Deep Feature Learning for Tibetan Speech Recognition using Sparse Auto-encoder

H. Wang, Y. Zhao, X.F. Liu, X.N. Xu, L.Wang, N. Zhou & Y.M. Xu
School of Information Engineering, Minzu University of China
China

***Abstract*—HMM models based on MFCC features are widely used by researchers in Tibetan speech recognition. Although the shallow models of HMM are effective, they cannot reflect the speech perceptual mechanism in human being's brain. In this paper, we propose to apply sparse auto-encoder to learn deep features based on MFCC features for speech data. The deep features not only simulate sparse touches signal of the auditory nerve, and are significant to improve speech recognition accuracy with HMM models. Experimental results show that the deep features learned by sparse auto-encoder perform better on Tibetan speech recognition than MFCC features and the deep features learned by MLP.**

Keywords-deep feature learning; sparse auto-encoder; tibetan speech recognition; MFCC features

I. INTRODUCTION

According to the knowledge of auditory phonetics, psycholinguistics and human ear structure, cochlea is equivalent to a filter bank, and the filtering effect of the cochlea is conducted on logarithmic frequency scale. An important part of the cochlea is the basement membrane, above which is the organ of corti. It is the main organ of mechanical to neural transduction. The variation of fluid waves' velocity in the cochlear can result in potential changes on both sides of the hair cell membrane in the organ of corti, which can activate and suppress auditory nerve under certain conditions. Then, the signals of auditory nerve are transmitted into the brainstem for speech recognition [1].

In the research of speech recognition, because MFCC (Mel Frequency Ceptral Coefficient) features represent the human auditory system's response to sound with respect to frequency approximately and reflect human auditory characters, it is widely used as inputs for speech recognition models, most of which are Hidden Markov Model(HMM) models. In Tibetan speech recognition, most researchers also apply this shallow modeling approach, for example, authors in [2,3,4] use 39-dim MFCC to build HMM models of phone and syllable recognition for Lhasa Tibetan. But the shallow models is a simple structure which switch input signals to special issue space[5]. Considering speech perception mechanism in brain, the shallow speech recognition models do not have the capability of simulating speech perception mechanism in brain, the main of which is that it cannot simulate how frequency signal is converted into trigger signal of the auditory nerve. So the accuracy of the speech recognition system based on shallow learning can be improved further.

However, some researchers use neural network to simulate speech recognition procedure[6,7] of brain, subject to gradient diffusion, local optimum of learning and non-sparsity constraint BP algorithm, traditional neural network are not very good at imitating the process of speech recognition.

In recent years, deep learning overcomes training problems of BP algorithm and builds neural network with unsupervised learning and sparsity constrain of hidden neurons[8,9]. Deep learning can reflect the mechanism of the signal processing in brain, and it can learn the deeper and more abstract features which represent the spatial pattern of brain nerve for input data. So, in this work we applied a simple and effective approach, i.e. sparse auto-encoder (SA), to learn deep features based on MFCC features to simulate how speech frequency signal is translated into trigger signal of the auditory nerve. MFCC representing auditory features of ear are converted into sparse auditory nerve signals by sparse auto-encoder, and finally the outputs of sparse auto-encoder are fed into HMM models to complete speech recognition.

The rest of this paper is organized as follows. Section 2 describes speech recognition architecture based on deep features learning. Section 3 introduces sparse auto-encoder method briefly. Section 4 gives the acoustic modeling algorithm of Tibetan speech recognition based on deep features learning. In Section 5, we report and analyze experimental results. Finally, we conclude the paper in Section 6.

II. SPEECH RECOGNITION SYSTEM BASED ON DEEP FEATURES LEARNING

The existing speech recognition systems based on MFCC and HMM model adopt the framework as shown in Figure 1.

