Mongolian Speech Recognition Based on Deep Neural Networks

Hui Zhang, Feilong Bao*, and Guanglai Gao

Collage of Computer Science, Inner Mongolia University, Hohhot, China, 010021
alzhu.san@163.com csfeilong@imu.edu.cn csggl@imu.edu.cn

Abstract. Mongolian is an influential language. And better Mongolian Large Vocabulary Continuous Speech Recognition (LVCSR) systems are required. Recently, the research of speech recognition has achieved a big improvement by introducing the Deep Neural Networks (DNNs). In this study, a DNN-based Mongolian LVCSR system is built. Experimental results show that the DNN-based models outperform the conventional models which based on Gaussian Mixture Models (GMMs) for the Mongolian speech recognition, by a large margin. Compared with the best GMM-based model, the DNN-based one obtains a relative improvement over 50%. And it becomes a new state-of-the-art system in this field.

Keywords: Mongolian, Deep Neural Networks (DNNs), Gaussian Mixture Models (GMMs), N-gram Language Model

1 Introduction

More than 7000 living languages are spoken in the world today [1]. However, Automatic Speech Recognition (ASR) systems have been built only for a small number of major languages, such as English, Chinese, Spanish and Arabic. Most other languages are virgin territory for the ASR research. Mongolian is one of the less studied languages for speech recognition.

Mongolian language is used mainly in Mongolia, parts of China (Inner Mongolia, Uugar), Russia (Buryat, Khalmyc) and their neighboring areas. Today, about 6 million people speak Mongolian [1]. There are two written systems in Mongolian language: 1) Traditional Mongolian scripts are used mainly in Inner Mongolia of China. 2) Cyrillic scripts are used mainly in Mongolia. A word can be written in both of the two scripts, and its pronunciation does not change. In this study, we focus on Traditional Mongolian. In Traditional Mongolian, the relationship between grapheme and phoneme is complex. Phonemes may insert, loss or vary during pronouncing. Furthermore, Mongolian has a very large vocabulary. Since it is an agglutinative language, new words can be formed by combining stem with a lot of optional suffixes. These make speech recognition for Mongolian difficult.

* Corresponding author.

The speech recognition research for Mongolian starts at 2003 in China [2–5], and it is just getting started in Mongolia [6]. All of these studies have achieved some success by employing the sophisticated speech recognition models which are the Gaussian Mixture Models and Hidden Markov Models (GMM-HMM). Recently, researchers have made a breakthrough on the ASR by introducing the Deep Neural Networks (DNNs) into this field. New DNN-HMM models are utilized in the ASR systems. The new models have been shown to outperform GMM-HMM models on a variety of speech recognition benchmarks, sometimes by a large margin [7]. In this study, we bring the success of DNN-HMM into the Mongolian ASR research, and build a Mongolian Large Vocabulary Continuous Speech Recognition (LVCSR) system. Experimental results show the DNN-HMM-based system outperforms the GMM-HMM-based one by a large margin.

The rest of the paper is organized as follows. In the next section, we present an overview of the Mongolian speech recognition framework. The experiments and results are presented in section 3. And we conclude the paper in section 4. In the last section, we list the future works.

2 Mongolian Speech Recognition Framework

In this study, we build a LVCSR system for Mongolian. The framework is shown in Fig. 1. Speech recognition is the translation of spoken words into text. We split the whole task into some steps. Each step recognizes some level of units in the speech. Specifically, we first recognize the phonemes, then use the recognized phonemes to make up words, finally output the recognized text.

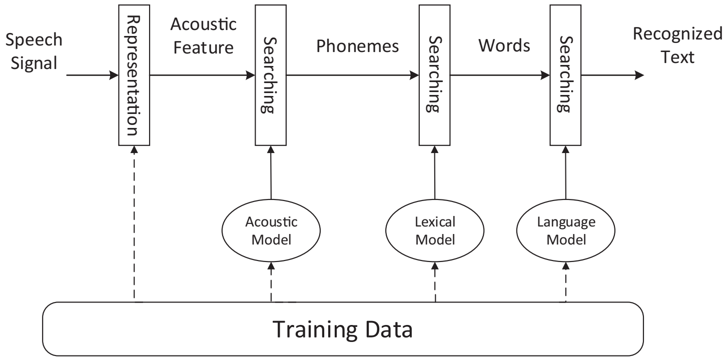

Fig. 1. Speech recognition framework

The entire processing can be split into two stages, the representation stage and the searching stage.

