then, a symmetric distance matrix $\mathbf{D} = (d_{i,j})$ is obtained as

$$\mathbf{D} = (1 - \overline{\mathbf{C}}) + (1 - \overline{\mathbf{C}})^\intercal, \tag{2}$$

where $d_{i,j}$ is now a pairwise distance between accent classes i and j. Finally, Kruskal's method for non-metric dimensional scaling (NMDS) [23] is applied to obtain a two-dimensional space where the Euclidean distances between points i and j, representing accent classes, are a monotonic transformation of the distances $d_{i,j}$.

The resulting NMDS configuration is shown in Figure 2. Upon closer inspection of Figure 2, we can identify three groups of accents that are well separable in the NMDS space: (A1) Beijing, Changchun; (A2) Chengdu, Jinan, Nanjing, Lanzhou, Tangshan, Xi'an, Zhengzhou; (A3) Changsha, Fuzhou, Guangzhou, Hangzhou, Nanchang, Shanghai. Not only are these regions characterized by geographical proximity (A1: north-east, A2: center-west, A3: south-east), but they can also be interpreted as follows: A1: regions where the local dialect is equal or close to Standard Mandarin; A2: regions where a dialect of Mandarin is native language; A3: regions where Mandarin is second language.

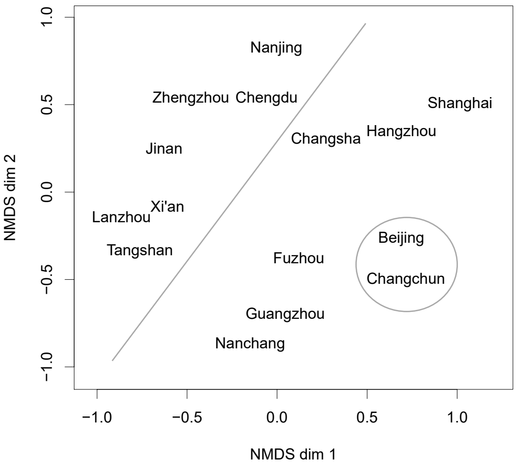


Figure 2: *Non-metric dimensional scaling applied to the distance matrix obtained from the class confusions of the 15-class bLSTM classifier. Gray lines indicate clustering into three accent groups (cf. Section 3.3).*

We can now evaluate the performance of the accent classifiers when mapping the predictions of the 15 accent classes to the groups A1, A2, A3. As performance measure, we opt for unweighted average recall (UAR) [3] since the number of instances is not uniformly distributed across the groups. Based on the predictions of the bLSTM classifier, we obtain 66.4 % UAR. Moreover, the DNN classifier with i-vector input achieves 66.0 % UAR, showing that the accent grouping obtained by analyzing the bLSTM confusions can be generalized.

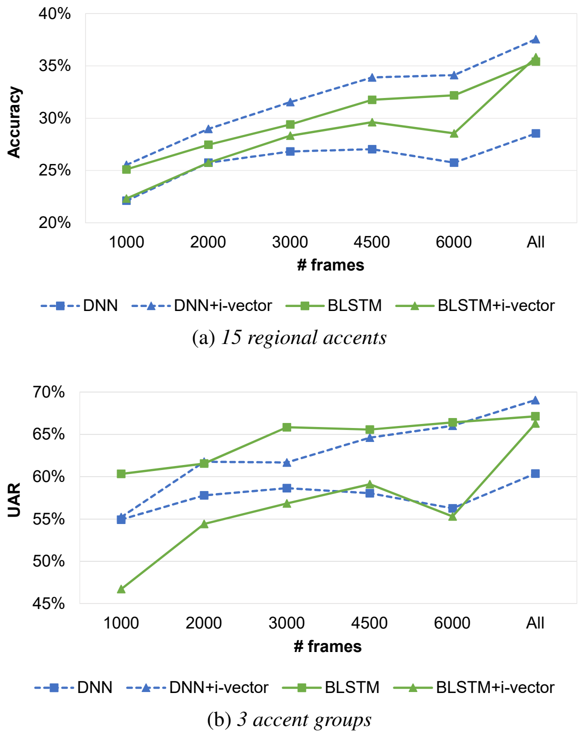
3.4. Effect of Limited Test Data Per Speaker

For practical applications, it is highly relevant to investigate the performance when only a limited amount of data is available per speaker in testing. Figure 3 shows the speaker-level accuracy obtained with varying $T_{\max}$. In case of predicting 15 classes, we see a large impact when varying $T_{\max}$ from 1000 (10 s) to the maximum amount (all frames per speaker, i.e., 10.9 minutes on average). The best performing classifier (DNN + i-vector) has only 25.5 % accuracy on 10 s of input. For predicting 3 classes, the bLSTM + i-vector system is most impacted by input length, especially when comparing 60 s to all frames. This points at overfitting of this classifier, since the i-vectors in training are estimated on the maximum amount of data per speaker. In contrast, the bLSTM without i-vector obtains the best performance (60.3 % UAR) among the classifiers at 10 s of input data, and its performances on 30 s and 60 s of data are very similar.