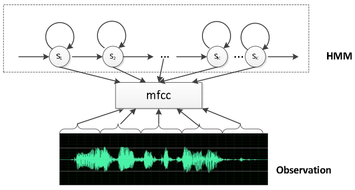
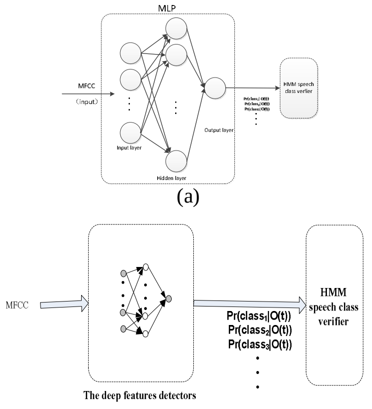


FIGURE I. SPEECH RECOGNITION SYSTEM BASED ON MFCC AND HMM MODEL.

In this system, MFCC features are viewed as input of HMM model, and a separate model for each speech class is built, which belongs to shallow learning and cannot simulate the process of analyzing and explaining input data in brains.

Another kind of speech recognition method uses artificial neural networks and HMM model, as shown in Figure 2(a).



(a)

(b)

FIGURE II. (a) SPEECH RECOGNITION SYSTEM BASED ON MFCC, NEURAL NETWORK AND HMM. (b) SPEECH RECOGNITION SYSTEM BASED ON MFCC, SA AND HMM FOR THE DEEP FEATURES LEARNING.

The system includes the input layer, hidden layers and output layer with only the neighboring layers connected, and its layered structure is similar to the structure of our brain. But traditional neural network training methods adopt back propagation algorithm, which uses iterative algorithm to train the entire network, initialize parameter randomly and calculate the output, then modifies parameters of each layer according to the difference between current output and real value until the system converges. There are major drawbacks of the training methods[10] . Firstly, it is sensitive to the initial value of parameters trapping in local optimum and over-fitting phenomenon easily. Secondly, the residual will become so small while propagating to the top of the layer and gradient diffusion will appear. Thirdly, because there are a large number of neurons in our brains, certain sounds can activate few neurons, so trigger signal of each layer should be sparse. But the BP neural network is not constrained to sparsity.

In this paper, we introduce deep learning to obtain the deep features of the input speech data with greedy layer-wise unsupervised pre-training from the bottom up (cognitive process) and weight tuning from the top down (generation process).Deep learning method avoids falling into local optimum, gradient diffusion and non-sparsity. Deep learning can get generation process and cognitive process to agree to ensure the top layer (output layer) recovering the bottom layer as accurately as possible. In our paper, we adopt a simple and effective method of deep learning, sparse auto-encoder[11], to learn the deep features, simulate perceptive signal of the auditory nerve on speech and gather information into HMM models to recognize speech class.

Figure 2(b) shows the speech recognition system framework based on sparse auto-encoder, where sparse auto-encoder is used as the deep feature detectors, and each speech class verifier is modeled by a three-state left-to-right HMM.

III. SPARSE AUTO-ENCODER

In this section, we will introduce the sparse auto-encoder for deep feature learning. Sparse auto-encoder is an unsupervised learning algorithm that setting the target values to be equal to the inputs. We describe a single-layer sparse auto-encoder shown as Figure 3, which includes input layer, hidden layer and output layer.

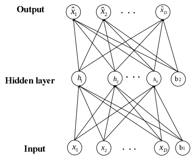


FIGURE III. A ONE-LAYER SPARSE AUTO-ENCODER.

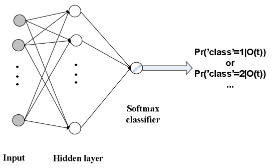


FIGURE IV. A SPARSE AUTO-ENCODER WITH ONE HIDDEN LAYERS AND A SOFTMAX CLASSIFIER LAYER.

The input vector is $\mathbf{x} \in \Re^D$, and hidden variable $\mathbf{h} \in \Re^N$ represents deep features. The mapping relation between the input layer and the hidden layer is as follows:

$$\mathbf{h} = \sigma(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \tag{1}$$

