Research on Mongolian speech recognition based on FSMN

Yonghe Wang, Feilong Bao✉, Hongwei Zhang and Guanglai Gao

College of Computer Science, Inner Mongolia University
Huhhot, China, 010021
cswyh92@163.com;csfeilong@imu.edu.cn; ndzhhw@163.com;
csggl@imu.edu.cn

Abstract. Deep Neural Network (DNN) model has been achieved a significant result over the Mongolian speech recognition task, however, compared to Chinese, English or the others, there are still opportunities for further enhancements. This paper presents the first application of Feed-forward Sequential Memory Network (FSMN) for Mongolian speech recognition tasks to model long-term dependency in time series without using recurrent feedback. Furthermore, by modeling the speaker in the feature space, we extract the i-vector features and combine them with the Fbank features as the input to validate their effectiveness in Mongolian ASR tasks. Finally, discriminative training was firstly conducted over the FSMN by using maximum mutual information (MMI) and state-level minimum Bayes risk (sMBR), respectively. The experimental results show that: FSMN possesses better performance than DNN in the Mongolian ASR, and by using i-vector features combined with Fbank features as FSMN input and discriminative training, the word error rate(WER) is relatively reduced by 17.9% compared with the DNN baseline.

Keywords: Mongolian, speech recognition, DNN, FSMN, i-vector, sequence-criterion training.

1 Introduction

Mongolian language has a wide influence in the world, it is the first or the second official language, about 6 millions persons who can speak that language in the Mongolia, Inner Mongolia of China and other districts. However, Mongolian speech recognition is still at its initial research stage. Deep learning methods have been widely used in the field of speech recognition, such as Deep Neural Network (DNN), recurrent neural networks (RNNs) or LSTM-based models [1-3]. But the research of Mongolian speech recognition tasks based on the depth learning method is relatively rare.

Recently, the approaches investigated in [4] proposed a simpler structure for memory neural networks, namely feedforward sequential memory networks (FSMN), which has been proven to have advanced performance on the LVCSR task than DNN and LSTM [4,5]. FSMN extends the standard feedforward neural networks with some

learnable memory blocks in hidden layers. The memory blocks are used to store a fixed-size long context information temporarily as short-term memory mechanism, which can learn long-term dependencies in sequence data. In this paper, we will introduce FSMN into acoustic modeling in Mongolian speech recognition.

The speech recognition research for Mongolian starts at 2003 in China. The first Mongolian speech recognition system is established by Gao [6]. Then some researches on the design and optimization of the Mongolian acoustic model are undertaken by [7,8]. The approaches investigated in [9] solve the large vocabulary problem in Mongolian, where segmentation-based approach is applied to the system. More recently, in [10] DNN-based acoustic model was first applied to the Mongolian ASR research, it reduces the Word Error Rate (WER) over 50% against the GMM-HMM. However, compared with other languages such as Chinese and English, there is still a lot of room for optimization in Mongolian speech recognition acoustic model.

In order to improve the performance of the Mongolian language recognition acoustic model, firstly this paper applies the FSMN structure to the acoustic model of the Mongolian speech recognition system and we investigate the FSMN architectures using of different hidden layers (including memory blocks). Further, i-vector based adaptation has been shown to be effective in reverberant environments and can be used for rapid adaptation of the neural network [11], in this paper we use i-vector based neural network adaptation. And lastly, discriminative training was conducted over the FSMN by using maximum mutual information (MMI) and state-level minimum Bayes risk (sMBR), respectively [12].

The rest of the paper is organized as follows: Section 2 describes the structure of FSMN. Section 3 draws the i vector extraction method. Section 4 shows MMI and sMBR sequence discriminative training, Section 5 details the experimental setup and Section 6 reports the experiments. The conclusions are presented in Section 7.

