深度神经网络建模方法用于数据缺乏的带口音普通话语音识别的研究

谢旭荣[1,3] 隋 相[1,3] 刘循英[1,2] 王 岚[1,3]

[1](中国科学院深圳先进技术研究院人机智能协同系统重点实验室 深圳 518055)
[2](剑桥大学工程系 剑桥 CB2 1TN)
[3](香港中文大学 香港 999077)

摘 要 众所周知中文普通话被众多的地区口音强烈地影响着，然而带不同口音的普通话语音数据却十分缺乏。因此，普通话语音识别的一个重要目标是恰当地模拟口音带来的声学变化。文章给出了隐式和显式地使用口音信息的一系列基于深度神经网络的声学模型技术的研究。与此同时，包括混合条件训练，多口音决策树状态绑定，深度神经网络级联和多级自适应网络级联隐马尔可夫模型建模等的多口音建模方法在本文中被组合和比较。一个能显式地利用口音信息的改进多级自适应网络级联隐马尔可夫模型系统被提出，并应用于一个由四个地区口音组成的、数据缺乏的带口音普通话语音识别任务中。在经过序列区分性训练和自适应后，通过绝对上 0.8% 到 1.5%(相对上 6% 到 9%)的字错误率下降，该系统显著地优于基线的口音独立深度神经网络级联系统。

关键词 语音识别；决策树；深度神经网络；口音；自适应
中图分类号 TP 391.4 **文献标志码** A

Investigation of Deep Neural Network Acoustic Modelling Approaches for Low Resource Accented Mandarin Speech Recognition

XIE Xurong[1,3] SUI Xiang[1,3] LIU Xunying[1,2] WANG Lan[1,3]

[1](*Key Laboratory of Human-Machine Intelligence-Synergy Systems*, *Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*Cambridge University Engineering Department*, *Cambridge* CB2 1TN, *U.K.*)
[3](*The Chinese University of Hong Kong*, *Hong Kong* 999077, *China*)

Abstract The Mandarin Chinese language is known to be strongly influenced by a rich set of regional accents, while Mandarin speech with each accent is of quite low resource. Hence, an important task in Mandarin speech recognition is to appropriately model the acoustic variabilities imposed by accents. In this

Received: 2015-08-12 **Revised:** 2015-09-14
Foundation: National Natural Science Foundation of China (NSFC 61135003); Shenzhen Fundamental Research Program (JCYJ20130401170306806, JC201005280621A)
Author: Xie Xurong, Research Assistant. His research interests are speech recognition and machine learning; Sui Xiang, Master. Her research interest is speech recognition; Liu Xunying, Senior Research Associate. His research interests include large vocabulary continuous speech recognition, language modelling, noise robust speech recognition, speech and language processing; Wang Lan (corresponding author), Professor. Her research interests are large vocabulary continuous speech recognition, speech visualization and speech centric human-machine interaction, E-mail: lan.wang@siat.ac.cn.

paper, an investigation of implicit and explicit use of accent information on a range of deep neural network (DNN) based acoustic modeling techniques was conducted. Meanwhile, approaches of multi-accent modelling including multi-style training, multi-accent decision tree state tying, DNN tandem and multi-level adaptive network (MLAN) tandem hidden Markov model (HMM) modelling were combined and compared. On a low resource accented Mandarin speech recognition task consisting of four regional accents, an improved MLAN tandem HMM systems explicitly leveraging the accent information was proposed, and significantly outperformed the baseline accent independent DNN tandem systems by 0.8%-1.5% absolute (6%-9% relative) in character error rate after sequence level discriminative training and adaptation.

Keywords　speech recognition; decision tree; deep neural network; accent; adaptation

1 Introduction

An important part of the Mandarin speech recognition task is to appropriately handle the influence from a rich set of diverse accents. There are at least seven major regional accents in China[1,2]. The related variabilities imposed on accented Mandarin speech are complex and widespread. The resulting high mismatch can lead to severe performance degradation for automatic speech recognition (ASR) tasks. To handle this problem, ASR systems can be trained on large amounts of accent specific speech data[3]. However, collecting and annotating accented data is very expensive and time-consuming. Hence, the amount of available accent specific speech data is often quite limited.

