Tibetan acoustic model research based on TDNN

Jinghao Yan* Hongzhi Yu† Guanyu Li‡
* Northwest Minzu University, Lanzhou, China
E-mail: yjh527a@163.com Tel/Fax: +86-18067528276
† Northwest Minzu University, Lanzhou, China
E-mail: yhz1947@163.com
‡ Northwest Minzu University, Lanzhou, China
E-mail: xxlgy@xbmz.edu.cn

***Abstract*—Deep neural network (DNN) has been significantly improved in Tibetan speech recognition tasks, however, it still requires improvement when compared with that in Mandarin, English, or other languages. This paper examines a Tibetan acoustic model based on deep neural network and extracts the i-Vector features by modeling the speaker in the feature space. After combining the MFCCs and i-Vector features, we train a time-delayed neural network (TDNN) based Tibetan acoustic model, compared to deep neural network, it can get better performance. At the same time, we study the transfer learning from Mandarin to Tibetan and prove its effectiveness.**

I. Introduction

In recent years, end-to-end speech recognition technology has become research hotspot. Traditional HMM based methods require separate components and separate training of acoustic models and language models. The end-to-end model jointly trains these components and reduces the assumption of traditional methods. Among them, Alex Graves, Navdeep Jaitly proposed a CTC-based model[1] in 2014, and in 2016, chan and Bahdanaua et al. introduced an attention-based model[2, 3]. But this technology requires a lot of data to get better results. Therefore, the end-to-end technology is not obvious for the speech recognition in low resource environment, so we used the traditional HMM based methods.

In speech recognition, a deep learning based acoustic model replaces the traditional Gaussian mixture model and becomes the mainstream, the error rate is greatly reduced[4]. However, the study on deep learning methods in Tibetan speech recognition tasks is scarce. Although it is acknowledged that Tibetan acoustic model based on deep neural network structure has been greatly improved, this structure can hardly be used in modeling longer context. Then some deep learning models that can be able to identify latent temporal dependants are used in speech recognition(LSTM RNN, TDNN). These models can achieve lower error rates in speech recognition than DNN models. Time-Delay Neural Network[5] was proposed by Hinton in 1989, and is mainly used in solving the problem that the traditional method HMM in speech recognition cannot adapt to dynamic time-domain changes in speech signals. It extends the output of each hidden layer in the time domain, that is, the input received by each hidden layer is not only the output of the previous layer at the current time, but also that before and after it, so longer context information can be modeled. Therefore, this paper uses the TDNN model for acoustic model modeling.

Speaker adaptation techniques have been proved to be effective in deep neural networks, but speaker adaptation techniques like fmllr[6] require two decoding stages, making it difficult to use for online speech recognition applications. Despite from this, the i-Vector with environmental information and speakers information is useful for the neural network's instantaneous and discriminative adaptation[7]. We study the effect of i-Vector in the TDNN model. Data augmentation is a common strategy adopted in increasing the quantity of training data, avoiding overfitting and improving robustness of the models. Vocal tract length perturbation(VTLP) has shown gains on TIMIT phoneme recognition task[8]. But Tom Ko et al. showed speed perturbation which emulates both VTLP and temp perturbation, it can give more word error rate(WER) improvement than either of those methods[9].

Transfer learning is a machine learning method that aims to develop a better system for a new task in a quick way. There is a rich survey of transfer learning in the literature[10,11,12]. Generally speaking, transfer learning involves all methods that utilize some auxiliary resources such as data, model, labels and so on. Dong Wang et al. gives a comprehensive survey of methods in speech recognition[13].

Tibetan is an alphabetic writing developed on the basis of Sanskrit, and contains 30 consonants and 4 vowels. Its character stream is a two-dimensional one and it is written horizontally from left to right. And it's listed in the Tibetan branch of the Tibetan-Burmese language group of the Sino-Tibetan language family and is mainly divided into three major dialects (Ü-Tsang, Kham, Amdo). In this work, we investigated the speech recognition of the Ü-Tsang dialect which has Lhasa accent.

This paper integrates the TDNN into the acoustic modeling of Tibet dialect speech recognition task. Because the Ü-Tsang dialect has tones, pitch features are extracted to combine with the mfcc features, at the same time, the i-Vector features are introduced to help with excavating more voice information. The experimental results prove that the best word error rate for Tibet dialect speech recognition, which is based on TDNN. Finally, we investigate the effect of transfer learning from Mandarin to Tibetan.