Where $\sigma(t) = (1+e^{-t})^{-1}$ is the sigmoid function, $\mathbf{W}^{(1)} \in \Re^{N \times D}$ is a weight matrix with N hidden units, $\mathbf{b}^{(1)} \in \Re^N$ is an encoding bias. The output $\hat{x}$, i.e. the reconstruction of the input x , is obtained by

$$\hat{\mathbf{x}} = \sigma(\mathbf{W}^{(2)T}\mathbf{h} + \mathbf{b}^{(2)}) \tag{2}$$

where $\mathbf{W}^{(2)} \in \Re^{N \times D}$ is the decoding matrix and $\mathbf{b}^{(2)} \in \Re^D$ is an decoding bias. An " over-complete " nonlinear features (greater than the number of input) are learned by minimizing the reconstruction error of the likelihood function with a sparsity constraint:

$$L(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{2}\sum_{i=1}^{m} \|\hat{\mathbf{X}} - \mathbf{X}\|^2 + \beta \sum_{j=1}^{N} KL(\rho \,\|\, \hat{\rho}_j) \tag{3}$$

where β is the weight of sparsity penalty, ρ is the average activation of $\mathbf{h}$, $\hat{\rho}_j = \frac{1}{m}\sum_{i=1}^{m}[h_j]_i$ is the average activation of over the training data, and the Kullback-Leibler divergence is written

$$KL(\rho \| \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho)\log \frac{1-\rho}{1-\hat{\rho}_j} \quad (4)$$

In our paper, we add a softmax classifier on the top layer in sparse auto-encoder to output a vector of speech class posterior probability, which is treated as observation and is fed into a HMM. The vector of speech class posterior probability maximize the separation between speech classes in the output space. This separation leads to improved discrimination by a GMM model, which describes the output space associated with each HMM state[12]. A sparse auto-encoder with one hidden layers and a softmax classifier layer is shown in Figure 4.The parameters for a sparse auto-encoder are obtained by greedy layer-wise training. This method consists of two phase: pre-training and fine-tuning. In pre-training, we use unlabeled data samples to train the sparse auto-encoder for parameters $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, $\mathbf{b}^{(1)}$, $\mathbf{b}^{(2)}$ with unsupervised methods. After the phase of pre-training is complete, back-propagation can be used to improve the results by tuning the parameters at the same time with supervised manner based on labeled data .

IV. The acoustic modeling algorithm of Tibetan speech recognition based on deep feature learning

The learning algorithm of deep feature learning and speech recognition modeling is summarized as follows.

a. Train a sparse auto-encoder on the MFCC features of input speech data U with unsupervised methods, and then feed the speech data U into trained sparse auto-encoder to obtain the output of the hidden layer as the deep feature $\mathbf{h}^{(1)}$ of input speech data U .

b. Feed the features $\mathbf{h}^{(1)}$ as the input into a softmax classifier and train them to map the speech classes.

c. Fine-tune the sparse auto-encoder with one hidden layer and a softmax classifier layer using labeled speech data U.

d. Output the well-trained sparse auto-encoder.

e. Feed speech data U into the well-trained sparse auto-encoder and get the speech class posterior probabilities, and then we put posterior probabilities as input observation for HMM models, and train the models of speech classes to perform speech recognition.

V. Experimental results

We evaluate the sparse auto-encoder method on extracting the deep features for Tibetan isolated word speech recognition based on MFCC features. In the experiments, we compared the deep feature learned by the single layer sparse auto-encoder with the feature learned by Multi-layer perceptron (MLP) and MFCC features to train HMM models for speech recognition.

In this paper, two data sets were used for evaluation. One is 8 speakers (4 males and 4 females) saying the Tibetan digits 0 to 9 for 4 times, and the other is 10 speakers (5 males and 5 females) saying 34 Tibetan letters for 4 times. The Tibetan letter speech data are clean audio, and digital speech was recorded with background noise for audio-video data.

All original speech data is sampled to 8 KHz. 39 MFCC features of each observation frame were extracted using a 32 ms window with 10 ms overlaps. 39 MFCCs are composed of 12 MFCCs plus the energy coefficient and their first-second-order time derivatives. Sparse auto-encoder and MLP have 39 input layer nodes and only one hidden layer. MLP has 50 hidden layer nodes. Sparse auto-encoder contains 100 nodes in hidden layer, where the weight of sparsity penalty β =3, the mean activation ρ =0.1, and the weight decay parameter for back propagation optimization λ =0.003. MLP and sparse auto-encoder select sigmoid type nodes in the hidden layers. The recognition accuracy on test set using three speech recognition systems, i.e. MFCC and HMM; MFCC, neural network and HMM; MFCC, SA and HMM.