2 Acoustic modeling based on FSMN

Feedforward sequential memory network is a multi-layer (normally more than three) feed-forward neural network model with single or multiple memory blocks in the hidden layer, which can learn long-term dependencies in sequence data. Fig.1. shows an FSMN structure diagram with two memory block added into hidden layer. These memory blocks are used to encode the information from the preceding and the subsequent frames. And these information make it possible to model long term dependency in the speech sequence.

Given a sequence $X=\{x_1; x2;...;x_t\}$, each $x_t \in X$ represents an input data at time instance t. The corresponding hidden layer outputs are denoted as $H=\{h_1; h2;...;h_t\}$. The structure of a memory block is illustrated in Fig.2, it uses a tapped-delay structure to encode h_t and its previous N_1 histories activities and N_2 posterior activities into a fixed-sized representation, which is fed into the next hidden layer along with the current hidden activity.

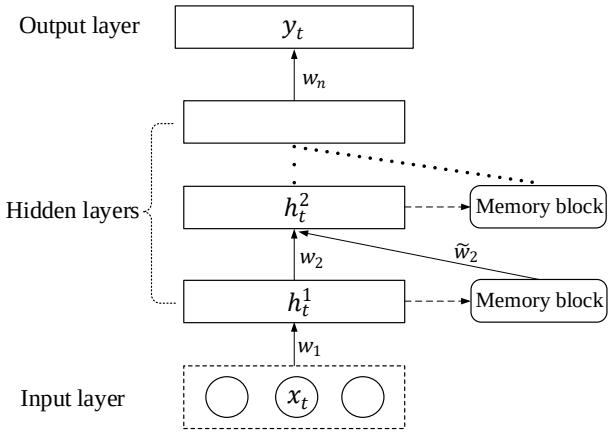

Fig. 1. The structure of FSMN with 2 Memory block.

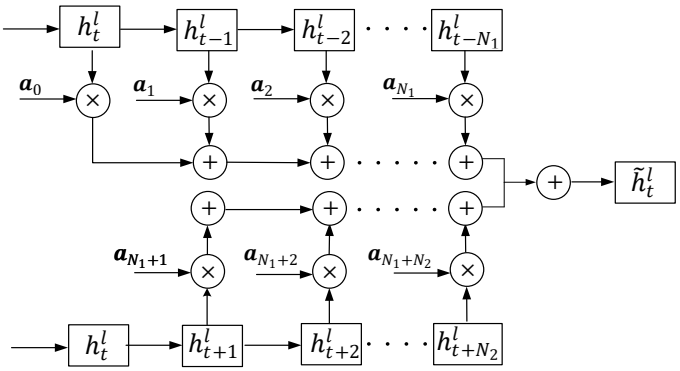

Fig. 2. The structure of Memory block in FSMN.

As the Fig.2. shows, in which the preceding N_1 frames' hidden layer output $h_{t-1}^l,...,h_{t-N_1}^l$, the current frames' hidden layer output h_t^l and the posterior N_2 frames' hidden layer output $h_{t+1}^l,...,h_{t+N_2}^l$ are summed up by the trained weight parameters ∂ into a context code $\tilde{h}_t^l$. Depending on the encoding method to be used, the weight parameters ∂ can be a scalar or a vector.

If the weight coefficient ∂ is set to a scalar, then FSMN is called scalar FSMN (sFSMN). The formula is defined as:

$$\tilde{h}_t^l = \sum_{i=0}^{N1} a_{t,i}^l \cdot h_{t-i}^l + \sum_{j=1}^{N2} a_{t,M_1-1+j}^l \cdot h_{t+j}^l \tag{1}$$

If the weight coefficient ∂ is set to a vector, then FSMN is called vector FSMN (vFSMN), and the formula is defined as:

$$\tilde{h}_t^l = \sum_{i=0}^{N1} a_{t,i}^l \odot h_{t-i}^l + \sum_{j=1}^{N2} a_{t,M_1-1+j}^l \odot h_{t+j}^l \tag{2}$$

vFSMN has better modeling power[5]. We adopt the vFSMN in this study, and refer it as the FSMN for short.