An alternative approach is to exploit the accent independent features among standard Mandarin speech data, which are often available in large amounts, to improve robustness and generalization. Along this line, two categories of techniques can be used. The first category of techniques aims to directly adapt systems trained on standard Mandarin speech data[4-8]. The second category uses standard Mandarin speech to augment the limited in-domain accent specific data in a multi-style training framework[9]. For example, an accent dependent phonetic decision tree tying technique was proposed[10-12]. It allows the resulting acoustic models to explicitly learn both the accent independent and the accent specific characteristics in speech.

Recently, deep neural networks (DNNs) have become increasing popular for acoustic modelling, due to their inherent robustness to the highly complex factors of variabilities found in natural speech[13-15]. These include external factors such as environment noise[16-18] and language dependent linguistic features[19-21]. In order to incorporate DNNs, or multi-layer perceptrons (MLPs) in general, into hidden Markov model (HMM)-based acoustic models, two approaches can be used. The first uses a hybrid architecture that estimates the HMM state emission probabilities using DNNs[22]. The second approach uses an MLP or DNN[17], which is trained to produce phoneme posterior probabilities, as a feature extractor. The resulting probabilistic features[23] or bottleneck features[24] are concatenated with standard front-ends and used to train Gaussian mixture model (GMM)-HMMs in a tandem fashion. As GMM-HMMs remain as the back-end classifier,

the tandem approach requires minimum change to the downstream techniques, such as adaptation and discriminative training, while retaining the useful information by the bottleneck features.

Using limited amounts of accented data alone is insufficient to obtain sufficient generalization for the resulting acoustic models, including DNNs. Therefore, a key problem in accented Mandarin speech recognition with low resources, as considered in this paper, is how to improve coverage and generalisation by exploiting the commonalities and specialties among standard and accented speech data during training. Using conventional multi-style DNN training based on a mix of standard and accented Mandarin speech data, accent independent features found in both can be implicitly learned[25,26].

Inspired by recent works on multi-lingual low resource speech recognition[19,20,21,27], this paper aims to investigate and compare the explicit as well as the implicit uses of accent information in state-of-the-art deep neural network (DNN) based acoustic modelling techniques, including conventional tied state GMM-HMMs, DNN tandem systems and multi-level adaptive network (MLAN)[27,28] tandem HMMs. These approaches are evaluated on a low resource accented Mandarin speech recognition task consisting of accented speech collected from four regions: Guangzhou, Chongqing, Shanghai and Xiamen. The improved multi-accent GMM-HMM and MLAN tandem systems explicitly leveraging the accent information during model training significantly outperformed the baseline GMM-HMM and DNN tandem HMM systems by 0.8%-1.5% absolute (6%-9% relative) in character error rate after minimum phone error (MPE) based discriminative training and adaptation.

The rest of this paper is organized as follows. Standard acoustic accent modelling approaches are reviewed in section 2. These include multi-accent decision tree state tying for GMM-HMM systems, and multi-accent DNN tandem systems. MLAN tandem systems with improved pre-training for accent modeling are presented in section 3.2. Experimental results are presented in section 4. Section 5 draws the conclusions and discusses future work.

2 Acoustic modelling for accented speech

2.1 Multi-style accent modelling

Multi-style training[9] is used in this paper for accent modelling. This approach uses speech data collected in a wide range of styles and domains. Then, it exploits the implicit modelling ability of mixture models used in GMM-HMMs and, more recently, deep neural networks[16,20,21] to obtain a good generalization to unseen situations. In the accented speech modelling experiments of this paper, large amount of standard Mandarin speech data is used to augment the limited accented data during training to provide useful accent independent features.

