The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Low-Resource Tibetan Dialect Acoustic Modeling Based on Transfer Learning

Jinghao Yan[1], *Zhiqiang Lv*[2], *Shen Huang*[3], *Hongzhi Yu*[4]

[14]Northwest Minzu University, Lanzhou, China
[23]Tencent Research, Beijing, China

544822903@qq.com, {zhiqianglv, springhuang}@tencent.com, yhz1947@163.com

Abstract

Deep neural network (DNN) based acoustic model has made great breakthroughs in speech recognition. However, low-resource Sino-Tibetan languages such as Tibetan still need further studies, especially when dealing with dialects. Based on a TDNN acoustic model trained according to lattice-free M-MI criteria, this paper demonstrates baseline systems for two Tibetan dialects: Ü-Tsang and Amdo. Transfer learning is also employed to improve our systems. Experiment results show that for low resource Tibetan dialect recognition, transfer learning can consistently outperform the baseline.

Index Terms: Tibetan, acoustic model, TDNN, transfer learning, low-resource

1. Introduction

Earlier speech recognition techniques mainly studied speaker-dependent speech recognition, isolated word recognition, and keyword recognition, using methods such as dynamic time warping (DTW) [1] and linear predictive coding (LPC). As the research progressed further, the proposed acoustic model of GMM-HMM promoted the development of speech recognition technology and became the mainstream for a long time, but the effect is not ideal in a real environment.Until Hinton et al. proposed deep learning in 2006 [2], people began to use deep learning in the field of speech recognition. After the DNN-HMM model framework was proposed, the error rate of speech recognition has significantly reduced compared with the traditional acoustic model[3], and it have already been used in many practical scenes. However, the DNN model lacks the ability to model time dependencies and it is difficult to further improve the performance of speech recognition tasks. Then some deep learning models that can be able to identify latent temporal dependants are used in speech recognition tasks (LSTM RNN, TDNN). These models can achieve lower error rates in speech recognition than DNN models.

In recent years, end-to-end speech recognition technology has became a 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. The end-to-end model approach greatly simplifies the process of establishing a speech recognition systems. Among them, Alex Graves and Navdeep Jaitly proposed a CTC-based model[4] in 2014. In 2016, Chan and Bahdanaua et al. introduced an attention-based model[5, 6]. However, the end-to-end model training is more difficult and requires a large amount of data to achieve good results. For a low-resource language, end-to-end methods show no advantage. Traditional H-MM based structure can achieve more stable results in the low-resource language set up.

In the field of speech recognition, cross-entropy criterion(CE) is usually used to minimize the expected frame error in frame-by-frame training, but speech recognition is essentially a sequence classification problem. Then discriminative training criteria such as MMI, MPE, and MBR are proposed and proven to significantly improve the performance of speech recognition. Furthermore, the sequence discriminative criterion pays more attention to the classification of the sequence itself than the cross entropy criterion. In the actual application process, it is usually necessary to use the CE criterion to train first, and then use the discriminative training criterion to retrain after the lattice is generated, resulting in very long training time. In 2016, Dan Povey et al. proposed a lattice-free MMI method[7] that can greatly speed up the training of discriminative models and avoid the problem of generating word lattices.

Transfer learning is a machine learning method that aims to develop a better system for a new task in a quick and efficient way, which mainly maintain and utilize knowledge learned from one or more similar tasks[8]. In many real world environments, it is often difficult to obtain sufficient matching data for a new task. In this case, transfer learning becomes very important. In speech recognition, the study of transfer learning focuses on multilingual and cross-language studies[9], multi-task training of DNNs, and robust speech recognition using both audio and video information. Due to the more abstract and invariable features represented by the neural network hidden layers, neural networks are very suitable for transfer learning.

Tibetan belongs to the Tibetan branch of Tibetan-Burmese language group of the Sino-Tibetan language. Tibetan is an alphabetic writing developed on the basis of Sanskrit, and containing 30 consonants and 4 vowels. There are three major dialects in Tibetan (Ü-Tsang, Amdo, Kham). Tibetan is a unified written language, but so far no ethnically recognized oral standard has been formed. Among the three dialects, the difference between Ü-Tsang dialect and Amdo dialect is the largest, and it is worth studying in depth. In the past, Tibetan speech recognition studies are all based on the modeling of individual dialect and did not make use of resources in other rich languages. Because there are few data resources in Tibetan, it is necessary to study transfer learning from Mandarin to Tibetan. By studying the transfer between Tibetan dialects, we can further promote the performance of Tibetan speech recognition.