3 i-vector

Separating speech and acoustic environment information from the speech data is important to improve the robustness against the speaker and environment various. i-vector encapsulate all the relevant information about a speaker's identity in a low dimensional fixed-length representation, which are widely used in speaker adaptation of speech recognition. In this paper, we use the standard Gaussian Mixture Model-Universal Background Model (GMM-UBM) to extract the i-vector. Assuming that the speaker and acoustic environment information is mapped to a GMM super vector space, total variability space between the speaker and acoustic environment information is trained by super vector space. The i-vector extraction process is shown in Fig.3.

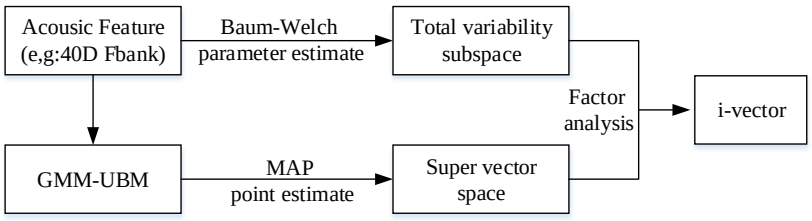

Fig. 3. The structure of i-vector feature extraction

The high-dimensional mean super-vector of a period of speech *h* with the speaker and acoustic environment information is expressed as:

$$M(h) = m + Tw(h) \tag{3}$$

where *m* represents the UBM super vector that is not related to the speaker and the environment. *T* is a total variability subspace matrix and *w* is termed as i-vector. The *M* and *m* are obtained from the GMM-UBM.

We train GMM-UBM on top of features processed with applying cepstral mean and variance normalization (CMVN) and then transforming with an LDA+MLLT matrix. The GMM parameters are initialized by setting the variance to the global variance of the features, the means to distinct randomly chosen frames and using a part of features to do multiple iterations of training by using expectation maximization (EM)

algorithm. For the total variability subspace matrix *T*, the features used to obtain the Gaussian posteriors are based on sliding-window Cepstral Mean Normalization (CMN), but the actual i-vector extractor sees the original features without CMN. The purpose is that we hope that the appropriate offset can be learned by the i-vector extractor itself.

In this study, we use the i-vector as an additional input feature of neural networks that was proposed by [13]. We illustrate the used FSMN structure and the i-vector feature inputs in Fig.4. We use Fourier-transform-based log filter-bank (Fbank) coefficients as the input feature. Consequently, a context window of *d*-dimensional frames of *m* dimensional acoustic features is augmented with *k*-dimensional i-vector resulting in a *dm+k* dimensional as input to the neural network.

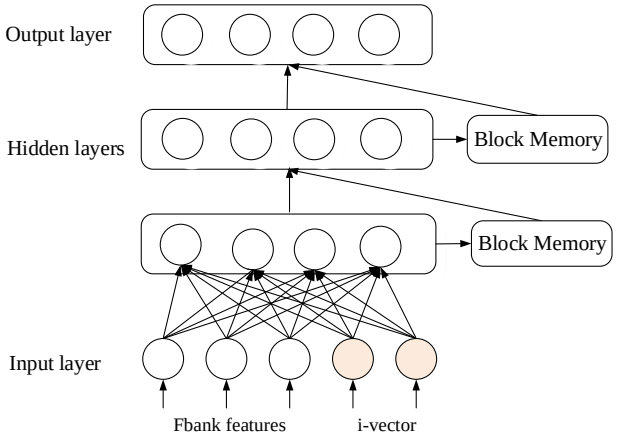

Fig. 4. Diagram of a 2-hidden layer neural network with inputs augmented with i-vectors

4 MMI and sMBR

In the speech recognition system, neural networks are trained to estimate the HMM states posteriors of each frame. To train the neural networks, the cost function is important. Cost function guides the training procedure by setting an optimization direction, and leads the model to a certain optimum. Traditional cost function used in the neural network training only consider the current input. Because they treat every frame as an independent observation, but do not care about the relationship among frames, we called this type of cost functions as frame-level training cost function. Cross-Entropy (CE) is one of the most common choice of frame-level training cost function for the speech recognition system. The CE cost function is defined as:

$$\mathrm{F}_{CE} = -\sum_{u=1}^{U}\sum_{t-1}^{T_u}\log p_\theta(\mathrm{s}_{ut}/x_{ut}) \tag{4}$$

where T_u represents the total time of utterance u. S_u is the reference state label at time t for utterance u. And X_u is given input at time t for utterance u.