2.2 Multi-accent decision tree state tying

As the phonetic and phonological realization of Mandarin speech is significantly different between regional accents, inappropriate tying of context dependent HMM states associated with different accents can lead to poor coverage and discrimination for GMM-HMM based acoustic models. In order to handle this problem, decision tree clustering[10, 11] with multi-accent branches is tried in this paper. In order to effectively exploit the commonalities and specificities found in standard and accented Mandarin data, accent dependent (AD) questions are

used together with conventional linguistic questions during the clustering process. A sample of the accented branches is shown in red part of Fig. 1.

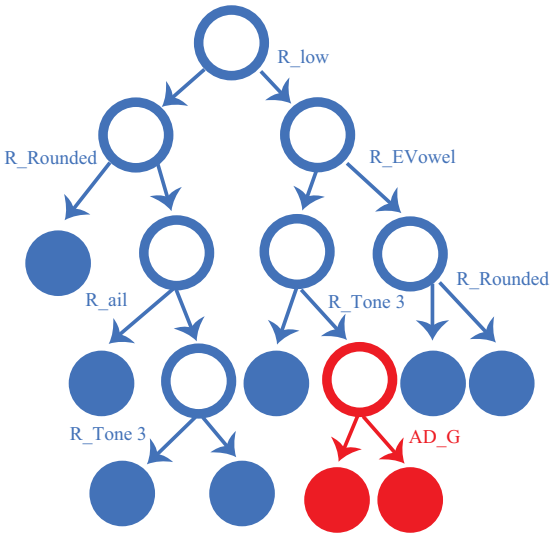

Fig. 1　A part of multi-accent decision tree. Blue: conventional branches; Red: accented branches

In common with standard maximum likelihood (ML) based phonetic decision tree tying[13], the questions giving highest log-likelihood improvement are chosen when splitting tree nodes. The algorithm iterates until no more splitting operations can yield a log-likelihood increase above a certain threshold. Therefore, the multi-accent information is explicitly used during states tying. As expected, the use of accent dependent questions dramatically increases the number of context-dependent phone units to consider during training and decoding. As not all of them are allowed by the lexicon, following the approach proposed in Liu's report[29], only the valid subset under the lexical constraint is considered in this paper.

2.3　Multi-accent DNN tandem systems

In this paper, DNNs are trained to extract bottleneck features to be used in both DNN tandem and MLAN tandem systems. They are trained to model frame posterior probabilities of context-dependent phone HMM state targets. The inputs to DNNs consist of a context window of 11 successive frames of features for each time instance. The input to each neuron of each hidden layer is a linearly weighted sum of the outputs from the previous layer, before fed into a sigmoid activation function. At the output layer a softmax activation is used to compute posterior probability of each output target. The networks were first pre-trained by layer-by-layer restricted Boltzmann machine (RBM) pre-training[14,15], then globally fine-tuned to minimize the frame-level cross-entropy by back-propagation. Moreover, the last hidden layer is set to have a significantly smaller number of neurons[24]. This narrow layer introduces a constriction in dimensionality while retaining the information useful for classification. Subsequently, low dimensional bottleneck features can be extracted by taking neuron values of this layer before activation. The bottleneck features are then appended to the standard acoustic features and used to train the back-end GMM-HMMs in tandem systems.

3　Multi-accent MLAN tandem systems

3.1　Multi-level adaptive network tandem systems

A multi-level adaptive network (MLAN) was first proposed for cross domain adaptation[27,28], where large amounts of out-of-domain telephone and meeting room speech were used to improve the performance of an ASR system trained on a limited amount of in-domain multi-genrearchive broadcast data. The MLAN approach explored the useful domain independent characteristics in the out-of-domain data to improve in-domain modelling performance, while reducing the mismatch across different domains. In this paper, the MLAN approach

was further exploited to improve the performance of accented Mandarin speech recognition systems.

An MLAN system consists of two component subnetworks. The first-level network is trained with acoustic features of large amounts of accent independent standard Mandarin speech data. The acoustic features of target accented speech data are then fed forward through the first-level network. The resulting bottleneck features are then concatenated with the associated standard acoustic features and used as input to train the second-level network. After both of two component networks are trained, the entire training set, including both standard and accented Mandarin speech data, is fed forward through two subnetworks in turn. The resulting set of bottleneck features are then concatenated with the standard front-ends and used to train the back-end GMM-HMMs.