4. Accent Robustness of ASR

We now proceed to showing that the accent classifier can be used to select accent specific ASR models for improved robustness. Our study is based on an ASR system using a hybrid DNN-HMM AM, with optional online speaker adaptation by i-vectors [24]. The AM has more than 20 M weights and is trained on several thousands of hours of collection and field data from the in-car domain. The language model includes an n-gram and a recurrent neural network component and is targeted for the domain of in-car large vocabulary speech recognition.

We use specific pronunciation models for each of the groups A2 and A3, while we use a Standard Mandarin one for the group



(a) *15 regional accents*

(b) *3 accent groups*

Figure 3: *Speaker-level accuracy / unweighted average recall (UAR) of accent identification for score averaging across varying numbers of frames per speaker.*

A1. From prior knowledge and examining ASR outputs, we know that group A2 is mostly characterized by tone confusions. Thus, we designed a specific pronunciation model for the group A2 that includes possible tone confusions (e.g., if speakers from a certain region regularly use the 4th tone instead of the 1st tone, we add 4th tone pronunciations for the 1st tone characters). Similarly, we created a pronunciation model for the group A3 that allows mainly consonant confusions (e.g., *zh* is often pronounced as *z* by these speakers) along with minor tonal ones. The most suitable pronunciation model for each speaker is selected by using the prediction of the 3-class bLSTM accent classifier (without i-vector input, for $T_{\max} = 6\,000$). We compare this to an oracle experiment where the true accent label is used to switch between pronunciation models.

Table 2 shows the CER achieved on the accented Mandarin speech database, subdivided into the 15 locations of collection as well as the three accent groups. As expected, the group A1, whose accent is close to Standard Mandarin, exhibits the lowest CER. Moreover, A2 shows higher CER than A3. An intuitive explanation for this is that speakers from the A3 group consciously switch between Standard Mandarin and their mother tongue, while there is a fuzzy boundary between local dialect and Standard Mandarin in the regions corresponding to A2, which leads to a stronger accent on average.

First, we discuss the results without speaker adaptation. Using the true accent label for model selection, we observe 13 % relative CER reduction (CERR) for the group A2, i.e., the speakers with heavy accents. A similar CERR (11 %) for this group can be obtained when using the predicted accent group, which is notable given the challenge of correct accent identification. However, for the group A3, accent model selection shows no benefit overall. We obtained improvements for part of the speakers in A3, which were, however, canceled by degradation for other speakers. In contrast, we found that the accent model selection helped uniformly for almost all of the heavily accented speakers (A2). On group A1, there is a slight degradation (4.32 to 4.44 % WER) when using the accent prediction instead of the true label, which can be explained by misclassification of Standard Mandarin speech into one of the accent classes, which causes a less precise pronunciation model to be selected. We could likely avoid some of this degradation by using a confidence threshold for selecting accent specific models, trading in some of the gain on heavily accented speech.

Table 2: *Character error rate (CER) on accented speech data summarized by accent groups and classes, using ASR systems with and without i-vector speaker adaptation, and additionally with accent model selection using ground truth accent labels (Label) or predictions (Pred) of the 3-class bLSTM classifier.*

Spk. adaptation		–	–	–	✓	✓
Model selection		–	Label	Pred	–	Pred
A1	Beijing	4.2	4.2	4.4	**3.7**	3.9
	Changchun	4.4	4.4	4.5	**3.8**	3.8
	Avg.	4.3	4.3	4.4	**3.8**	3.9
A2	Chengdu	5.6	6.2	6.1	**5.1**	5.4
	Jinan	8.8	7.8	7.9	7.6	**7.2**
	Lanzhou	13.7	11.7	12.3	11.4	**10.4**
	Nanjing	9.5	8.3	8.6	7.8	**7.4**
	Tangshan	7.2	6.7	6.7	5.9	**5.7**
	Xi'an	12.5	9.5	10.0	10.5	**8.9**
	Zhengzhou	10.1	8.4	8.6	8.8	**7.8**
	Avg.	9.6	8.4	8.6	8.2	**7.5**
A3	Changsha	6.4	6.3	6.2	5.6	**5.5**
	Fuzhou	5.5	5.5	5.7	**4.8**	5.0
	Guangzhou	5.9	6.0	6.1	**5.1**	5.2
	Hangzhou	6.7	6.5	6.5	**5.6**	5.7
	Nanchang	6.7	6.8	7.0	**5.5**	5.8
	Shanghai	7.0	7.1	7.2	**5.8**	6.1
	Avg.	6.4	6.4	6.4	**5.4**	5.5

With i-vector speaker adaptation, we obtain relative CERRs of 13.2 %, 15.3 % and 14.6 % for the groups A1, A2 and A3 respectively. This shows that speaker adaptation helps regardless of the accent, as expected. On top of i-vector speaker adaptation, accent model selection notably delivers 8.5 % relative CERR for the group A2. The CER for the A3 and A1 groups is unchanged, which is similar to the case without speaker adaptation.