The paper is organized as follows. Section 2 mentions

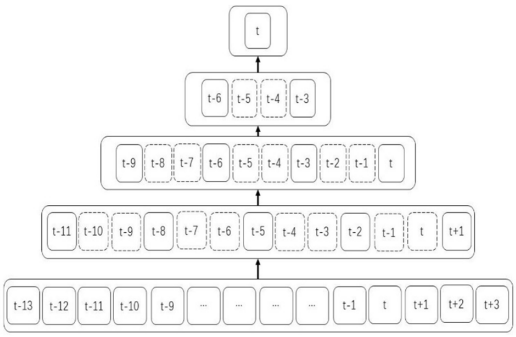

Fig. 1. Computation in TDNN with sub-sampling

relevant works and the neural network architecture. Section 3 describes the experimental setup. Section 4 presents the experimental results and analyzes the experimental results. Section 6 presents the conclusions.

II. Related work

A. Model structure

Time delay neural network[5] is a multi-layer artificial neural network structure whose purposes are to: 1) classify the patterns of mobile invariance; 2) context modeling based on each layer in the network. DNN usually obtains a certain context by splicing adjacent frames, but longer-term context cant be reached. Compared with DNN, TDNN is more capable in learning long context as well as in analyzing its relationship among layers than DNN. It learns short context in the first layer, the bigger the layer number is, the more context it will learn, thus, a convolutional neural network starts to come into being.

A sub-sampling based TDNN model is applied in this article[14], shown as in Figure 1. Partitioned frames can be used in TDNN hidden layer, and a classic TDNN model will calculate the hidden layer excitation for all adjacent frames. However, there are many overlapping contexts in the hidden layer input, which may increase the amount of calculation when training, in order to solve this, sub-sampling method is used to reduce the connecting number by abandoning adjacent frames in the hidden layer, thus to reduce the amount of computation in the training process. The solid line in Fig.1 is a TDNN model using sub-sampling.

The model input layer uses {-2,-1,0,1,2} five frames splicing, and the hidden layers use {-2,1}, {-3,3}, {-6,-3} for splicing, hidden layer usually only spliced two frames, thus the overlap of the underlying context are removed. Traditional neural network models usually use nonlinear activation functions such as Tanh and Sigmoid functions.It was observed that using a linear activation unit, ReLU, can improve its capability on neural networks. This paper uses ReLU activation functions.

B. Speaker adaptation and Audio augmentation

I-Vector is an information-rich low-dimensional fixed length vector extracted from the feature sequence[7]. Because i-Vector represents information about both speaker and acoustic environment of the corresponding segment and therefore, this technique effectively adapts speech recognition system to both speaker and acoustic environment[15]. In order to make the mean-offset information to be encoded in the i-Vector, we estimate the i-Vector on top of features that have not been mean-normalized. At the same time, in order to ensure that there is sufficient and diverse i-Vector in the training data, rather than a separate i-Vector per speaker that we extract through the online form. At the same time, in order to save the calculation, we extracted every 10 frames instead of extracting a single i-Vector in one sentence.

It is necessary to enhance the model to make it more stable when encountering different data disturbances. The article [9] shows that the trained model, which use speed perturbation for data training, achieves a relative improvement of 4.3% on the LVCSR task. It was observed that volume perturbation can achieve better performance than by using only speed perturbation. In order to implement speed perturbation and volume perturbation, we resample the signal using the speed function of the Sox audio manipulation tool. We create two data with speed of 0.9 and 1.1 based on the original training data through the Sox tool. Due to the increased amount of data, we use GMM model to align these data.

C. Transfer learning

Transfer learning methods can be summarized as case learning, feature representation based transfer, parameter based transfer, and correlation based transfer[10]. The parameter transfer method assumes that the model on the relevant task should share some parameters, prior distribution or hyper-parameters, and the structure of the deep neural network is very suitable for parameter transfer[16]. In the framework of deep neural networks, transfer learning can be expressed as how to represent the knowledge to be transfer in deep neural networks and how to use knowledge of other languages. This paper use the method of weight transfer to train a TDNN model of a source language, then transfer the hidden layers to the TDNN model of the target language, and add two newly layers, finally retrained all layers. We choose Mandarin as the source language for transfer learning, because Mandarin and Tibetan are in the same language family. The weight transfer can be represented by Figure 2.

