Research on Acoustic Model of Multi-Task Learning for Speech Recognition

Chunwei Gong¹, Guanyu Li² and Lisai Luo¹*
¹National Languages Information Technology Department, Northwest Minzu University, Lanzhou, Gansu Province, 730000, China
²National Languages Information Technology Department, Northwest Minzu University, Lanzhou, Gansu Province, 730000, China
*Corresponding author’s e-mail: hdd_gcw@163.com

Abstract. Applying the idea of multi-task learning, this paper proposes a speech recognition acoustic model structure, which can share information between multiple tasks to improve the performance of each task. Applying this structure to the acoustic model modeling of resource-poor languages has achieved good recognition results.

1. Introduction

Establishing a speech-specific automatic speech recognition (ASR) acoustic model is a mature technology that can be used to train acoustic models by collecting and transcribing large amounts of speech data. However, when it comes to addressing ASR in some resource-poor languages, data collection is difficult and the cost of tagging is high, so other solutions are often needed. We use the idea of multi-tasking learning to share acoustic information between existing resource-rich languages and resource-poor languages to build a large set of acoustic models. In this way, we solve the problems caused by insufficient data.

Multi-task learning is a machine learning technique designed to improve the generalization ability of a model by jointly learning multiple related tasks [1]. The key to the successful application of multi-task learning is the high correlation between tasks. In this paper, we take Chinese speech recognition and Tibetan speech recognition as two tasks. First, Chinese and Tibetan belong to the same language family, so they have a high correlation. Secondly, the Tibetan language is relatively scarce and the Chinese language resources are relatively easy to obtain. In order to improve the Tibetan speech recognition performance, we have performed relevant experiments on these two data sets and achieved good experimental results.

In fact, this method of information sharing using the relationship between languages has long been recognized by researchers. For example, it has succeeded in large-vocabulary ASR multilingual acoustic modeling research[2] and it has also been well applied in cross-lingual speech recognition [3] and It also has a good application in low-resource speech recognition.[4]. The above studies have used information sharing between different languages. The phoneme sets used are public. This is the biggest difference from this article. The approach taken in this article is that the phoneme sets used for various tasks are not fused, and each uses its own.

This paper treats Chinese speech recognition and Tibetan speech recognition as two different tasks. The core idea of this research is to treat the hidden layer of the neural network as a feature extractor, where the input and output correspond to each other, and the hidden layers of the two tasks are shared.
The main purpose of doing this is to use the neural network to discover some kind of internal connection in speech. We use this multi-task learning framework to improve the performance of Tibetan speech recognition. The advantage of jointly training acoustic models is that it can effectively use the characteristics between languages to improve the speech recognition performance of some resource-poor languages.

The rest of the paper is introduced as follows, the second part introduces the framework of single-task and multi-task models, the third part shows the experimental results; the fourth part gives the research conclusions and future work.

2. Model

2.1. Single task model

First introduce the single-task model of speech recognition we use. The ASR structure selected in this paper is a time-delay neural networks (TDNN). TDNN will stitch the current output of the hidden layer with the output of several times before and after it, and use it as the input of the next hidden layer to model longer historical information. TDNN has achieved good results in speech recognition.

Since the TDNN structure will calculate the hidden layer stimulus of all adjacent frames, the hidden layer input will contain many overlapping contexts, and the amount of calculation is large during training. Therefore, this paper uses the subsampling method. In this way, TDNN can use non-adjacent frames, reduce the context overlap and greatly reduce the amount of calculation during training. The structure of the TDNN using the subsampling method is shown in Figure 1 below.

![Figure 1. Time-delay neural networks using subsampling method](image)

In the TDNN using the sub-sampling method, the first layer of stitching [-2,2], that is, two frames before and after the current frame is spliced; the second layer of stitching [-1, -2], uses only the previous time of the current frame. Frames and frames at the second moment in the future; the third layer fights the frame [-3, +3], using only the frames at the third moment in the current frame and the frames at the third moment in the next frame; the fourth layer Frame {-7, +2}, that is, the frame from the seventh moment before and the frame from the second moment after using the current frame. Through the sub-sampling method, the TDNN model is simplified and the context information is reduced.

2.2. Multitasking Model

The multi-task learning model consists of three parts: 1) shared hidden layers as task-independent feature extractors; 2) different task-specific output layers as task-related classifiers or regressors; and 3)
task-specific input layers. The key to the multitasking model is that the hidden layer is shared between different tasks. As shown in the figure below, we divide multi-task learning into the following three types: multi-label, multi-data, multi-label + multi-data. The black layer is a layer shared by each task, the blue layer is a task-related input layer, and the red layer is a task-related output layer [5].

As shown in Figure 4 above, the multi-tasking structure used in this article is a multi-label + multi-data structure. The basic idea is to input the main task and auxiliary tasks to the shared layer of the network in the same way. Different tasks correspond to different output layers. In the shared layer, the main task and auxiliary tasks learn from each other and constrain each other to achieve a certain effect. The structure of the multi-task acoustic model used in this study is the same as the structure used for the single task above, and it is also a TDNN structure. We divide multiple tasks into main tasks and auxiliary tasks, corresponding to their respective weights.

3. Experiment
This method has been tested in the AISHELL database and the Tibetan database, both of which already have text labels. We first show the ASR baseline effect of a single task, and then show the experimental results of a multi-tasking model. All experiments were performed using the Kaldi toolkit.