In this paper, the TDNN-LSTM baseline system is established separately for the Ü-Tsang dialect with tones and the Amdo dialect without tones, and the lattice-free MMI criterion is used for training. Then we use the Mandarin TDNN-LSTM model as the initial model to study the effect of Mandarin transfer on these two Tibetan dialects. Finally, in the two Tibetan dialects, a model transfer experiment is conducted, that is, a dialect trained model is used as the other dialect initial model to study the internal migration effects.

The paper is organized as follows. Section 2 mentions relevant work and the neural network architecture, Section 3 de-

10.21437/SLTU.2018-2

scribes the experimental setup. Section 4 presents the experimental results. Section 5 analyzes the experimental results. Section 6 presents the conclusions.

2. Related work

2.1. Model structure

The TDNN model is a deep neural network model proposed by Hinton in 1989[10]. It is considered to be the predecessor of convolutional neural networks. Compared to DNN(deep neural networks), it can model longer context information. As hidden layers increase, the context information per layer will increase. The classic TDNN model adopts a full connection method, which leads to more overlapping context information contained in the upper layer, and it results in greater redundancy. Vijayaditya Peddinti et al. show that using sub-sampling can greatly reduce the speed of model training under the premise of guaranteeing performance[11]. Through this technique, TDNN can use cut-off frames, thus reduce the number of overlapping contexts contained in the hidden layer input.

The RNN(recurrent neural network) model is a neural network which can model sequence data, that is, the current output of a sequence is also related to the previous output. The specific form of expression is that the network will memorize previous information and apply it to the calculation of the current output. However, the effect is not ideal in practical use because the vanishing gradient and exploding gradient problem in RNN structures. The LSTM model[12,13] is an improved model based on the RNN model which solves the gradient related problem. It introduces the concept of a memory cell and controls the information in the cell by adding three gates. The forget gate determines what old information we want to delete from the cell state, the input gate determines what new information we want to add in the cell state, and finally the control gate controls what information should be output. To reduce model calculations, we use the LSTM with recurrent projection layer structure[14], which adds a projection layer to the original LSTM and connects this layer to the LSTM input. This paper uses a hybrid model of TDNN-LSTM. The model structure is shown in Figure 1.

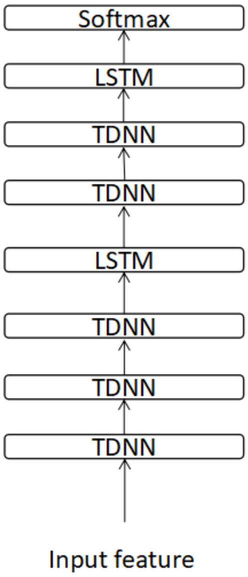

Figure 1: *TDNN-LSTM structure*

2.2. Lattice-free MMI

Povey et al. introduced the idea of CTC based on MMI and proposed lattice-free MMI criterion[7]. The lattice-free MMI criterion computes all possible annotation sequences at the output layer of the neural network, then computes the corresponding MMI information and associated gradients, and finally completes the training by the gradient propagation algorithm. Because the lattice-free MMI training criterion can directly calculate the posterior probability of all possible paths during the training process, it eliminates the needs to generate lattices in advance before discriminative training and reduces the time. At the same time, using the lattice-free MMI criterion for training, there is no need to use the CE criterion in advance. we use three regularization methods(cross-entropy regularization, output l_2-norm regularization, and leaky HMM) to prevent overfitting and reduce the decodding frame rate to one-third and use a simpler HMM structure.