3.2 Improved MLAN tandem systems for accent modelling

The MLAN framework can be considered as stacked DNNs that consists of multi level of networks[21,20]. The second level network of stacked DNNs uses the information of first level network only in the input features, while weights and biases in the second level network are randomly initialized before pre-training and training. One important issue associated with conventional MLAN systems is the robust estimation of the second level DNN parameters. When using limited amounts of in-domain, accent specific speech data to adapt the second level DNN, as is considered in this work, a direct update of its associated weight parameters presents a significant data sparsity problem and can lead to poor generalization performance[21,25,26]. In order to address this issue, an improvement form of pre-training initialization is used in this paper for the second level DNN.

First, all the hidden layers parameters of the second level accent adaptive DNN, and its input layer parameters associated with the standard acoustic features (shown as red and orange parts in Fig. 2) are initialized using those of the first level DNN trained on sufficient amounts of accent independent speech data. Second, the remaining input layer weights and biases connecting the input bottleneck features generated from first level DNN are initialized using RBM pre-training (shown as green in Fig. 2).

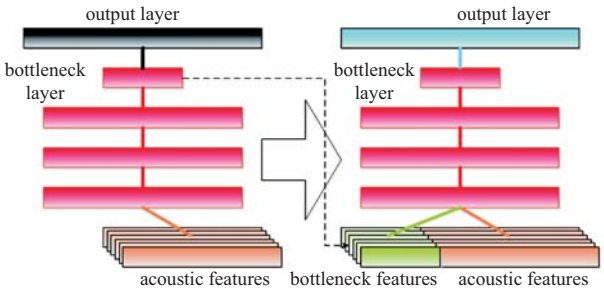

Fig. 2 Improved MLAN training for tandem systems. Left: first level DNN network; Right: second level DNN network

When training the second level DNN, the parameters between bottleneck layer and output layer are updated first (shown as blue in Fig. 2), while fixing the rest of the second level network. The entire second level network is then globally fine-tuned using back-propagation. Similar to the multilingual DNN adaptation approach investigated in Grezl's report[21], the proposed method aims to adapt the second level network parameters based on those of a well trained first level network.

4 Experiments and results

4.1 Data description

In this section the performance of various accented

Mandarin speech recognition approaches are evaluated. 43 hours of standard Mandarin speech[30] and 22.6 hours of accented Mandarin speech containing Guangzhou, Chongqing, Shanghai and Xiamen regional accents[31] released by CASIA and RASC863 databases respectively were used in training. Four testing sets associated with each of these four accents respectively were also used. More detailed information of these data sets was presented in Table 1.

4.2 Experiment setup

Baseline context-dependent phonetic decision tree clustered[13,32] triphone GMM-HMM systems with 16 Gaussians per state were trained using 42 dimensional acoustic features consisting of heteroskedastic linear discriminant analysis (HLDA) perceptual linear predictive (PLP) features and pitch parameters. These were used as the input features, and to produce accent independent state level alignment to train DNNs with 2 048 neurons in each non-bottleneck hidden layer using the Kaldi toolkit[33]. Meanwhile the bottleneck layer had 26 neurons. All DNNs were trained with initial learning rate of 0.008 and the commonly used new bob annealing schedule. Mean normalization and principle component analysis (PCA) de-correlation were applied to the resulting bottleneck features before being appended to the above acoustic features.

4.3 Performance of multi-accent GMM-HMM systems

The performance of multi-accent GMM-HMM systems were first evaluated on Guangzhou accented speech data. These are shown in Table 2. In this table, the “Model AD” column denotes accent dependent questions were used in decision tree state tying. This table shows that the multi-accent HMM model (System (2)) trained by adding all four types of accented speech to the standard Mandarin data outperformed folding in Guangzhou accent data only (System (1)). In addition, the explicit use of accent information during decision tree clustering (System (3)) obtained a further character error rate (CER) reduction of 2.7% absolute from 17.77% down to 15.07%.