2.1 Representation

In the GMM-HMM-based ASR systems, the speech signal is typically represented by the Mel-Frequency Cepstral Coefficients (MFCCs) [8] computed from the raw waveform. And the MFCCs usually are concatenated with their first and second order temporal differences to get some dynamic properties [9]. MFCC is designed to discard the large amount of information in waveforms that is considered to be irrelevant for recognition and to express the remaining more useful information.

But in DNN-HMM-based ASR systems, the Fourier-transform-based log filter-bank (fbank) coefficients outperform the MFCCs [10], because fbank remains more information. Compared with the MFCCs, the fbank are strongly correlated. Modeling fbank with GMM is computationally expensive. Therefore conventional ASR systems use the MFCCs rather than the fbank.

Both the MFCCs and the fbank are nonadaptive. We can also employ some acoustic feature adaptations by utilizing the training data. Linear Discriminant Analysis (LDA) is one of such methods. LDA transform the raw feature into a new one which facilitates discrimination by supervised training. And $f$eature space Maximum Likelihood Linear Regression (fMLLR) [11] is another feature adaptation method but focuses on speaker adaptation, which aimed to eliminate the mismatch caused by the difference of speakers.

In this study, we use MFCCs for GMM-HMM-based system and fbank coefficients for the DNN-HMM-based system. The MFCCs contains 23 coefficients together with their first and second order temporal differences. The fbank feature contains 40 coefficients (and energy) distributed on a mel-scale. The LDA and fMLLR are also applied.

2.2 Searching

After getting the acoustic feature, the problem left to the ASR system is searching for the most optimal answer: the text which is most probable to generates the input acoustic feature. It can be implemented by dynamic programming, such as the Viterbi algorithm [12]. And pruning techniques are usually used to accelerate the searching. The most important task in this stage is modeling the probability of the text generates the input acoustic feature, $p(acoustic\_feature|text)$. To make the task easier, we decompose it into a series steps, and model the probability by a series individual models, which are acoustic models, lexical models and languages models.

The acoustic model models the relationship between the acoustic features and phonemes. Usually, a GMM-HMM or a DNN-HMM model is used here. The lexical model models the relationship between the phonemes and the words. It is a pronouncing dictionary or some grapheme-to-phoneme models. The language model assigns a probability to a sequence of words. High probability indicates the sequence is more likely to occur in the language. The language model provides context to distinguish between words and phrases that sound similar. The language model is usually implemented in a N-gram fashion which models the probability of a word occurs in a context of $N - 1$ words.

2.3 Acoustic Model Based on GMM-HMM

Acoustic model is the fundamental component of the ASR system. The GMM-HMM-based acoustic model achieves its great success in ASR, since the introduction of the expectation-maximization (EM) algorithm for joint training of GMMs and HMMs. The GMMs give acoustic feature a score which indicates the probability of the acoustic feature is generated by a HMM state. Then the scores are used in HMM to decode the input into phonemes. Despite the success of the GMM-HMM models, GMMs have some shortcomings, especially on the modeling efficiencies. While, a discriminative model, like DNN, can do better.

2.4 Acoustic Model Based on DNN-HMM

Deep Neural Network (DNN) is an artificial neural network with multiple hidden layers between the input and output layers. DNN is a powerful classifier, and has shown its strength in the speech recognition [7], object recognition [13], natural language understanding [14], and so forth.

In the ASR research, DNN is used as an alternative of the GMM. It assigns scores for each acoustic feature to HMM states. Those scores are then used for HMM decoding. A DNN-based acoustic model is shown in Fig. 2. The input of the DNN is a window of frames of real-valued acoustic coefficients. And the output layer is a softmax layer that contains one unit for each possible state of each HMM. The DNN is trained to predict the HMM state corresponding to the central frame of the input window. These targets are obtained by using a baseline GMM-HMM system to produce a forced alignment.

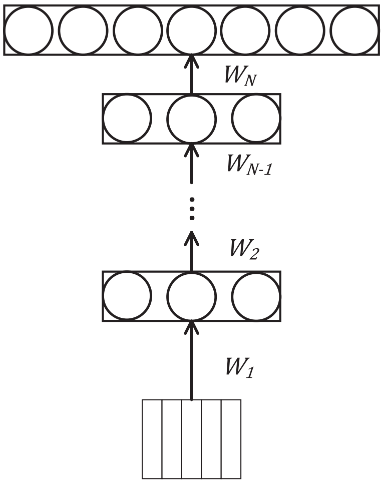

Fig. 2. DNN-based acoustic model.

3 Experiments and Results

We implement the Mongolian LVCSR system based on the Kaldi speech recognition toolkit [15] and use the SRI Language Modeling Toolkit (SRILM) [16] to train the language model.