Speech is sequence data, where the every frame has strong relationships with each other. Therefore the frame-level training cost function is not enough for the speech recognition system. Different from the frame-level training cost function, Sequence-discriminative training cost function base on the whole sequences rather than frames. In recent years, a variety of sequence-level cost functions have been proposed. Among them, MMI and sMBR are the most common choice.

The MMI criterion is the mutual information between the distributions of the observation and word sequences. The MMI criterion is:

$$\mathrm{F}_{MMI} = \sum_{u=1}^{U} \log \frac{p(X_u / S_u)^k P(W_u)}{\Sigma_w p(X_u / S)^k P(W)} \tag{5}$$

where X_u as the sequence of all observation; W_u as the word-sequence in the reference for utterance u; S_u is the sequence of states corresponding to W_u; and k is the acoustic scaling factor.

The sMBR is designed on the basis of the compliance with the Bayes risk rule to minimize the expected error corresponding to different granularity of labels, which is defined as:

$$\mathrm{F}_{\mathrm{sMBR}} = \sum_{u=1}^{U} \frac{\Sigma_w p(X_u / S)^k P(W) A(W, W_u)}{\Sigma_{w'} p(X_u / S)^k P(W')} \tag{6}$$

where X_u as the sequence of all observation; $A(W,W_u)$ is the raw accuracy of the number of state labels that corresponding to the word sequence W with respect to that corresponding to the reference W_u.

In this study, we use a modified sMBR cost functions that was proposed by [14]. In this modified sMBR cost function, silence, vocalized noise and non-spoken noise are treated as silence. And the silence is treated as any other phone, except that all pdfs of silence phones are collapsed into a single class for the frame-error computation. Therefore, it is not to be penalized to replace one silence phone with another silence phone, but it will be penalized to insert a non-silent phone into a silent area.

5 EXPERIMENTS SETUP

5.1 Dataset

We use a Mongolian speech corpus for acoustic model training, this dataset contains 78 hours of speech sampled at 16KHz and the corpus involves 193 speakers. We divided the corpus into two parts randomly which include a training set and a test set. The training set is about 88% of the whole corpus and each test set is about 12% of the whole corpus. The dataset pronunciation dictionary uses an alphabet of 63 phonemes which include 37 vowels and 26 consonants. For the language model training,

a 3-gram language model with about 85 million tokens of Mongolian training text from Mongolian web sites was used.

5.2 ASR system

We implement the Mongolian LVCSR system based on the Kaldi speech recognition toolkit [15]. The base acoustic features are 40-dimensional Fbank feature vector, these features are then processed using utterance level cepstral mean and variance normalization (CMVN) and features in time taking a context size of 7 frames (i.e., ±3) were spliced and projected to 40 dimensions with linear discriminant analysis (LDA). A maximum likelihood linear transform (MLLT) was estimated on the LDA features to train the LDA+MLLT model and speaker adaptive training was performed with FMLLR transform. FMLLR features were then used for training DNN, FSMN and i-vector.

Traditional neural networks usually employ a non-linear operation such as Sigmoid, Tanh function as an activation function; however, it has been shown that rectifier linear unit (ReLU) can improve the performance of NNs [16]. In this paper, all the NNs were trained using a ReLU nonlinearity activation function. In our experiment, parallel training of the NNs using up to 4 GPUs was done, while uses greedy layer-wise supervised training, preconditioned stochastic gradient descent (SGD) technique updates.