4.4 Performance of multi-accent DNN tandem systems

A second set of experiments comparable to those

Table 1　Standard and accented Mandarin speech data sets

Data Source	Database	Train (h)	Test (h)
Standard(st) Mandarin	CASIA	42.9	—
Guangzhou (GZ) accent	RASC863	6.0	1.7
Chongqing (CQ) accent	RASC863	6.0	1.6
Shanghai (SH) accent	RASC863	5.5	1.4
Xiamen (XM) accent	RASC863	5.4	1.5

Table 2　Performance of baseline GMM-HMM systems trained on standard Mandarin speech plus Guangzhou accent data only, or all four accents of Table 1, and evaluated on Guangzhou accent test set

System	Model	Model AD	Baseline training	CER (%)
(1)	Baseline HMM	×	st＋GZ	20.06
(2)	Baseline HMM	×	st＋all	17.77
(3)	Baseline HMM	√	st＋all	15.07

shown in Table 2 were then conducted to evaluate the performance of four tandem systems on the Guangzhou accent test set. In addition to the standard Mandarin speech data, the Guangzhou accent data, or all four accent types, were also used in DNN training. All DNNs here had 4 hidden layers including the bottleneck layer. These are shown in Table 3. The multi-accent trained DNN tandem system (System (4) in Table 3), which used both accent dependent questions in decision tree based HMM state clustering, and included all four accent types in DNN training, gave the lowest character error rate of 13.16%.

4.5 Performance of multi-accent MLAN tandem systems

The performance of various MLAN tandem systems on Guangzhou accent speech data are shown in Table 4. In addition to the standard Mandarin speech data, all four accent types were used in both baseline HMM and the first level DNN training. The first level DNN had 4 hidden layers. The first four MLAN tandem systems used a conventional random initialization of the second level DNN with 2 or 4 hidden layers prior to pre-training and full network update on the target accent data. As discussed in sections 1 and 3.2, when using limited amounts of accent specific speech data to estimate the second level DNN, a direct update its associated weight parameters can lead to unrobust estimation and poor generalization. This is shown in the first four lines of Table 4. Increasing the number of hidden layers from 2 to 4 for the second level DNN led to further performance degradation. Compared with the best DNN tandem system shown in the bottom line of Table 3, a performance degradation of 0.92% absolute was observed.

In contrast, when the improved pre-training based MLAN tandem system discussed in section 3.2 was used, as shown in the last two rows in Table 4, consistent improvements were obtained using both the accent independent and dependent MLAN tandem configurations over the comparable DNN tandem systems shown in the last two rows of Table 3.

Table 3 Performance of DNN tandem systems on Guangzhou accent test set

System	Model	Model AD	Baseline training	DNN training	CER (%)
(1)	DNN tandem	×	st+GZ	st	17.14
(2)	DNN tandem	×	st+GZ	st+GZ	15.85
(3)	DNN tandem	×	st+all	st+all	14.12
(4)	DNN tandem	√	st+all	st+all	13.16

Table 4 Performance of MLAN tandem systems on Guangzhou accent test set

System	Model	2nd DNN		Model AD	CER (%)
		Initial	Hidden		
(1)	MLAN tandem	random	2	×	14.35
(2)	MLAN tandem	random	4	×	15.42
(3)	MLAN tandem	random	2	√	13.24
(4)	MLAN tandem	random	4	√	14.08
(5)	MLAN tandem	1st DNN	4	×	14.00
(6)	MLAN tandem	1st DNN	4	√	12.96