3.1 Dataset

We build a Mongolian speech corpus which contains about 78 hours recordings. The material includes Mongolian dialogues, Mongolian news and articles from Mongolian text books of junior school. The corpus involves 193 speakers, in which there are 110 male speakers and 83 female speakers. The corpus is divided into training set and test set randomly, where the test set is about 10% of the whole corpus. There are no overlap between the training and test set.

We also build a Mongolian text corpus for language model training. This corpus is a collection of Mongolian web pages, which contains about 85 million tokens.

A pronouncing dictionary is built based on widely used Mongolian dictionaries. It contains about 40,000 items which cover a variety of daily used Mongolian words. And the pronunciation is described with 63 phonemes which include 37 vowels and 26 consonants as listed in Table 1.

Table 1. Mongolian phonemes.

Vowels	**Consonants**
ɑ ə ɪ i ɔ ʊ o u æ œ ʏ ɑː əː ɪː iː ɔː ʊː oː uː æː eː œː ɑ̆ ɛ̆ ɪ̆ ǐ ɔ̆ ŏ ʊɪ ui ʊæ ue ʊɑ y ɿ ʅ ɚ	b p w m s d t l r ʤ ʧ ʃ ɟ ɡ x ŋ f k x ɬ ʥ ʦ dʐ tʂ ʂ ʐ

3.2 Models

We follow an iterative training scheme, and train a series of models. The latter models are trained based on the alignment of the former models. The dependence relationships among the models are illustrated in Fig. 3.

The name of models start with a description of model, then a level index, and end up with a letter to distinguish the models. The configurations of each model are introduced as follows.

`mono1` is a mono-phone model based on GMM-HMM. The feature used here is MFCC with its first and second order temporal difference. Training starts with an equally spaced alignment.

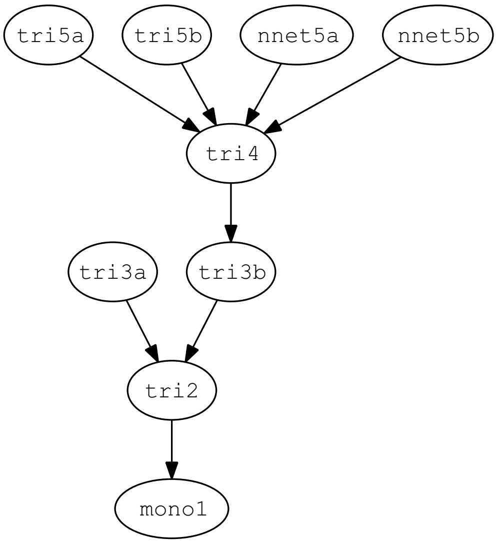

Fig. 3. Dependence relationships among the models. The arrows point to the model, which provides alignment results.

`tri2` is a cross-word tri-phone model based on GMM-HMM. `tri2` is trained based on the alignment of `mono1`. And the feature is same as the `mono1`.

`tri3a` is same as `tri2`, but is trained based on the alignment of `tri2`.

`tri3b` is same as `tri3a`, but apply a speaker adaptation using fMLLR.

`tri4` is same as `tri3b`, but is trained based on the alignment of `tri3b`.

`tri5a` is same as `tri4`, but is trained based on the alignment of `tri4`.

`tri5b` is same as `tri5a`, but used 40-D LDA feature based on the MFCC.

`nnet5a` is a tri-phone model based on DNN-HMM. The model is trained based on the alignment of `tri4`. This model uses 40-D fbank feature with a context window of 9 frames (the preceding 4 frames and the following 4 frames). Then the features are converted into 360-D by LDA. The DNN is composed of 1 tanh hidden layer, which contains 1024 nodes. The output transformation is softmax. There are approximately 3.7 million trainable parameters in the DNN.

`nnet5b` is also a tri-phone model based on DNN-HMM. But the DNN in `nnet5b` is much larger than the one in `nnet5a`. This model uses 40-D fbank feature with a context window of 15 frames (the preceding 7 frames and the following 7 frames). Then the features are converted into 900-D by LDA. The DNN is composed of 6 tanh hidden layers, in which the former 3 layers contain 3762 nodes in each layer and the latter 3 layers contain 1536 nodes in each layer. There are about 50 million trainable parameters in the DNN.

3.3 Evaluation

We employ the Word Error Rate (WER) as the evaluation metric and generate decode results with the 2-gram and 3-gram language model. The results are listed in Table 2.

Table 2. Performances of different models: %WERs.