6 EXPERIMENTS

6.1 Baseline Experiments

In our DNN-HMM acoustic model, we used context dependent tri-phones as the units of acoustic modeling. They share a total of 3762 unique context dependent states, which corresponds to the output dimension of the DNN acoustic model. This model uses 40-dimensional Fbank feature, where each feature vector is concatenated with a context window of 15 frames (7+1+7) to yield an input feature vector size of 600 as DNN input. The labels for training data are created by forced alignment with a classic GMM-HMM acoustic model trained on that data. Neural networks denote the standard 6 layers of fully connected neural networks using ReLU activation functions, each of layers has 1024 nodes or 2048 nodes were used, respectively. We use RBM-based pre-training to initialize DNN layer by layer. The mini-batch size is fixed to 256, and the initial learning rate parameter was set to 0.05 and 0.008 in the pre-training and fine-tuning, separately.

We used a FSMN with 6 hidden layers, each hidden layer has 1024 neurons, in which the former three hidden layers have block memories and the latter three layers are normal hidden layer. We also extract 40-dimensional Fbank feature, because of the inherent memory mechanism of FSMN, it does not need to concatenate too many consecutive frames, therefore these features are augmented with the neighboring frames in 3 context window (i.e., ±1) as input features to FSMN[5]. For each block memory, we set the memory range from the preceding 5 frames to the subsequent 5

frames. The neural network is randomly initialized during training, without using any pre-training method. We use SGD with a mini-batch size of 256 for training task. The initial and final learning rates were specified by hand and equal to 0.05 and 0.008 respectively.

Table 1 shows the results of DNNs and FSMN based acoustic modeling trained on the Mongolian dataset. We can see that the FSMN based acoustic modeling can significantly outperform the DNNs. The hidden layer of DNN contains 2048 nodes showed great improvement of word accuracy compared with 1024 nodes. While the hidden unit of FSMN is 1024, but the performance is still better than the DNN contains 2048 nodes, WER is brought down to 12.90% from 11.94% (relative 7.4%), demonstrating its advantage in Mongolian acoustic modeling.

Table 1. The results (%WER) of DNNs and FSMN, "1024L6" is to be interpreted as 6 hidden layers of 1024 hidden units.

Model	Scale	WER(%)
DNN	1024L6	14.56
DNN	2048L6	12.90
FSMN	1024L6	**11.94**

6.2 Experiments on different structure for FSMN

In this section, we investigate the effect of the number of hidden layers of FSMN containing memory blocks on the final Mongolian speech recognition performance. Every FSMN architecture with 6 hidden layers, each of which has 1024 nodes. And all hidden units adopt the rectified linear activation function. In the experiment, instead of adding block memory to each hidden layer, we start adding memory blocks from the hidden layer containing 3 block memories. We use FSMN_3 denotes the former three hidden layers have block memories and the latter three layers are normal hidden layer, FSMN_4 denotes the former four hidden layers have block memories and the latter two layers are normal hidden layer, and so on. The configuration of the neural network in training is the same as that of the baseline FSMN training.

Table 2. The results (%WER) of FSMN for various structure

Model	WER(%)
FSMN_3	11.94
FSMN_4	11.56
FSMN_5	**11.50**
FSMN_6	12.28

From the experimental results in Table 2, we can see that with the increase of block memory in the hidden layer, the WER is decreased significantly. This is because the increase of the block memory would entitle FSMN to acquire more fixed-size long term temporal contexts relationships information. However, when the number of the

block memory in the hidden layer is increasing to 6, the accuracy of speech recognition decreases. The reason is that due to the small amount of the Mongolian training data, the data sparseness would be triggered when the number over block memory of hidden layers reaches 6, which would cause that the neural network cannot learn enough information and the performance of the acoustic model would be compromised.

Eventually, using FSMN network architecture achieved 4.8%-10.8% relative WER reduction over the baseline DNN model. The best performance is obtained using the FSMN_5, the WER is reduced from 12.90% to 11.50%. Overall, experimental results indicate that the effectiveness of FSMN in Mongolian speech recognition, being able to model long term temporal contexts in short-term Mongolian speech features.