4.6 Performance evaluation on multiple accent test sets

A full set of experiments were finally conducted to evaluate the performance of various multi-accent systems on four accent test sets: Guangzhou, Chongqing, Shanghai and Xiamen. The performances of systems are presented in Table 5 and Table 6 for the multi-accent GMM-HMM, DNN tandem and improved MLAN tandem systems. "＋MPE" denotes MPE discriminative training[34] performed on the maximum likelihood trained "ML" model, and "＋MLLR" denotes a subsequent maximum likelihood linear regression (MLLR) adaptation[35] on the "＋MPE" model. Moreover, System (0) used only out of domain data, namely standard Mandarin data, to train the GMM-HMMs, which denoted by "$HMM^{(out)}$". Meanwhile, "$HMM^{(ma)}$" denotes multi-accent GMM-HMM systems trained with all accented data as well as standard Mandarin data. Both DNN tandem and improved MLAN tandem systems utilized "$HMM^{(ma)}$ ML" models as their baselines. All DNNs here had 6 hidden layers including the bottleneck layer. "DNN AD" denotes DNN trained with accent dependent state alignment, while all DNNs used in MLAN tandem systems were trained with accent independent state alignment.

A general trend can be found in Table 5 and 6 that the explicit use of accent information in training lead to consistent improvements for GMM-HMM, DNN tandem and MLAN tandem systems. For example, by explicitly using accent information during model training, an absolute CER reduction of 1.5% (relative 9%) was obtained on the GMM-HMM systems (System (2) compared to System (1) in Table 5). Although the improved MLAN tandem systems got less improvement from MPE training than the DNN tandem systems, they got more significant amelioration when MLLR adaptation was utilized. This indicates that the improved MLAN framework is not exclusive to the MLLR adaptation. The best

Table 5　Performance of baseline multi-accent GMM-HMM and DNN tandem systems evaluated on all four accent test sets

System	Model	Model AD	DNN AD	Back-end	CER (%)				
					GZ	CQ	SH	XM	Average
(0)	$HMM^{(out)}$	×	×	ML	27.65	29.78	33.64	37.98	32.26
(1)	$HMM^{(ma)}$	×	×	ML	17.77	20.02	21.57	22.59	20.49
		×	×	＋ME	15.30	17.36	20.12	20.87	18.41
		×	×	＋MLLR	13.75	15.48	17.81	18.94	16.50
(2)		√	×	ML	15.07	18.11	20.71	21.88	18.94
		√	×	＋MPE	12.87	15.48	18.70	19.41	16.62
		√	×	＋MLLR	11.97	13.92	16.56	17.60	15.01
(3)	DNN tandem	×	×	ML	13.44	15.87	19.13	18.05	16.62
		×	×	＋ME	12.71	14.79	18.37	16.97	15.71
		×	×	＋MLLR	11.91	13.77	16.38	16.09	14.54
(4)		√	×	ML	12.83	15.00	18.79	17.67	16.07
		√	×	＋ME	12.16	13.91	17.92	16.49	15.12
		√	×	＋MLLR	11.45	12.83	16.28	15.54	14.03
(5)		√	√	ML	12.94	15.17	18.78	17.57	16.12
		√	√	＋MPE	12.25	13.94	17.94	16.75	15.22
		√	√	＋MLLR	11.36	12.86	16.20	15.72	14.04

performance was obtained using the improved MLAN tandem system with accent dependent modelling (System (4) in Table 6). Using this improved MLAN tandem system, an average CER reduction of 0.8% absolute (6% relative) was obtained over the baseline DNN tandem system trained without explicitly using any accent information (System (3) in Table 5).

Comparing the results to previous works evaluated also on RASC863 database, Zhang et al.[36,37] used the augmented HMM and dynamic Gaussian mixture selection (DGMS), instead of multi-style accent modelling HMM and multi-accent decision tree state tying used in this paper. Their error rate for Guangzhou (Yue), Chongqing (Chuan) and Shanghai (Wu) accented Mandarin ASR stayed above 40% in syllable level (SER), and the best relative SER reduction against HMM trained with standard Mandarin (Putonghua) was about 20%. Although SER is not directly comparable to CER, but can still be seen as a reference. Meanwhile, for these three accents the comparable HMM system in this paper (System (2) in Table 5) obtained ML CER of about 18%, which had relative reduction of more than 40% against system (0) in Table 5. It might because that information of standard Mandarin cannot complement the low resource accented Mandarin in the augmented HMM and DGMS approaches.