3.1. Database
Training set: AISHELL data set includes 340 speakers and 120098 sentences; Tibetan data set includes 152 speakers and 73845 sentences. This set is used to train two TDNN-based single-task systems and a multi-task system.

Test set: aishell test set consists of 20 speakers and 7176 sentences; Tibetan test set consists of 4 speakers and 1631 sentences. These two test sets are used to evaluate the performance of single-tasking and multi-tasking ASR.

3.2. Single task ASR baseline system
The ASR system is basically constructed according to the formula of Kaldi WSJ s5 nnet3. We use the TDNN structure of the subsampling method. The size of the cells is 1024. The model was trained using Natural Stochastic Gradient Descent (NSGD) algorithm [17]. Features use 40-dimensional High-resolution MFCC features to stitch 3-dimensional Pitch features and 100-dimensional ivector features.

Because training TDNN-HMM model requires frame-level labeling. So before training the TDNN-HMM model, you need GMM-HMM to force the alignment of the features. The specific steps are: training the monophone GMM-HMM model using the training set, training using the Flat-start method for rapid initialization, and initial alignment using the uniform alignment method. After each iteration, an alignment operation is performed, and then the single-phone GMM-HMM is used to force the alignment of the training set, and then the training set is used to train the three-phone GMM-HMM.
model, and then the trained three-phone GMM-HMM model pair is used. Align the training set, use the training set to train a three-phone GMM-HMM model, repeat this process for LDA + MLLT three-phone GMM-HMM modeling, speaker adaptive training (SAT) GMM-HMM modeling, Finally, a good GMM-HMM model based on speaker adaptive training is used to provide alignment for the TDNN-HMM model.

The Chinese ASR output layer consists of 3064 units, and the Tibetan ASR output layer consists of 2952 units, which is equal to the total number of pdf files trained in the traditional GMM system to guide the TDNN model. The baseline performance report is shown in the following table.

| Table 1. Single task ASR baseline GMM-HMM experimental results. |
|---------------------------------------------------------------|
| **GMM**                                                      |
| Aishell: 12.79                                               |
| Tibetan: 17.09                                               |

| Table 2. Single task ASR baseline TDNN-HMM experimental results. |
|---------------------------------------------------------------|
| **TDNN**                                                      |
| Aishell: 9.72                                                 |
| Tibetan: 13.87                                                |

3.3. Multi-task joint training

The multi-task TDNN structure used in this paper is modified on the basis of nnet3. The specific structure is the same as the single-task structure. The features used by the main task and auxiliary tasks are exactly the same as those used by the single task. Due to the flexibility of the multi-tasking TDNN structure, it is not possible to evaluate all configurations. We selected typical results and report the results in the table below. We just show the ASR results on the aforementioned combined dataset.

| Table 3. Multi-task ASR baseline TDNN-HMM experimental results. |
|---------------------------------------------------------------|
| **Font** | **Spacing** |
|-----------|-------------|
| TDNN(5 Hidden Layer+1 Output Layer) | 10.20 | 9.28 |
| TDNN(5 Hidden Layer+2 Output Layer) | 10.15 | 9.10 |
| TDNN(5 Hidden Layer+3 Output Layer) | 10.09 | 9.02 |

From the results shown in the table above, we first observe that the multi-tasking TDNN structure has a certain improvement compared to the single-tasking structure, which indicates that the two tasks in the multi-tasking TDNN structure promote and promote each other. However, this observation is based on current experiments only. With more data, more information may bring additional benefits.

Again, these observations are based on a relatively small database. With more data, performance with different configurations can become different.

4. Summary

This paper proposes a TDNN structure for multi-task learning speech recognition. This structure integrates speech recognition in multiple languages. This experiment has obtained very good experimental results in Tibetan and Chinese. The proposed method can simultaneously learn the inherent information of different languages, thereby improving the speech recognition performance of each language. Future work includes research using different features, such as plp features; Analyze the study of data on different shared layers; Voiceprint recognition of the same language will also be tested in subsequent research. This article aims to train relatively good speech recognition systems with less labeled data.
Acknowledgments
The author thanks Li Guanyu and Luo Lisai for their valuable suggestions. We thank the Northwest University for Nationalities for its research and innovation project "Research and Implementation of Tibetan Speech Keyword Detection Technology Based on LSTM" for its support.

References
[1] Lin H, Deng L, Yu D, et al. A study on multilingual acoustic modeling for large vocabulary ASR[J]. 2009.
[2] Yu D, Deng L, Liu P, et al. Cross-lingual speech recognition under runtime resource constraints[C]// Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan. IEEE, 2009.
[3] Thomas S, Ganapathy S, Hermansky H. Multilingual MLP features for low-resource LVCSR systems[C]// IEEE International Conference on Acoustics. IEEE, 2012.
[4] W. Byrne, et al., “Towards language independent acoustic modeling”, Proc. ICASSP, 2000 [2]. P. Cohen, et al., “Towards a universal speech recognizer for multiple languages”, in Proc. ASRU 1997. [3] L. Deng, “Integrated-multilingual speech recognition using universal phonological features in a functional speech production model,” Proc. ICASSP, Munich, Germany, 1997, Vol. 2, pp. 1007-1010.
[5] Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, Marc’Aurelio Ranzato, Matthieu Devin, and Jeffrey Dean. Multilingual acoustic models using distributed deep neural networks. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 8619–8623. IEEE, 2013.