III. Dataset and setup

A. Dataset

This experiment is based on the Tibetan-language Lhasa dataset built by Northwest Minzu University, the dataset is 33 hours' long with the sampling rate of 16K, it contains 31 hours' train corpus and 2 hours' test corpus. The lexicon contains 5933 syllables, basically containing all common syllable types. The language model is a 3-gram model which trained use train corpus text. We use 178 hours' AISHELL

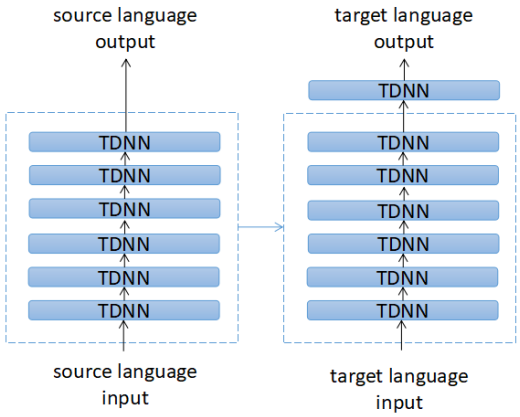

Fig. 2. Weight transfer

Open Source Mandarin Speech Corpus as the source language train corpus of transfer learning.

B. Setup

First, we extract 13 dimensions mfcc coefficient, then use 39 dimensions mfcc feature, which have done twice difference on the 13 dimensions mfcc coefficient and have been normalized by CMVN, to train a monophone GMM, then, we use the monophone model to align the training data and use this data to train a tri-phone model. According to the process, we used LDA and MLLT feature to train the tri-phone model, and use fmllr transform to make the speaker adaptation, thus to get a best GMM baseline system to align the input feature of the neural network.

For the DNN-HMM baseline system, the input features are 40 dimensions Fbank features which have been normalized by cepstral mean variance normalization(CMVN). By retaining 13-dimensional mfcc features and doing twice difference. Compared with mfcc features, Fbank does not need to be discrete cosine transformed to remove their correlations and keep more information. Finally, the adjacent frame and the current frame are stitched together as input to the DNN model, thus to contain certain context information. In this experiment, the preceding 5 frames and the following 5 frames (11 frames in total) work as the input to the DNN model, finally, update parameters by using stochastic gradient descent method, with cross-entropy criteria. There are 6 hidden layers in the model with 1024 neurons per layer in it, and the learning rate is 0.008.

For the TDNN-HMM baseline system, First, we extract 40 dimensions high resolution mfcc feature(mfcc-hires), then splice the adjacent four frames and the current frame to obtain the 200 dimensions input feature, finally, we use the method of LDA affine to transform it into 450 dimensions feature and put it into TDNN model. For the TDNN model, we set 6 hidden layers, and use the sub-sampling technology to splice frames according to {0}, {-1, 1}, {-1, 1}, {-1, 1 }, {-3, 3}, {-6,-3}, {0} on hidden layers. {0} represents a full join. At the same time, we compare the performance of TDNN models from 450 to 1280 different neuron settings.

We use the speaker adaptation and audio augmentation technology to enhance model performance and make it robust. For speaker adaptation, we extract 100 dimensions i-Vector feature, then splice it with mfcc-hires feature. For audio augmentation, we use speed perturbation to create 0.9, 1.0, and 1.1 times the rate of data, then use volume perturbation on training data where each recording is scaled with a random variable drawn from a uniform distribution over [0.125,2]. Finally, we re-extract mfcc-hires feature.

In this paper, greedy layer-wise supervised training is performed by using preconditioned stochastic gradient descent (SGD) technology. The TDNN model uses four GPUs for simultaneous training. In the contrast, the DNN model uses only one single GPU for training. In the training process, cross-entropy criterion is used.

IV. Experiments and Results

A. Baseline system

TABLE I
BASELINE SYSTEMS

Model	Hidden layers	Hidden neurons	Wer(%)	Params
DNN	6	1024	25.71	8.7M
TDNN-A	6	450	24.72	4.0M
TDNN-B	6	650	24.55	7.5M
TDNN-C	6	850	24.52	12.0M
TDNN-D	6	1024	24.55	16.8M
TDNN-E	6	1280	24.65	25.3M

Table.1 shows speech recognition results of Tibetan used TDNN and DNN as acoustic models. Compared to DNN model, the recognition error rate of TDNN-B model as an acoustic mode reduces by 4.5%. The reason is that TDNN model can model longer context information. At the same time, TDNN also performs stitching frames in the hidden layer, therefore, the TDNN model parameters more than the DNN model when they have same model setup. TDNN-A to TDNN-E in the experiment show that when the number of hidden layer neurons is small, the model is easy to be underfitting, and on the contrary, it is easy to be overfitting.

B. Speaker adaptation and audio augmentation

In this work, we used speaker adaptation and audio augmentation to improve model performance, and studied the effect of pitch features on Tibetan.