5 Conclusions

In this paper, implicit and explicit accent modeling approaches were investigated for low resource accented Mandarin speech recognition. The improved multi-accent GMM-HMM and MLAN tandem systems significantly outperformed the baseline GMM-HMM and DNN tandem HMM systems by 0.8%-1.5% absolute (6%-9% relative) in character error rate after MPE training and adaptation. Experimental results suggest the proposed techniques may be useful for accented speech recognition. Future work will focus on modelling a larger and more diverse set of accents.

Table 6 Performance of improved MLAN tandem systems evaluated on all four accent test sets

System	Model	Model AD	1st DNN trn	Back-end	CER (%)				
					GZ	CQ	SH	XM	Average
			st	ML	13.79	15.96	19.14	18.55	16.86
(1)			st	+MPE	13.36	15.37	18.87	18.14	16.44
			st	+MLLR	12.20	13.98	16.71	16.57	14.87
		×	st+all	ML	13.39	15.14	18.66	17.78	16.24
(2)			st+all	+MPE	13.10	14.57	18.32	17.26	15.81
	MLAN		st+all	+MLLR	11.96	13.19	16.17	15.85	14.29
	tandem		st	ML	12.74	15.01	18.85	17.84	16.11
(3)			st	+MPE	12.50	14.62	18.63	17.39	15.79
			st	+MLLR	11.58	12.98	16.40	15.80	14.19
		√	st+all	ML	12.56	14.53	18.12	17.29	15.63
(4)			st+all	+MPE	12.20	14.04	17.94	16.86	15.26
			st+all	+MLLR	11.10	12.55	15.70	15.46	13.70

References

[1] Liu Y, Fung P. Partial change accent models for accented Mandarin speech recognition [C] // IEEE Workshop on Automatic Speech Recognition and Understanding, 2003: 111-116.

[2] Li J, Zheng TF, Byrne W, et al. A dialectal Chinese speech recognition framework [J]. Journal of Computer Science and Technology, 2006, 21(1): 106-115.

[3] Fisher V, Gao YQ, Janke E. Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer [C] // The 5th International Conference on Spoken Language Processing, Incorporating, 1998: 787-790.

[4] Oh YR, Kim HK. MLLR/MAP adaptation using pronunciation variation for non-native speech recognition [C] // IEEE Workshop on Automatic Speech Recognition &Understanding, 2009: 216-221.

[5] Wang ZR, Schultz T, Waibel A. Comparison of acoustic model adaptation techniques on non-native speech [C] // 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003: 540-543.

[6] Tomokiyo LM, Waibel A. Adaptation methods for non-native speech [J]. Multilingual Speech and Language Processing, 2003: 6.

[7] Liu M, Xu B, Hunng T, et al. Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling [C] // 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000: 1025-1028.

[8] Zheng Y, Sproat R, Gu L, et al. Accent detection and speech recognition for Shanghai-accented Mandarin [C] // Interspeech, 2005: 217-220.

[9] Lippmann RP, Martin E, Paul DB. Multi-style training for robust isolated-word speech recognition [C] // IEEE International Conference on Acoustics, Speech, and Signal Processing, 1987: 705-708.

[10] Young SJ, Odell JJ, Woodland PC. Tree-based state tying for high accuracy acoustic modeling [C] // Proceedings of the Workshop on Human Language Technology, 1994: 307-312.

[11] Reichl W, Chou W. A unified approach of incorporating general features in decision tree based acoustic modeling [C] // 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999: 573-576.

[12] Sim KC, Li H. Robust phone set mapping using decision tree clustering for cross-lingual phone recognition [C] // 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2008: 4309-4312.

[13] Seide F, Li G, Yu D. Conversational speech transcription using context-dependent deep neural networks [C] // Interspeech, 2011: 437-440.

[14] Dahl GE, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition [J] IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 30-42.