6.3 Experiments with i-vector Features

To obtain the i-vector estimation, the GMM-UBM was trained with 512 mixtures computed from 40 perceptual linear prediction coefficients with delta and delta-delta features appended. The LDA and MLLT transforms obtained from GMM-HMM system were used to GMM-UBM. The matrix *T* was randomly initialized using LDA and updated using the EM algorithm with 10 iterations.

In this experiment, we trained DNNs and FSMNs using the baseline experiment configuration. All neural networks use the ReLU activation function, and we extract *k*-dimensional i−vector append to the context window of *d* frames of 40-dimensional Fbank acoustic features resulting in a $40d+k$ dimensional at each time step and the network is fed with these features with no stacking of acoustic frames. In this experiment, we do not apply the cepstral mean normalization to the Fbank features, because i-vector can supply the information about any mean offset of the speakers data, then the network itself is able to do any feature normalization that is needed .To explore the influences of the i-vecotr size, we trained different neural networks with different i-vecotr size. The experimental results are listed in Table 3.

Table 3. The results (%WER) of different dimension i-Vector features using in DNN and FSMN.

Model	i-vector dimensions (k)				
	0	50	100	150	200
DNN	12.90	12.32	**12.24**	12.48	12.54
FSMN	11.94	**11.38**	11.48	11.66	11.97

The eventual experimental results showed that: Compared with the baseline (in first column), by combining the i-vector feature with the Fbank feature as the input, both of the DNNs and FSMNs obtain a considerable improvement. Meanwhile, we observe that different i-vector size will lead to different performance of the neural networks models. For the DNN, the best experimental result is that with a 100-dimensional i-vector led to a WER of 12.24%. For the FSMN, with a 50-dimensional i-vector can obtain the best experimental results, the WER is reduced from 11.94% to 11.38%. However, as the i-vector dimension increases, the accuracy of speech recognition is

reduced. Because the increasing amount of information content of the i-vectors presented to the network would lead to the overfitting. On the whole, the experiment shows that using the i-vector as input features provides the neural networks with valuable information in Mongolian speech recognition.

6.4 Sequence-discriminative training

In this set of experiments, sequence-discriminative training is investigated on Mongolian large vocabulary continuous speech recognition tasks. Different sequence discriminative criteria MMI and sMBR are compared. The initial models used for sequence-discriminative training are trained from baseline experiments using the Maximum likelihood estimation (MLE) as training criteria. The training start from a set of alignments and lattices that are generated by decoding the training data with a unigram language model. The experimental results in Table 4 indicate that the effectiveness of sequence-discriminative training in Mongolian speech recognition, Different training criteria have little effect on the experimental results. For the DNN, the WER is reduced from 12.90% to 12.06% (relative 6.5%).For the FSMN, the WER is reduced from 11.94% to 11.28% (relative 5.5%).

Table 4. The results (%WER) of DNN and FSMN trained using different criteria

Model	training criteria		
	MLE	MMI	sMBR
DNN	12.90	12.12	**12.06**
FSMN	11.94	**11.28**	11.34

Finally, we investigated experiments with FSMN system which is trained using i-vector features and sequence-discriminative criteria on the Mongolian large vocabulary continuous speech recognition task. We select the FSMN model in the baseline experiments as the starting point. As shown in Table 5, we can see that for each auxiliary vector, the finally results consistently are better than the baseline FSMN model. And using the sequence-discriminative training can obtain better results than the i-vector features. Moreover these auxiliary vectors are complementary. The best performance is obtained using the i-vector + sequence training setup. The WER is reduced from 11.94% to 10.59% (relative 11.3%).