2.3. Transfer learning

According to learning methods, transfer learning can be divided into sample-based transfer, feature-based transfer, model-based transfer, and relationship-based transfer. In the framework of deep neural networks, transfer learning can be expressed as how to represent the knowledge to be transferred in deep neural networks and how to use knowledge of other languages[15,16]. We study the model based transfer method, which uses model parameters to represent knowledge to be transferred. We transfer the knowledge of the target language to the two Tibetan dialect models by sharing the hidden layer of neural network. In this paper, the model-based transfer is used to transfer the hidden layers of the neural network trained in the original domain language to the Tibetan language model. That is, the Mandarin TDNN model is used as the initial model, the softmax layer is replaced with the Tibetan softmax layer, and then the data is retrained using the data of the Amdo dialect and the Ü-Tsang dialect. The transfer between dialects uses the other dialect trained model as the initial model, and then also transfer the hidden layer to the target dialect. Figure 2 shows a deep neural network model transfer method based on shared hidden layers.

3. Dataset and setup

3.1. Dataset

The Amdo dialect data set includes a 58 hours' training corpus, a 3 hours' development corpus(dev-Amdo), and a 1 hours' test corpus(test-Amdo). The Ü-Tsang dialect data set includes a 44 hours' training corpus, a 3 hours' development corpus(dev-Ü-Tsang), and a 5 hours' test corpus(test-Ü-Tsang). The language model is a 4-gram model. Modeling units for acoustic models and language models are syllables. We use a Ü-Tsang Lexicon with 5k Ü-Tsang syllables for two dialects.

3.2. Experimental setup

This paper use the speed-perturbation technology to enhance the model and create 0.9, 1.0, and 1.1 times the rate of data,

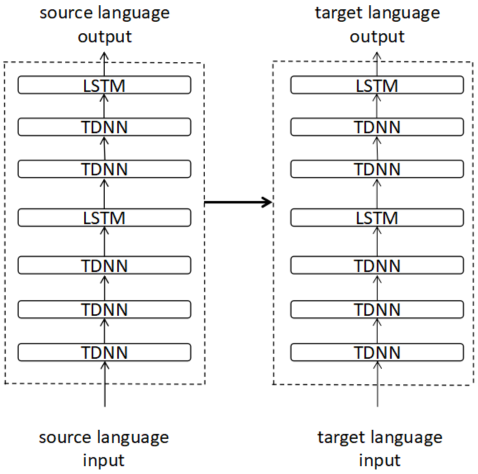

Figure 2: *Model transfer*

making the model more stable to encounter different data disturbances. The article [17] shows that after using speed perturbation for training data, the trained model achieved a relative WER improvement of 4.3% on the LVCSR task. When use volume perturbation technology to do volume perturbation, the volume perturbation technology can achieve a 1.5% WER improvement on the test set compared to that uses only the speed perturbation technology. Through two enhancement techniques, training data has been expanded.

We need to transfer recorded audio files into features. The mfcc feature can be used to describe the non-linear characteristics of human's ears, which is widely used in the field of speech recognition. For the input feature of the GMM model, we first extract the 13 dimensional mfcc coefficients, then obtain the 39 dimensional mfcc feature through two differential operations, and finally perform the cepstral mean and variance normalization on the 39 dimensional mfcc feature. For the input features of the TDNN model, we first use the speed-perturbation to amplify the data, and then extract the 40 dimensional high-resolution mfcc feature. Since both Mandarin and Ü-Tsang dialect have tones, we extract the 3 dimensional pitch features after the 40 dimensional mfcc feature. Finally, we extract the 100 dimensional i-Vector feature with speaker information and environmental information and send it to the TDNN model at the same time. In this way, the 40 dimensional mfcc feature does not need to do cepstral mean and variance normalization.

We use the 39 dimensional mfcc feature to train the monophone based GMM model first. Then, we use alignments from the monophone GMM model to train a triphone based GMM model. Finally, a good SAT based GMM model is trained to provide alignments for the TDNN model. The TDNN model sets 6 hidden layers which use sub-sampling techniques, and the hidden layer splicing configuration is {0}{-1,0,1}{-3,0,3}{-7,0,2}{-3, 0, 3}{0},({0} means no splicing). The TDNN-LSTMP model sets up 7 TDNN layers and 3 LSTMP layers. The training criterion is the lattice-free MMI.