Table.2 shows the results of the experiment on TDNN-C. The only use of both technologies can improve the performance of the acoustic model, and the two technologies can be used simultaneously for better performance. This is due to the enhanced data and the introduction of speaker related information. At the same time, experimental results show that adding pitch features can improve the performance of the model in a variety of situations. This is because the Ü-Tsang dialect is a tonal language, and the extraction of the pitch

TABLE II
SPEAKER ADAPTATION AND AUDIO AUGMENTATION

Model	Pitch	Wer(%)
TDNN-C	N	25.71
TDNN-C + sp + vp	N	23.83
TDNN-C + i-vector	N	23.77
TDNN-C + sp + vp + i-vector	N	22.94
TDNN-C	Y	23.95
TDNN-C + sp + vp	Y	23.34
TDNN-C + i-vector	Y	23.63
TDNN-C + sp + vp + i-vector	Y	**22.92**

feature introduces prosody information. Finally, the best WER is 22.92%.

C. Weight transfer

In this work, we finally investigate the effect of transfer learning on Tibetan. First, we train a Mandarin TDNN model, then we transfer hidden layer into Tibetan TDNN model and add two newly layers, finally, we use one-third of the learning rate re-train the layers.

TABLE III
WEIGHT TRANSFER

Model	Number of transferred layers	Wer(%)
TDNN-F	0	22.92
TDNN-F	3	22.96
TDNN-F	4	22.91
TDNN-F	5	**22.58**

* TDNN-F : TDNN-C + sp + vp + i-vector

In this experiment, we transfer a different number of hidden layers and train them using the same parameters and methods. And the experimental results show that transferring more hidden layers can achieve better model performance. This may because both the source language data and the Tibetan data are in the form of aloud reading, and the data is relatively clean. At the same time, the experimental results also show that the transfer from Mandarin to Tibetan is effective. This is because Mandarin and Tibetan are in the same language family, and there are many similarities in the Ü-Tsang dialect and Mandarin. For example, both Ü-Tsang dialect and Mandarin are tonal languages, they have the same retroflex(zh, ch, sh, and r) and ancient voiced stops and voiced affricate have evolved into unvoiced sounds. Finally, the best WER is 22.58%.

V. CONCLUSIONS

The acoustic model based on TDNN has achieved better results in the task of Ü-Tsang dialect speech recognition. In a low resource data environment, performance can be improved through speaker adaptation and speed perturbation. At the same time, using trained acoustic models to perform transfer learning has also achieved some improvement, which demonstrates that Mandarin to Tibetan language transfer learning is effective. In subsequent studies, we will increase Tibetan data to train a better TDNN-LSTM hybrid model and use multiple languages to train an initial model for transfer learning simultaneously.

ACKNOWLEDGMENT

Supported by the Natural Science Foundation of China under Project (Grant No. 61633013), the Graduate Research(Practice) Innovation Project of Northwest Minzu University, China(Grant No. yxm2016128). And Fundamental Research Funds for the Central Universities(31920170145).

The authors would like to thank all the people who have helped us during the studying process of this work. And give our thanks to the organization of this conference giving us an opportunity to communicate with other craft brother.

REFERENCES

[1] Graves, A., & Jaitly, N. (2014, January). Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning (pp. 1764-1772).

[2] Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on (pp. 4960-4964). IEEE.

[3] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016, March). End-to-end attention-based large vocabulary speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on (pp. 4945-4949). IEEE.

[4] Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.

[5] Waibel, A., Hanazawa, T., Hinton, G. E., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(3), 328-339.

[6] Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., & Visweswariah, K. (2008). Boosted MMI for model and feature-space discriminative training. international conference on acoustics, speech, and signal processing.

[7] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. Academic Press, Inc.

[8] N. Jaitly and G. E. Hinton, Vocal tract length perturbation (VTLP) improves speech recognition, in International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech, and Language Processing, 2013.

[9] Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In Sixteenth Annual Conference of the International Speech Communication Association.

[10] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

[11] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, Transfer learning using computational intelligence: a survey, Knowledge-Based Systems, vol. 80,pp. 14C23, 2015.

[12] Y. Bengio et al., Deep learning of representations for unsupervised and transfer learning. ICML Unsupervised and Transfer Learning, vol. 27, pp. 17C36, 2012.

[13] Wang, D., & Zheng, T. F. (2015). Transfer learning for speech and language processing.

[14] Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts.. conference of the international speech communication association.

[15] Karafiat, M., Burget, L., Matejka, P., Glembek, O., & Cernocky, J. (2011). iVector-based discriminative adaptation for automatic speech recognition. ieee automatic speech recognition and understanding workshop.

[16] Huang, J. T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013, May). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 7304-7308). IEEE.