Model	2-gram	3-gram
`mono1`	69.71	57.65
`tri2`	41.69	29.22
`tri3a`	38.87	27.46
`tri3b`	36.10	25.40
`tri4`	40.22	28.24
`tri5a`	36.32	25.79
`tri5b`	**35.46**	**25.12**
`nnet5a`	19.01	13.82
`nnet5b`	**16.47**	**12.37**

From Table 2, as we expected, we see DNN-HMM-based models outperform the GMM-HMM-based ones. And the 3-gram language model is better than the 2-gram language model. For 2-gram language model, the best DNN-HMM-based model achieves a relative improvement of 53.55% compared with the best GMM-HMM-based model. And for 3-gram language model, the relative improvement is 50.76%.

The improvement comes from three factors. 1) The fbank is better than MFCC when modeling by DNN. 2) The input context of DNN is larger than that in GMM. 3) DNN is more powerful than GMM.

4 Conclusion

In this study, we introduce the DNN-HMM models into the Mongolian speech recognition. The new ASR system achieves significant performance gains as in other languages in previous works [7]. It obtains word correct recognition rate of 87.63% in the test set. As far as we know, the DNN-HMM-based Mongolian ASR system becomes the state-of-the-art one in this field.

5 Future Works

We plan to build a practical Mongolian LVCSR system. This paper is just an initial report of this long term project. There are some works to do in the future.

1) In this study, we only deal with Traditional Mongolian. And Cyrillic Mongolian should also be taken into consideration. Some conversion systems

have been proposed [17, 18] to convert Traditional to Cyrillic Mongolian and vice versa. With those conversion systems, the recognized results can be output in any written systems.

2) Mongolian has a very large vocabulary. Although the pronouncing dictionary used in this study covers the frequently-used Mongolian, the vocabulary should be expanded for a wider range of application. And [5] proposes a segmentation-based method, which seems a promising way to combat the large vocabulary problem.

3) Recurrent Neural Network (RNN) is used more frequently in ASR now [19–22]. We plan to explore its power both in the language model and the acoustic model.

Acknowledgements

This research was supported in part by the China national nature science foundation (No.61263037), Inner Mongolia nature science foundation (No. 2014BS0604) and the program of high-level talents of Inner Mongolia University.

References

1. Lewis, M. Paul, G. F. Simons, and C. D. Fennig, “Ethnologue: Languages of the world, eighteenth edition,” *Dallas, TX: Sil International*, 2015. [Online]. Available: http://www.ethnologue.com
2. G. Gao, Biligetu, Nabuqing, and S. Zhang, “A mongolian speech recognition system based on HMM,” in *Computational Intelligence.* Springer, 2006, pp. 667–676.
3. H. Qilao and G. Gao, “Researching of speech recognition oriented mongolian acoustic model,” in *Pattern Recognition, 2008. CCPR 2008. Chinese Conference on.* IEEE, 2008, pp. 406–411.
4. F. Bao and G. Gao, “Improving of acoustic model for the mongolian speech recognition system,” in *Pattern Recognition, 2009. CCPR 2009. Chinese Conference on.* IEEE, 2009, pp. 616–620.
5. F. Bao, G. Gao, X. Yan, and W. Wang, “Segmentation-based mongolian LVCSR approach,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* IEEE, 2013, pp. 8136–8139.
6. A. Ayush and B. Damdinsuren, “A design and implementation of HMM based mongolian speech recognition system,” in *Strategic Technology (IFOST), 2013 8th International Forum on*, vol. 2, June 2013, pp. 341–344.
7. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.
8. S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.
9. S. Furui, “Cepstral analysis technique for automatic speaker verification,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 2, pp. 254–272, 1981.

10. A.-r. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.* IEEE, 2012, pp. 4273–4276.
11. M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.
12. G. D. Forney Jr, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.
13. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
14. Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," *Innovations in Machine Learning*, pp. 137–186, 2006.
15. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," 2011.
16. A. Stolcke *et al.*, "SRILM-an extensible language modeling toolkit," in *INTERSPEECH*, 2002.
17. F. Bao, G. Gao, X. Yan, and H. Wang, "Language model for Cyrillic Mongolian to Traditional Mongolian conversion," in *Natural Language Processing and Chinese Computing.* Springer, 2013, pp. 13–18.
18. F. Bao, G. Gao, X. Yan, and H. Wei, "Research on conversion approach between Traditional Mongolian and Cyrillic mongolian," *Computer Engineering and Applications*, pp. 206–211, 2014.
19. T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," 2010, pp. 1045–1048.
20. M. Sundermeyer, I. Oparin, J. Gauvain, B. Freiberg, R. Schlüter, and H. Ney, "Comparison of feedforward and recurrent neural network language models," 2013, pp. 8430–8434.
21. A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deepspeech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.
22. W. Chan and I. Lane, "Deep recurrent neural networks for acoustic modelling," *arXiv preprint arXiv:1504.01482*, 2015.