[15] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups [J] IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[16] Seltzer ML, Yu D, Wang Y. An investigation of deep neural networks for noise robust speech recognition [C] // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: 7398-7402.

[17] Yu D, Seltzer ML. Improved bottleneck features using pretrained deep neural networks [C] // The 12th Annual Conference of the International Speech Communication Association, 2011: 237-240.

[18] Xie X, Su R, Liu X, et al. Deep neural network bottleneck features for generalized variable parameter HMMs [C] // The 15th Annual Conference of the International Speech Communication Association, 2014: 2739-2743.

[19] Thomas S, Seltzer ML, Church K, et al. Deep neural network features and semi-supervised training for low resource speech recognition [C] // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: 6704-6708.

[20] Knill KM, Gales MJF, Rath SP, et al. Investigation

of multilingual deep neural networks for spoken term detection［C］// 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, 2013: 138-143.

［21］ Grezl F, Karafiat M, Vesely K. Adaptation of multilingual stacked bottle-neck neural network structure for new language［C］// 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, 2014: 7654-7658.

［22］ Bourlard HA, Morgan N. Connectionist Speech Recognition: A Hybrid Approach［M］. USA: Academic Publishers, 1993.

［23］ Hermansky H, Ellis DW, Sharma S. Tandem connectionist feature extraction for conventional HMM systems［C］// 2000 IEEE International Conference on Acoustics, Speech and Signal Processing, 2000: 1635-1638.

［24］ Grezl F, Karafiat M, Kontar S, et al. Probabilistic and bottle-neck features for LVCSR of meetings［C］// 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, 2007: 757-760.

［25］ Wang H, Wang L, Liu X. Multi-level adaptive network for accented Mandarin speech recognition［C］// The 4th IEEE International Conference on Information Science and Technology, 2014: 602-605.

［26］ Sui X, Wang H, Wang L. A general framework for multi-accent Mandarin speech recognition using adaptive neural networks［C］// The 9th IEEE International Symposium on Chinese Spoken Language Processing, 2014: 118-122.

［27］ Bell P, Swietojanski P, Renals S. Multi-level adaptive networks in tandem and hybrid ASR systems［C］// 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: 6975-6979.

［28］ Bell PJ, Gales MJF, Lanchantin P, et al. Transcription of multi-genre media archives using out-of-domain data［C］// 2012 IEEE on Spoken Language Technology Workshop, 2012: 324-329.

［29］ Liu X, Gales MJF, Hieronymus JL, et al. Investigation of acoustic units for LVCSR systems［C］// 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, 2011: 4872-4875.

［30］ 中国科学院自动化研究所. CASIA 北方口音语音库［OL］.［2015-07-28］. http://www.chineseldc.org/doc/CLDC-SPC-2004-015/intro.htm.

［31］ 中国社会科学院语言所. 四大方言普通话语音语料库［OL］.［2015-08-02］. http://www.chineseldc.org/doc/CLDC-SPC-2004-004/intro.htm.

［32］ Young SJ, Evermann G, Gales MJF, et al. The HTK Book（Revised for HTK version 3.4. 1）［M］. Cambridge University, 2009.

［33］ Ghoshal A, Povey D. The Kaldi speech recognition toolkit［EB/OL］. 2013-02-03［2015-08-02］. http://kaldi.sourceforge.net.

［34］ Povey D, Woodland PC. Minimum phone error and I-smoothing for improved discriminative training［C］// 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002: 105-108.

［35］ Gales MJ, Woodland PC. Mean and variance adaptation within the MLLR framework［J］. Computer Speech & Language, 1996, 10(4): 249-264.

［36］ Zhang C, Liu Y, Xia Y, et al. Reliable accent specific unit generation with dynamic Gaussian mixture selection for multi-accent speech recognition［C］// 2011 IEEE International Conference on Multimedia and Expo, 2011: 1-6.

［37］ Zhang C, Liu Y, Xia Y, et al. Discriminative dynamic Gaussian mixture selection with enhanced robustness and performance for multi-accent speech recognition［C］// 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, 2012: 4749-4752.