4. Experiments and results

4.1. Baseline

Table 1 below shows the baseline system results for the Amdo dialect and the Ü-Tsang dialect. In the Amdo dialect model, the TDNN and LSTMP hybrid model is used and a relative 5.9% reduction in WER can be achieved in test-Amdo corpus compare with the TDNN model. In the Ü-Tsang dialect model, the hybrid model can achieve a relative 16% WER decrease in test-Ü-Tsang corpus compared with the TDNN model. From the results, we can see that the TDNN-LSTMP model trained with lattice free MMI can reduce the WER in the test set. The experimental results of Amdo dialect show that the lexicon of Ü-Tsang dialect can also achieve good results in Amdo dialect.

Table 1: *Amdo baseline system*

Model	dev-Amdo(WER)	test-Amdo(WER)
TDNN	25.30%	59.33%
TDNN-LSTMP	25.54%	**55.81%**

Table 2: *Ü-Tsang baseline system*

Model	dev-Ü-Tsang(WER)	test-Ü-Tsang(WER)
TDNN	24.45%	48.98%
TDNN-LSTMP	26.66%	**40.85%**

4.2. Mandarin to Tibetan transfer

First, we use a Mandarin data set of 1700 hours to train a TDNN-LSTMP hybrid model. Then we remove the softmax layer of the model and create a new softmax layer of Tibetan language to connect the hidden layer. One-third of the initial learning rate is adjusted, and the final learning rate remains unchanged. Then we retrain the model by using the data of Amdo dialect and the Ü-Tsang dialect. The results are shown in Table 3 and Table 4.

Table 3: *Mandarin to Amdo transfer*

Model	dev-Amdo(WER)	test-Amdo(WER)
TDNN-LSTMP	25.54%	55.81%
TDNN-LSTMP(M-A)	24.28%	56.00%

In the Amdo dialect model which use Mandarin model transfer, the experimental results in the dev-Amdo corpus achieved a relative 4.9% WER decrease compared to the Amdo baseline system, and there is no significant change in WER on the test-Amdo corpus. In the Ü-Tsang dialect model which use Mandarin model transfer, the experimental results showed a relative decrease of 7.7% in the WER compared to the Ü-Tsang baseline system on the dev-Ü-Tsang corpus, and a relative 16% decrease in the WER on the test-Ü-Tsang corpus. The results show that the model transfer from Mandarin to Ü-Tsang dialect can significantly improve the performance. However, the model transfer from Mandarin to the Amdo dialect doesn't show obvious effects, and only little improvement is made in the dev-Amdo set.

4.3. Tibetan dialect transfer

Finally, we study the effect of using transfer learning between the two dialects. The first is transferring the Ü-Tsang dialect to the Amdo dialect. By using the baseline TDNN-LSTMP model

Table 4: *Mandarin to Ü-Tsang transfer*

Model	dev-Ü-Tsang(WER)	test-Ü-Tsang(WER)
TDNN-LSTMP	26.66%	40.85%
TDNN-LSTMP(M-Ü)	24.60%	**34.29%**

of the Ü-Tsang dialect as an initial model, the softmax layer of the Ü-Tsang dialect is removed, the softmax layer of the Amdo dialect is newly established and connected, finally, we use the training corpus of the Amdo dialect to train the new model. In the same way, the initial learning rate is adjusted to one third, and the termination of learning rate remains unchanged. The steps for the transfer of the Amdo dialect to the Ü-Tsang dialect are the same.

Table 5: *Ü-Tsang to Amdo transfer*

Model	dev-Amdo(WER)	test-Amdo(WER)
TDNN-LSTMP	25.54%	55.81%
TDNN-LSTMP(M-A)	24.28%	56.00%
TDNN-LSTMP(Ü-A)	23.18%	**54.76%**

Table 6: *Amdo to Ü-Tsang transfer*

Model	dev-Ü-Tsang(WER)	test-Ü-Tsang(WER)
TDNN-LSTMP	26.66%	40.85%
TDNN-LSTMP(M-Ü)	24.60%	**34.29%**
TDNN-LSTMP(A-Ü)	28.41%	39.14%