Table 5. The results (%WER) of different auxiliary information for FSMN

Model	WER(%)
FSMN	11.94
FSMN + sequence training	11.28
FSMN + iVectors	11.38
FSMN + iVectors + sequence training	**10.59**

7 Conclusions

In this paper, we presents the first application of FSMN networks for Mongolian large vocabulary continuous speech recognition tasks, the experimental results show that the FSMN can obtain better performance than the DNN. Compared among different FSMN model architectures, the former three hidden layers with block memory and the latter three layers being normal hidden layer was found to be optimal, WER was relatively reduced by 10.8%. Further, we show the i-vector features and the sequence discriminative training are both effective in the Mongolian speech recognition system. And these auxiliary vectors are complementary, by combining these two method, we can get a relative improvement of 11.3%. Finally, by using the FSMN, i-vector features and the sequence discriminative training, comparing with the DNN baseline, we obtain a WER relative reduction of 17.9%. And the WER score of the best system is 10.59% which is a very good performance record of the Mongolian speech recognition system.

Acknowledgements. This research was supports in part by the China national natural science foundation (No.61563040, No.61773224) and Inner Mongolian nature science foundation (No. 2016ZD06).

References

1. Hinton G, Deng L, Dong Y, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29(6), 82–97 (2012).
2. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In: 38th ICASSP, pp. 6645–6649. IEEE Press, Vancouver(2013)
3. Sak H, Senior A W, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: 15th INTERSPEECH, pp.338–342. Singapore (2014)
4. Zhang S L, Jiang H, Wei S, et al. Feedforward sequential memory neural networks without recurrent feedback [J]. Computer Science: 1510.02693 (2015)
5. Zhang S, Liu C, Jiang H, et al. Feedforward sequential memory networks: A new structure to learn long-term dependency [J]. Computer Science:1512.08301 (2015)
6. Gao, G., Zhang, S.: A Mongolian speech recognition system based on HMM. In: International Conference on Intelligent Computing, pp. 667--676. Springer, Heidelberg (2006)
7. Qilao, H., Gao, G. L.: Researching of speech recognition oriented mongolian acoustic model. In: 2th Pattern Recognition, 2008. CCPR'08. Chinese Conference, pp. 1--6. IEEE Press, Beijing (2008)
8. Bao, F., Gao, G.: Improving of acoustic model for the mongolian speech recognition system. In: 2th Pattern Recognition, 2009. CCPR 2009. Chinese Conference, pp. 1--5. IEEE Press, Nanjing (2009)
9. Bao, F., Gao, G., Yan, X., Wang, W.: Segmentation-based Mongolian LVCSR approach. In: 38th ICASSP 2013, pp. 1--5. IEEE Press, Vancouver (2013)
10. Zhang, H., Bao, F., Gao, G.: Mongolian speech recognition based on deep neural networks. In: 15th Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 180--188. Springer International Press, Yantai (2015)

11. Alam, M. J., Gupta, V., Kenny, P., Dumouchel, P.: Use of multiple front-ends and I-vector-based speaker adaptation for robust speech recognition. REVERB workshop. (2014)
12. Xue, S., Abdel-Hamid, O., Jiang, H., Dai, L. Fast adaptation of deep neural network based on discriminant codes for speech recognition. IEEE/ACM Transactions on Audio Speech & Language Processing. 22(12), 1713-1725 (2014)
13. Senior, A., Lopez-Moreno, I.: Improving DNN speaker independence with I-vector inputs. In: 39th ICASSP, pp. 225-229. IEEE Press, Florence (2014)
14. Peddinti, V., Chen, G., Povey, D., Khudanpur, S.: Reverberation robust acoustic modeling using i-vectors with time delay neural networks. In: 16th INTERSPEECH, pp. 2440-2444. Dresden (2015)
15. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., & Silovsky, J.: The Kaldi speech recognition toolkit. Workshop on automatic speech recognition and understanding (No. EPFL-CONF-192584). IEEE Signal Processing Society (2011)
16. Maas, A. L., Hannun, A. Y., Ng, A. Y.: Rectifier nonlinearities improve neural network acoustic models. In: 30th ICML Workshop on Deep Learning for Audio, Speech and Language Processing (2013)