5. Conclusions

In this study, we have analyzed various approaches based on deep learning and i-vector to identify accented Mandarin speech. The error analysis of accent classification led us to propose a 3-class grouping, which can be used to select accent-specific pronunciation models. We have demonstrated that model switching based on accent prediction can yield CER improvements for a state-of-the art ASR system, even if speaker adaptation is already in place. In future work, we aim to use tone information from ASR to improve accent classification. Moreover, we will explore using the accent classifier to spot accented training data for offline AM adaptation.

6. Acknowledgements

We thank Ran Xu and the Nuance China R&D team for specifying, driving and providing the data collection.

7. References

[1] J. Norman, *Chinese*. Cambridge, UK: Cambridge University Press, 1988.

[2] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via ivectors and dimensionality reduction," in *Proc. of 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Florence, Italy: ISCA, 2011, pp. 857–860.

[3] B. W. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. C. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Proc. of 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 2001–2005.

[4] A. Abad, E. Ribeiro, F. Kepler, R. F. Astudillo, and I. Trancoso, "Exploiting phone log-likelihood ratio features for the detection of the native language of non-native english speakers," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 2413–2417.

[5] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Frontend factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[6] M. Senoussaoui, P. Cardinal, N. Dehak, and A. L. Koerich, "Native language detection using the i-vector framework." in *Proc. of 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 2398–2402.

[7] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[8] Y. Liu and P. Fung, "Multi-accent Chinese speech recognition," in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, PA, 2006, pp. 133–136.

[9] M. Chen, Z. Yang, J. Liang, Y. Li, and W. Liu, "Improving deep neural networks based multi-accent mandarin speech recognition using i-vectors and accent-specific top layer," in *Proc. of 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Dresden, Germany: ISCA, 2015, pp. 3620–3624.

[10] J. Yi, H. Ni, Z. Wen, and J. Tao, "Improving BLSTM RNN based mandarin speech recognition using accent dependent bottleneck features," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*. Jeju, Korea: IEEE, 2016, pp. 1–5.

[11] D. Yang, I. Koji, and S. Furui, "Accent analysis for Mandarin large vocabulary continuous speech recognition," Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, Tech. Rep., 2008.

[12] C. Chen, R. C. Bunescu, L. Xu, and C. Liu, "Tone classification in mandarin chinese using convolutional neural networks," in *Proc. of 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 2150–2154.

[13] G.-A. Levow, "Context in multi-lingual tone and pitch accent recognition," in *Proc. of 9th European Conference on Speech Communication and Technology*. Lisbon, Portugal: ISCA, 2005, pp. 1809–1812.

[14] L. Zhang, C. Huang, M. Chu, F. Soong, X. Zhang, and Y. Chen, "Automatic detection of tone mispronunciation in Mandarin," in *Proc. of International Symposium on Chinese Spoken Language Processing*, ser. Lecture Notes in Computer Science, vol. 4274. Springer, 2006, pp. 590–601.

[15] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian mixture models," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Madonna di Campiglio, Italy: IEEE, 2001, pp. 343–346.

[16] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon, "Accent detection and speech recognition for Shanghai-accented mandarin," in *Proc. of 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2005, pp. 217–220.

[17] Y.-F. Liao, S.-C. Yeh, M.-F. Tsai, W.-H. Ting, and S.-C. Chang, "Latent prosody model-assisted Mandarin accent identification," in *Proc. of 21st Conference on Computational Linguistics and Speech Processing*. Taichung, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), 2009, pp. 125–136.

[18] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vancouver, Canada: IEEE, 2013, pp. 6645–6649.

[19] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Olomouc, Czech Republic: IEEE, 2013, pp. 55–59.

[20] K. Laskowski, M. Heldner, and J. Edlund, "The fundamental frequency variation spectrum," in *Proc. of FONETIK*. Gothenburg, Sweden: Citeseer, 2008, pp. 29–32.

[21] K. Chen and Q. Huo, "Training deep bidirectional LSTM acoustic model for LVCSR by a context-sensitive-chunk BPTT approach," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 7, pp. 1185–1193, 2016.

[22] A.-r. Mohamed, F. Seide, D. Yu, J. Droppo, A. Stoicke, G. Zweig, and G. Penn, "Deep bi-directional recurrent networks over spectral windows," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ, 2015, pp. 78–83.

[23] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[24] X. Li, Y. Pan, M. Gibson, and P. Zhan, "DNN online adaptation for automatic speech recognition," in *Proc. of 29th Conference on Electronic Speech Signal Processing (ESSV)*. Ulm, Germany: TUDpress, 2018.