Table 5 and Table 6 show the transfer learning effects of the two dialects. In the Amdo dialect model, using the Ü-Tsang dialect model as an initial model, The Amdo baseline system achieves a relative 9.2% WER decrease on the dev-Amdo corpus, and a relative 1% WER decrease on the test-Amdo corpus. In the Ü-Tsang dialect model, we use the Amdo dialect model as an initial model, which is less effective than the baseline system on the dev-Ü-Tsang corpus. But compared with the Ü-Tsang baseline system, the test-Ü-Tsang corpus achieves a relative 4.1% WER decrease. The transfer learning between the two dialects shows a certain effect on each test corpus. The results of the Amdo dialect experiment show that the model transfer using Ü-Tsang dialect model can achieve certain effects, but the improvement is not obvious on the test-Amdo set. However, in the Ü-Tsang dialect experiment, using Mandarin model transfer can achieve great improvement, while using Amdo model transfer promotion is not obvious, and there will even be negative transfer on the dev-Ü-Tsang set.

5. Discussion

This paper studies the effects of transfer learning in Tibetan acoustic models. From the results, it can be seen that the transfer from Mandarin to Tibetan has a particularly good effect on the Ü-Tsang dialect, but with little influence on the Amdo dialect. It may be that the Ü-Tsang dialect has similarities in articulation rules with Mandarin. For example, Ü-Tsang dialects and Mandarin are both languages with tones. The Ü-Tsang dialect data we collected is mainly a Lhasa spoken language. There are 4 retroflex (zh, ch, sh, and r) in the dialects of the Ü-Tsang dialects, which are consistent with the Mandarin. The fricatives and affricates of the Ü-Tsang dialect are all voiceless, and there are no voiced consonants. Ancient media and voiced affricate have evolved into unvoiced sounds, which is consistent with Mandarin as well. The Amdo dialect is not toned but there are consonants such as voiced plosives, voiced affricates and voiced fricatives as well as some complex consonant clusters. However, there is no consonant cluster in Mandarin and there is only consonant cluster with alveolar nasals in Lhasa spoken language. These lead to ineffective results in the transfer learning between the two dialects. Another reason is that we use a Ü-Tsang lexicon for two dialects, it may make the Amdo model less reliable. Therefore, the Amdo model doesn't show good results when it transfer to the Ü-Tsang model, and the Ü-Tsang model transfer to the Amdo model show the normal results.

6. Conclusion

In this paper, we first used the TDNN-LSTMP structure to model the dialects of Tibetan,Ü-Tsang and Amdo, and established a baseline system of the dialects. We used the method of model transfer in transfer learning to transfer the hidden layers of trained models to Tibetan models. Then we studied the effect of model transfer from Mandarin to Tibetan. The results show that the model transfer of Mandarin to Ü-Tsang dialect is effective on datasets. However, that of Mandarin to Amdo dialect does not work well. Finally, we studied the model transfer effects between Ü-Tsang dialect and Amdo dialect. Experimental results show that the model transfer between two Tibetan dialects is effective. Comparing the results of two transfer experiments, we found in the dialect of Amdo, we can get better results when using the Ü-Tsang model transfer to the Amdo model. While in the Ü-Tsang dialect, using the Mandarin model transfer to the Ü-Tsang model can achieve better results.

7. References

[1] Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083.

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

[3] Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on audio, speech, and language processing, 20(1), 30-42.

[4] 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).

[5] 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.

[6] 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.

[7] Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., ... & Khudanpur, S. (2016, September). Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In Interspeech (pp. 2751-2755).

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

[9] Ghoshal, A., Swietojanski, P., & Renals, S. (2013, May). Multilingual training of deep neural networks. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 7319-7323). IEEE.

[10] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang,Phoneme recognition using time-delay neural networks, IEEE Transactions on Acoustics, Speech, and Signal Processing,vol. 37, no. 3, pp. 328339, Mar. 1989.

[11] Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In Sixteenth Annual Conference of the International Speech Communication Association.

[12] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[13] Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In Acoustics, speech and signal processing (icassp), 2013 ieee international conference on (pp. 6645-6649). IEEE.

[14] Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth annual conference of the international speech communication association.

[15] 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.

[16] Ghoshal, A., Swietojanski, P., & Renals, S. (2013, May). Multilingual training of deep neural networks. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 7319-7323). IEEE.

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