TABLE I. Recognition accuracy rate on three speech recognition systems.

Speech recognition systems	34 phones	10 digits
MFCC and HMM	70.6%	43.75%
MFCC, MLP and HMM	95.6%	50%
MFCC, SA and HMM	100%	52.5%

As we can see from the table, all the models trained on the clean audio data (34 phones) have higher recognition accuracy than the models trained on audio data extracted from audio-video data. Because we should take into account that noise of recorded audio-video files is stronger than that of audio files.

In the table, new features obtained by MLP based on MFCC features have better performance on speech recognition than MFCC features. We observed that recognition accuracy increases by 25% on 34 phones data sets and 6.25% on 10 Tibetan digits. We can also see from the table that recognition accuracy on the deep features learned by sparse auto-encoder is higher than that by MLP. The experiments suggests that compared with MFCC features, neural network simulates the process of brains ’ speech recognition and improves recognition accuracy. However, traditional neural networks like MLP have many defects such as gradient diffusion, non-sparsity and so on. In this paper, we apply sparse auto-encoder based on MFCC features to learn features deeply. The features extracted represent the sparse trigger signal of the auditory nerve more approximately and actually improve speech recognition accuracy for HMM models.

VI. Conclusion

In this paper, we use sparse auto-encoder to extract the deep features for Tibetan speech recognition. The experimental results showed that the features extracted by deep learning method can better simulate the process of transforming frequency signals to sparse trigger signal of the auditory nerve and have a better performance on speech recognition than MFCC features and the features extracted by MLP.

In further work, we will use a large amount of Tibetan continuous speech data to evaluate the performance of the proposed deep feature learning method for speech recognition.

ACKNOWLEDGEMENTS

Fund for this paper was provided by MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No.12YJA630123). Research in this paper is also supported by the first-class university and first-rate discipline funds of Minzu University of China.

REFERENCES

[1] Han Jiqing, Zhang Lei, Zheng Tieran. Speech Signal Processing[M]. Tsinghua University press.2004.11-12,24-25.

[2] Pei Chunbao. Speech recognition technology for Lhasa Tibetan[D]. Tibet University. 2009.K. Elissa, “Title of paper if known,” unpublished.

[3] Meng, M. E. N. G. Research on Acoustic Model of Large-vocabulary Continuous Speech Recognition for Lhasa Tibetan, computer engineering,38, no. 5 (2012): 189-191.

[4] Xu hui. Lhasa voice dialing technology.based on hidden Markov model[D]. Northwest University for Nationalities. 2013.

[5] Morgan N, Bourlard H. Continuous speech recognition[J]. Signal Processing Magazine, IEEE, 1995, 12(3): 24-42.

[6] Dede G, Sazlı M H. Speech recognition with artificial neural networks[J]. Digital Signal Processing, 2010, 20(3): 763-768.

[7] Fu G. A Novel Isolated Speech Recognition Method Based on Neural Network[C]//Proceedings of the International Conference on Information Engineering and Applications (IEA) 2012. Springer London, 2013: 429-436.

[8] Ahmadi S, Ahadi S M, Cranen B, et al. Sparse coding of the modulation spectrum for noise-robust automatic speech recognition[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2014, 2014(1): 1-20.

[9] O’Donnell F, Triefenbach F, Martens J P, et al. Effects of architecture choices on sparse coding in speech recognition[M]//Artificial Neural Networks and Machine Learning–ICANN 2012. Springer Berlin Heidelberg, 2012: 629-636.

[10] SUN Zhi-jun, XUE Lei, XU Yang-ming, WANG Zheng. Overview of deep learning[J]. Application Research of Computers.2012,08:2806-2810.

[11] H. C, Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, “Stacked Auto-encoders for Unsupervised Feature Learning and Multiple Organ Detection in a Pilot Study Using 4D Patient Data,” IEEE Transaction Pattern Analysis and Machine Intelligence, Volume 35, Issue 8, pp.1930-1943, August 2012.

[12] J. Frankel, Ö. Çetin, and N. Morgan, “Transfer Learning for Tandem ASR Feature Extraction,” Machine Learning for Multimodal Interaction, Springer, pp. 227-236, 2008.